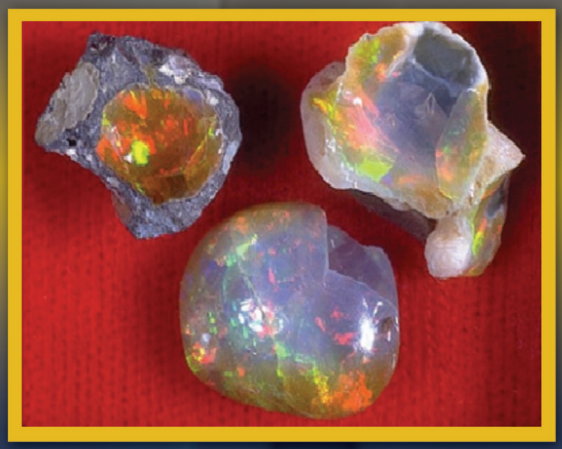
Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology



Marc J. Madou

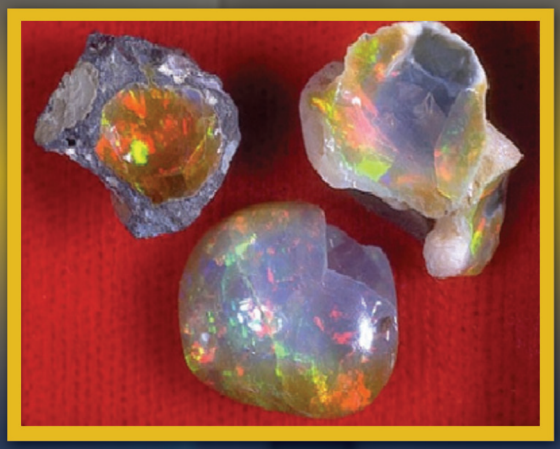
CRC Press
Taylor & Francis Group

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME I

Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology



Marc J. Madou

CRC Press
Taylor & Francis Group

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME I

Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME I

Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology

Marc J. Madou

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2012 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper

International Standard Book Number: 978-1-4200-5511-5 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

I dedicate this third edition of Fundamentals of Microfabrication to my family in the US and in Belgium and to all MEMS and NEMS colleagues in labs in the US, Canada, India, Korea, Mexico, Malaysia, Switzerland, Sweden and Denmark that I have the pleasure to work with. The opportunity to carry out international research in MEMS and NEMS and writing a textbook about it has been rewarding in terms of research productivity but perhaps even more in cultural enrichment. Scientists have always been at the frontier of globalization because science is the biggest gift one country can give to another and perhaps the best road to a more peaceful world.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Roadmap ix
Author xi
Acknowledgments xiii

INTRODUCTION
MEMS and NEMS Foundations

Introduction 2

1 Historical Note: The Ascent of Silicon, MEMS, and NEMS **5**

2 Crystallography 37

3 Quantum Mechanics and the Band Theory of Solids 75

4 Silicon Single Crystal Is Still King 215

5 Photonics 299

6 Fluidics 435

7 Electrochemical and Optical Analytical Techniques 517

Index **631**

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Roadmap

In *Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology* we lay the foundations for a qualitative and quantitative theoretical understanding of micro- and nanoelectromechanical systems, i.e., MEMS and NEMS. In integrated circuits (ICs), MEMS, and NEMS, silicon (Si) is still the substrate and construction material of choice. A historical note about the ascent of silicon, MEMS and NEMS is the topic of Chapter 1. The necessary solid-state physics background to understanding the electronic, mechanical, and optical properties of solids relied on in ICs, MEMS and NEMS is covered in Chapters 2–5. Many important semiconductor devices are based on crystalline materials because of their reproducible and predictable electrical properties. We cover crystallography in Chapter 2. The ultimate theory in modern physics today to predict physical, mechanical, chemical, and electrical properties of atoms, molecules, and solids is quantum mechanics. Quantum mechanics and the band theory of solids are presented in Chapter 3. The relevance of quantum mechanics in the context of ICs and NEMS cannot be underestimated, and the profound implications of quantum physics for nanoelectronics and NEMS are a recurring topic throughout this book. Given the importance of single-crystal Si (SCS) for IC, MEMS, and NEMS applications, we analyze silicon crystallography and band structure in more detail in Chapter 4. This chapter also elucidates all the single-crystal Si properties that conspired to make Si so important in electronic, optical, and mechanical devices that one might rightly call the second half of the 20th century the Silicon Age. Photonics, treated in Chapter 5, involves the use of radiant energy and uses photons the same way that electronic applications use electrons. We review the distinctive optical properties of bulk 3D metals, insulators, and semiconductors and summarize effects of electron and photon confinement in lower-dimensional structures. We show how evanescent fields on metal surfaces enable the guiding of light below the diffraction limit in plasmonics. Plasmonics is of growing importance for use in submicron lithography, near-field optical microscopy, enhancement of light/matter interaction in sensors, high-density data storage, and highly integrated optic chips. We also delve into the fascinating new topic of metamaterials, man-made structures with a negative refractive index, and explain how this could make for perfect lenses and could change the photonic field forever. In Chapter 6 we introduce fluidics, compare various fluidic propulsion mechanisms, and discuss the influence of miniaturization on fluid behavior. Given the high level of interest, fluidics for miniaturized analytical equipment is covered in this chapter as well. Chapter 7 combines a treatise on electrochemical and optical analytical processes whose implementation is often attempted in miniaturized components and systems.

Note to the Reader: *Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology* was originally composed as part of a larger book that has since been broken up into three separate volumes. *Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology* represents the first volume in this set. The other two volumes include *Manufacturing Techniques for Microfabrication and Nanotechnology* and *From MEMS to Bio-NEMS: Manufacturing Techniques and Applications*. Cross-references to these books appear throughout the text and will be referred to as Volume II and Volume III, respectively. The interested reader is encouraged to consult these volumes as necessary.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Author

Dr. Madou is the Chancellor's Professor in Mechanical and Aerospace Engineering (MEA) at the University of California, Irvine. He is also associated with UC Irvine's Department of Biomedical Engineering and the Department of Chemical Engineering and Materials Science. He is a Distinguished Honorary Professor at the Indian Institute of Technology, Kanpur, India and a World Class University Scholar (WCU) at UNIST in South Korea.

Dr. Madou was Vice President of Advanced Technology at Nanogen in San Diego, California. He specializes in the application of miniaturization technology to chemical and biological problems (bio-MEMS). He is the author of several books in this burgeoning field he helped pioneer both in academia and in industry. He founded several micromachining companies.

Many of his students became well known in their own right in academia and through successful MEMS start-ups. Dr. Madou was the founder of the SRI International's Microsensor Department, founder and president of Teknekron Sensor Development Corporation (TSDC), Visiting Miller Professor at UC Berkeley, and Endowed Chair at the Ohio State University (Professor in Chemistry and Materials Science and Engineering).

Some of Dr. Madou's recent research work involves artificial muscle for responsive drug delivery, carbon-MEMS (C-MEMS), a CD-based fluidic platform, solid-state pH electrodes, and integrating fluidics with DNA arrays, as well as label-free assays for the molecular diagnostics platform of the future.

To find out more about those recent research projects, visit http://www.biomems.net.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Acknowledgments

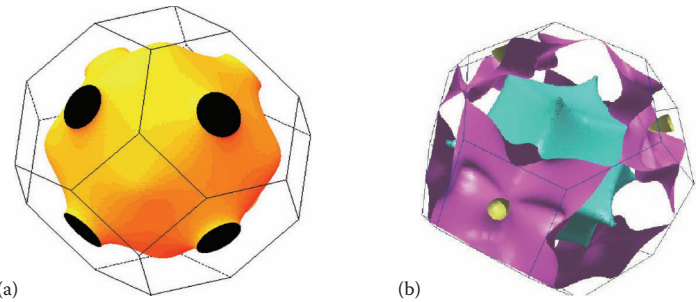
I thank all of the readers of the first and second editions of Fundamentals of Microfabrication as they made it worthwhile for me to finish this completely revised and very much expanded third edition. As in previous editions I had plenty of eager reviewers in my students and colleagues from all around the world. Students were especially helpful with the question and answer books that come with the three volumes that make up this third edition. I have acknowledged reviewers at the end of each chapter and students that worked on questions and answers are listed in the questions sections. The idea of treating MEMS and NEMS processes as some of a myriad of advanced manufacturing approaches came about while working on a WTEC report on International Assessment Of Research And Development In Micromanufacturing (http://www.wtec.org/micromfg/report/Micro-report.pdf). For that report we travelled around the US and abroad to visit the leading manufacturers of advanced technology products and quickly learned that innovation and advanced manufacturing are very much interlinked because new product demands stimulate the invention of new materials and processes. The loss of manufacturing in a country goes well beyond the loss of only one class of products. If a technical community is dissociated from manufacturing experience , such as making larger flat-panel displays or the latest mobile phones, such communities cannot invent and eventually can no longer teach engineering effectively. An equally sobering realization is that a country might still invent new technologies paid for by government grants, say in nanofabrication, but not be able to manufacture the products that incorporate them. It is naïve to believe that one can still design new products when disconnected from advanced manufacturing: for a good design one needs to know the latest manufacturing processes and newest materials. It is my sincerest hope that this third edition motivates some of the brightest students to start designing and making things again rather than joining financial institutions that produce nothing for society at large but rather break things.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Introduction

MEMS and NEMS Foundations

Miniaturization science is the science of making very small things. In top-down micro- and nanomachining, one builds down from large chunks of material; in bottom-up nanochemistry, one builds up from smaller building blocks. Both require a profound understanding of the intended application, different manufacturing options, materials properties, and scaling laws. The resulting three-dimensional structures, ranging in size from subcentimeters to subnanometers, include electronics, photonics, sensors, actuators, micro- and nanocomponents, and micro- and nanosystems.



(a) (b)

(a) Copper Fermi surface—FCC sixth band with eight short necks touching the eight hexagonal zone faces. (Fermi surface database at http://www.phys.ufl.edu/fermisurface.) (b) Platinum Fermi surface—FCC fourth, fifth, and sixth bands. (Fermi surface database at http://www.phys.ufl.edu/fermisurface.)

No one behind, no one ahead. The path the ancients cleared has closed. And the other path, everyone's path, easy and wide, goes nowhere. I am alone and find my way.

Dharmakirti
(7th century India)

Introduction

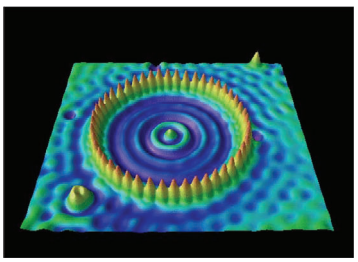
Chapter 1 Historical Note: The Ascent of Si, MEMS, and NEMS
Chapter 2 Crystallography
Chapter 3 Quantum Mechanics and the Band Theory of Solids
Chapter 4 Silicon Single Crystal Is Still King
Chapter 5 Photonics
Chapter 6 Fluidics
Chapter 7 Electrochemical and Optical Analytical Techniques

Introduction

In Volume I, we lay the foundations for a qualitative and quantitative understanding of micro- and nanoelectromechanical systems, i.e., MEMS and NEMS. In integrated circuits (ICs), MEMS, and NEMS, silicon (Si) is still the substrate and building material of choice. A historical note about the history of the ascent of silicon, MEMS, and NEMS is the topic of Chapter 1.

The necessary solid-state physics background of electronic, mechanical, and optical properties of solids relied on in MEMS and NEMS is covered in Chapters 2–5. Solid-state physics is the study of solids. A major part of solid-state physics is focused on crystals because the periodicity of atoms in a crystal facilitates mathematical modeling, but more importantly because crystalline materials often have electrical, optical, or mechanical properties that can be easier exploited for engineering purposes. In Chapter 2, we detail crystalline materials in which atoms are arranged in a pattern that repeats periodically in three dimensions. The materials covered here prepare the reader for Chapter 3, which explains the band theory of solids based on quantum mechanics. The relevance of quantum mechanics in the context of ICs and NEMS cannot be underestimated, and the profound implications of quantum physics for nanoelectronics and NEMS is a recurring topic throughout this book. This is followed in Chapter 4 by a description of the single-crystal Si band structure, the growth of single crystals of Si, its doping, and oxidation. In this chapter, we also review the single-crystal Si properties that conspired to make Si so important in electronic, optical, and mechanical devices so that one might rightly call the second half of the 20th century the Silicon Age. Although the emphasis in this book is on nonelectronic applications of miniaturized devices, we briefly introduce different types of diodes and two types of transistors (bipolar and MOSFET). In Chapter 5, we introduce photonics. We compare electron and photon propagation in materials and contrast electron and photonic confinement structures and the associated evanescent wave phenomena. We also delve into the fascinating new topic of metamaterials, artificially engineered materials possessing properties (e.g., optical, electrical) that are not encountered in naturally occurring ones. An introduction to diode lasers, quantum well lasers, and quantum cascade lasers concludes the photonics section.

Fluidics and electrochemical and optical analytical techniques are important current applications of MEMS and NEMS. In Chapter 6 we introduce fluidics, compare various fluidic propulsion mechanisms, and discuss the influence of miniaturization on fluid behavior. Given the current academic and industrial interest, fluidics in miniaturized analytical equipment is detailed separately at the end of this chapter. Chapter 7 combines a treatise on



STM image showing standing waves in a 2D electron gas trapped in a "quantum corral" made by positioning Fe atoms on a Cu (111) surface.[1]

electrochemical and optical analytical techniques. Using sensor examples, we introduce some of the most important concepts in electrochemistry, i.e., the electrical double layer, potentiometry, voltammetry, two- and three-electrode systems, Marcus' theory of electron transfer, reaction rate- and diffusion rate-controlled electrochemical reactions, and ultramicroelectrodes. Many researchers use MEMS and NEMS to miniaturize optical components or whole instruments for absorption, luminescence, or phosphorescence spectroscopy. Optical spectroscopy is concerned with the production, measurement, and interpretation of electromagnetic spectra arising from either emission or absorption of radiant energy by matter. The sensitivity of these optical sensing techniques and the analysis of how amenable they are to miniaturization (scaling laws) are also compared herein. Chapter 7 ends with a comparison of the merits and problems associated with electrochemical and optical measuring techniques.

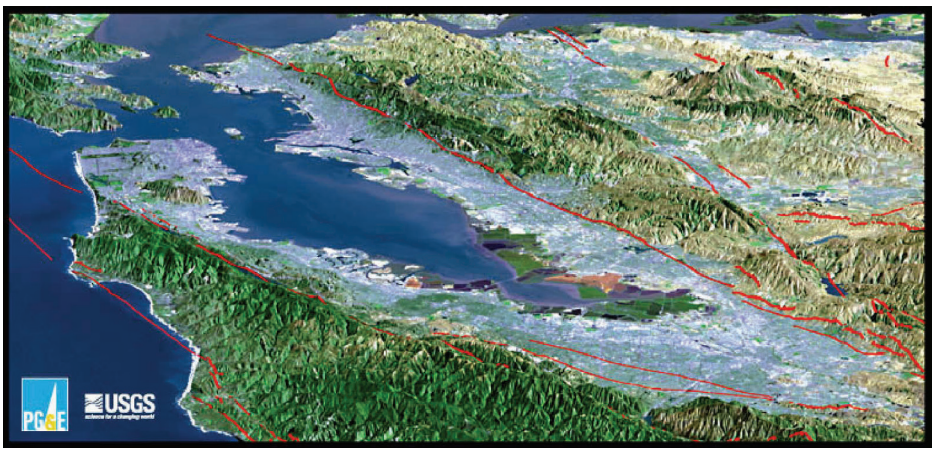
Reference

1. Crommie, M. F., C. P. Lutz, and D. M. Eigler. 1993. Confinement of electrons to quantum corrals on a metal surface. *Science* 262:218–20.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

Historical Note: The Ascent of Silicon, MEMS, and NEMS



Silicon Valley is the only place on Earth not trying to figure out how to become Silicon Valley (Robert Metcalfe, father of the ethernet). (From PG&E and USGS.)

Outline

Silicon in Integrated Circuits

MEMS

NEMS

Acknowledgments

Appendix 1A: International Technology Roadmap for Semiconductors (ITRS)

Appendix 1B: Worldwide IC and Electronic Equipment Sales

Questions

References

Silicon in Integrated Circuits

In 1879, William Crookes recounted his experiments on passing electric discharges through an evacuated glass tube for the Royal Society, thus describing the first cathode ray tube (CRT). Four years later, Thomas Alva Edison and Francis Upton discovered the "Edison effect." They introduced a metal plate into an incandescent electric light bulb (invented by Edison in 1879) in an attempt to keep the bulb from turning black (Figure 1.1). It did not work, but they discovered that there was a current between the lighted filament and the separate metal plate when the plate was positively charged but not when it was negatively charged. This led Edison and Upton to stumble on the basic principle of the operation of the vacuum tube (rectification!).

The first diode tube we owe to John Fleming, who, in 1904, filed a patent for a "valve" vacuum tube, also called a *Fleming valve* or *Fleming diode*.

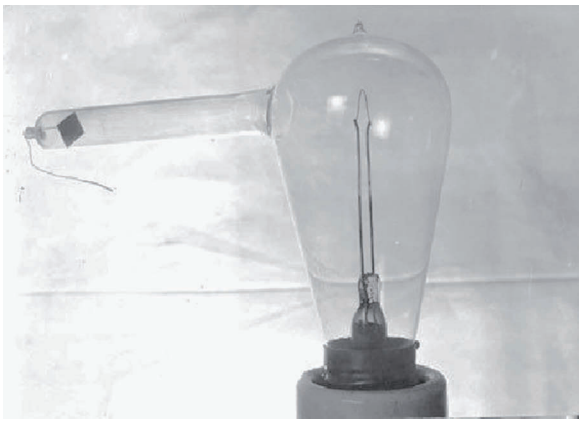


FIGURE 1.1 Edison bulb used to demonstrate the "Edison effect."

Early researchers thought of electricity as a kind of fluid, leading to the inherited jargon such as current, flow, and valve. Fleming recognized the importance of Edison and Upton's discovery and demonstrated it could be used for the rectification of alternating currents. Interestingly, Fleming, at first, tried to get reliable rectification from single-crystal rectifiers used in crystal set radios (Figure 1.2), but could never get them to work well enough, so he switched to tubes!

J.J. Thomson, in 1887, convincingly showed that an electrical current was really an electron flow, and Fleming could explain the rectification in his diode tube as electrons boiling of the heated filament and flowing to the metal plate (thermionic emission). Because the plate was not hot enough to emit electrons, no current could go in the opposite direction. Thus, the Edison effect always produced direct

FIGURE 1.2 Early crystal set radio with a galena (lead sulphide) and the "cat's whisker" (the small coil of wire) that was used to make contact with the crystal.

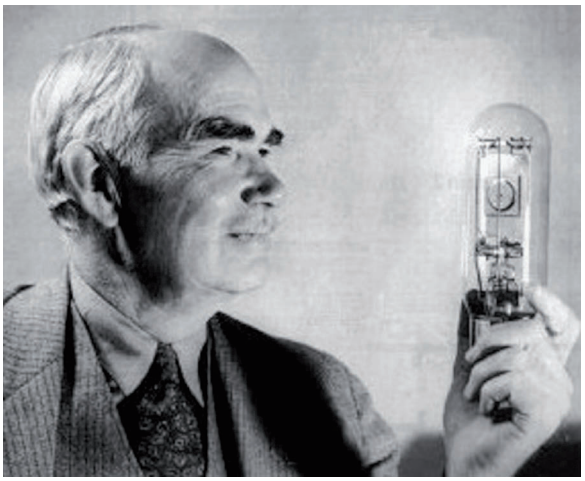


FIGURE 1.3 Lee De Forest in 1906 with the Audion, the first triode.

current only. In 1906, an American scientist, Lee De Forest (Figure 1.3), invented the vacuum tube amplifier or triode based on the two-element vacuum tube invented by Fleming. De Forest, reportedly a tireless self-promoter, added an electrode—the *grid*—to the Fleming diode, and inserted it between the anode and the cathode. With this grid the diode became an active device, i.e., it could be used for the amplification of signals (say, for example, in radios) and as a switch (for computers). Hence, the amplifying vacuum tube, the ancestor of the transistor, was born. A gate in a dam controls huge amounts of flowing water with relatively small movements. Similarly, a small signal applied to the grid controls the much larger signal between anode and cathode.

Vacuum tubes—miniature particle accelerators—dominated the radio and television industries until the 1960s, and were the genesis of today's huge electronics industry. However, tubes were fragile, large, very power hungry, and costly to manufacture. The industry needed something better. That today's world is largely electronic—e.g., automobiles, home appliances, even books, writing tablets, and tally sheets—is because of solid-state electronics,* not vacuum tubes.

It is true, albeit unfortunate, that World War II and the subsequent Cold War era is what spurred research and development in solid-state electronic devices. Human foibles led to faster development of

* Based on or consisting chiefly or exclusively of semiconducting materials, components, and related devices.

RAdio **D**etecting **A**nd **R**anging or RADAR, **SO**und **N**avigation **A**nd **R**anging or SONAR, and many other technological innovations. As we all know, the list of innovations made to feed human aggression did not abate. Alan Turing led the team in England that in 1943 built the Colossus coding and deciphering machine. The Colossus was a special-purpose computer used to break the German code ULTRA, encrypted using ENIGMA machines. Breaking the German code was one of the keys to the success of the D-Day invasion. The Harvard Mark I and later II, III, and IV were general-purpose electromechanical calculators (sponsored by the U.S. Navy) to compute artillery and navigation tables—the same purpose intended, 100 years earlier, by Babbage for his analytical engine (Figure 1.4).

John Mauchly and Presper Eckert started work on the first electronic computer, the ENIAC (Electronic Numerical Integrator and Computer), at the University of Pennsylvania in 1943. The ENIAC, having been a secret during the war, was unveiled in Philadelphia in 1946. This computer featured 17,468 vacuum tubes used as switches and consumed 174 kW of power, enough to light 10 homes! Several tubes would fail every day until the engineers decided to never turn off the machine. This increased the average time until a tube would fail to 2 days. ENIAC was designed to calculate munition trajectory tables for the U.S. Army. It was U-shaped, 25 m long, 2.5 m high, 1 m wide, and weighed more than 30 tons (see Figure 1.5a). Programming was done by plugging cables and setting switches. By the mid-1970s, identical ENIAC functions could be achieved by a 1.5- × 1.5-cm silicon die, and the original Pentium processor, if fabricated using ENIAC technology, would cover more than 10 square miles.

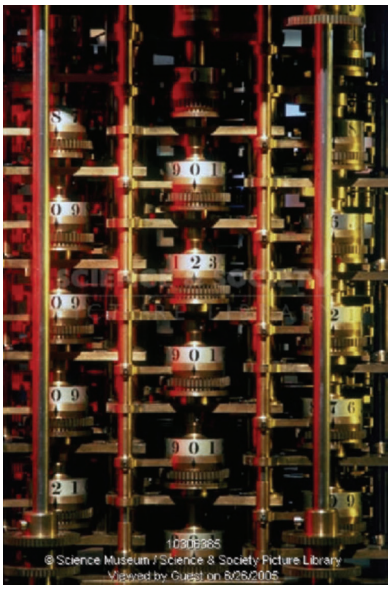


FIGURE 1.4 Charles Babbage (1791–1871) first conceived the idea of an advanced calculating machine to calculate and print mathematical tables in 1812. It was a decimal digital machine, with the value of a number being represented by the positions of toothed wheels marked with decimal numbers.

One of ENIAC's heirs was a computer called the UNIVAC (Universal Automatic Computer), considered by most historians to be the first commercially successful digital computer (Figure 1.5b). First constructed by the Eckert-Mauchly Computer Corporation (EMCC), it was taken over by Sperry-Rand. At 14.5 ft. long, 7.5 ft. high, and 9 ft. wide, the UNIVAC, priced at $1 million, was physically

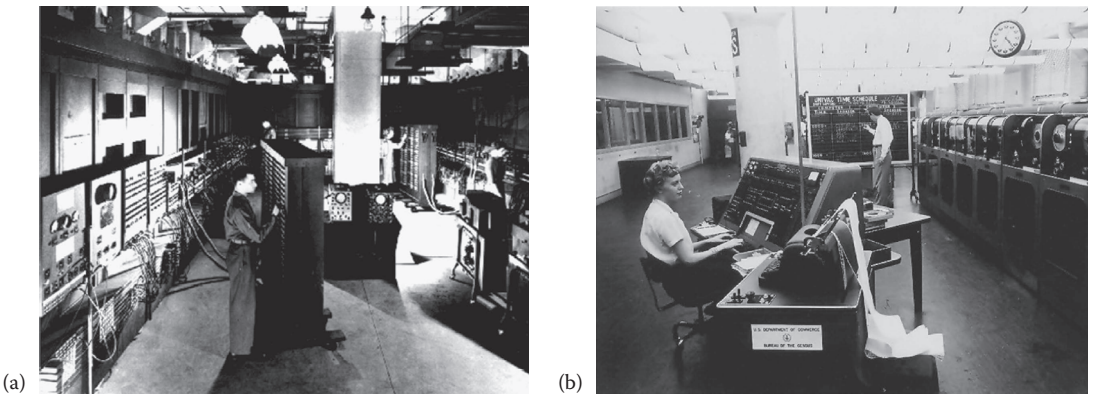


FIGURE 1.5 Electronic Numerical Integrator and Computer (ENIAC), the world's first large-scale, general-purpose electronic computer (a), and the UNIVAC, the first commercial computer (b). First-generation computers based on vacuum tubes.

FIGURE 1.6 Von Neumann in his living room. (Photograph by Alan Richards hanging in Fuld Hall, Institute for Advanced Study, Princeton, NJ. Courtesy of the Archives of the Institute for Advanced Study.)

FIGURE 1.7 The first point-contact germanium bipolar transistor. Notice the paper clip! Roughly 50 years later, electronics accounted for 10% ($4 trillion) of the world's aggregate GDP.

smaller than ENIAC but more powerful. ENIAC and UNIVAC constitute first-generation computers based on vacuum tubes.

It was the concept of the stored program, invented by John von Neumann in 1945 (Figure 1.6); the magnetic core memory, invented by An Wang at Harvard and used in grids by J.W. Forrester and colleagues at MIT for random access memory (RAM); and William Shockley's transistor, based on transistors for switches instead of tubes, that would make a second generation of computers possible, thus starting the computer revolution.

The year 1940 gave rise to an important milestone in solid-state electronics history with the invention of a silicon-based solid-state p-n junction diode by Russell Ohl at Bell Labs.[1] This device, when exposed to light, produced a 0.5 V across the junction and represented the first Si-based solar cell. Bell Labs, in 1945, established a group charged with developing an alternative to the vacuum tube. Led by Shockley (1910–1989) and including John Bardeen (1908–1991) and Walter Brattain (1902–1987), in 1947 the group made an odd-looking device consisting of semiconducting germanium (Ge), gold strips, insulators, and wires, which they called a transistor (subject of U.S. Patent #2,524,035 [1950][2]) (Figure 1.7; notice the paper clip!). For this invention Shockley, Bardeen, and Brattain were awarded the 1956 Nobel Prize for Physics (Bardeen went on to claim a second Nobel Prize for Physics in 1972 for superconductivity). Bardeen called Ohl's junction diode fundamental to the invention of the transistor. Brattain was the unassuming experimentalist, Bardeen the theorist, and Shockley* the inventor and leader (Figure 1.8).

This trio thus succeeded in creating an amplifying circuit using a point-contact bipolar transistor that *trans-ferred* resistance (hence *transistor*). Two

FIGURE 1.8 Shockley (seated), Bardeen, and Brattain.

* Unfortunately, Shockley became associated with racist ideas and briefly pursued a U.S. Senate seat (as a Republican).

wires made contact with the germanium crystal near the junction between the p- and n-zones just like the "cat whiskers"* in a crystal radio set. A few months later, Shockley devised the junction transistor, a true solid-state device that did not need the whiskers of the point-contact transistor (see also Figure 1.2). Junction transistors were much easier to manufacture than point-contact transistors, and by the mid-1950s the former had replaced the latter in telephone systems. G. Teal and J.B. Little, also from Bell Labs, were able to grow large single crystals of germanium by 1951, which led to the start of commercial production of germanium transistors in the same year. Christmas 1954 saw the first transistor radio (the Regency TR-1) built by Industrial Development Engineering Associates, which sold for $49.95 (Figure 1.9). This radio featured four germanium transistors from Texas Instruments. Although germanium was used in early transistors, by the late 1960s silicon, because of its many advantages, had taken over.

Silicon has a wider bandgap (1.1 eV for Si vs. 0.66 eV for Ge), allowing for higher operating temperatures (125–175°C vs. 85°C), a higher intrinsic resistivity (2.3×10^5 Ω cm vs. 47 Ω cm), and a better native oxide (SiO_2 vs. GeO_2 [water soluble!]).

FIGURE 1.9 Movie producer mogul Michael Todd (husband of Elizabeth Taylor in the mid-1950s) placed Regency TR-1s in gift books to commemorate his movie *Around the World in 80 Days*. The one pictured was for Shirley MacLaine.

* A cat whisker is a piece (often springy) of pointed metal wire.

The latter results in a higher-quality insulator that protects and "passivates" underlying circuitry, helps in patterning, and is useful as a mask for dopants. Finally, silicon is cheaper and much more abundant (sand!) than germanium. Second-generation computers relied on transistors instead of vacuum tubes for switches (logic gates). In recent years, germanium is making a comeback based mostly on its higher carrier mobility (three times higher than silicon-based ones), of great interest for faster circuitry, and because Ge has a lattice constant similar to GaAs, making it easier to integrate GaAs optical components with Ge-CMOS circuits.

Transistors perform functions similar to vacuum tubes, but they are much smaller, cheaper, less power hungry, and more reliable. Michael Riordan and Lillian Hoddeson's *Crystal Fire* gives, in the author's opinion, one of the best popular accounts of the invention of the transistor.[3]

The honeymoon with the transistor was quickly over; by the second half of the 1950s, new circuits on the drawing board were so big and complex that it was virtually impossible to wire that many different parts together reliably. A circuit with 100,000 components could easily require 1 million, mostly manual, soldering operations that were time consuming, expensive, and inherently unreliable. The answer was the "monolithic" idea, in which a single bloc of semiconductor is used for all the components and interconnects, invented by two engineers working at competing companies: Jack Kilby at Texas Instruments (Figure 1.10) and Robert Noyce (Figure 1.11) at Fairchild Semiconductor.

In 1958, Jack Kilby at Texas Instruments formed a complete circuit on a piece of germanium, landing U.S. Patent #3,138,743 (1959). His circuit was a simple IC oscillator with three types of components: transistors, resistors, and capacitors (Figure 1.12). Kilby got his well-deserved Nobel Prize for this work only in 2000. Technological progress and engineering feats are not often awarded a Nobel Prize, and if awarded at all they are often belated or controversial (see Kary Mullis, Nobel Laureate Chemistry 1993 for the invention of PCR).

Robert Noyce—Mr. Intel (Integrated Electronics)—then at Fairchild, introduced, with Jean Horni, planar technology, wiring individual devices together

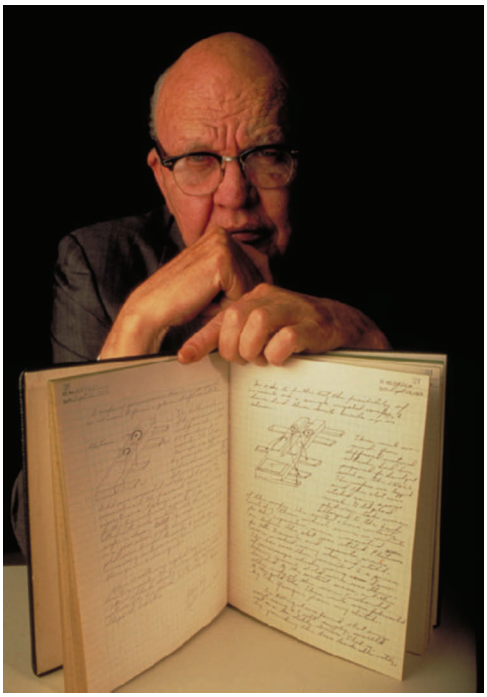


FIGURE 1.10 Jack Kilby with notebook. (TI downloadable pictures.)

on a silicon wafer surface. Noyce's "planar" manufacturing, in which all the transistors, capacitors, and resistors are formed together on a silicon chip with the metal wiring embedded on the silicon, is still used today. By 1961, Fairchild and Texas Instruments had devised methods whereby large

FIGURE 1.11 Robert Noyce in 1990.

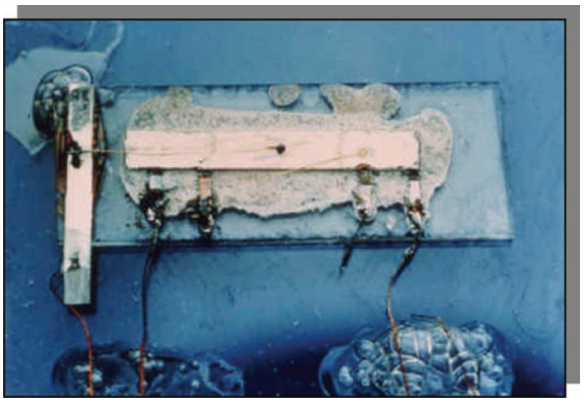


FIGURE 1.12 The first integrated circuit (germanium) in 1958 by Jack S. Kilby at Texas Instruments contained five components of three types: transistors, resistors, and capacitors.

numbers of transistors were produced on a thin slice of Si—and IC production on an industrial scale took off. The transistors on ICs were not the bipolar type but rather field effect transistor devices. The concept of a field effect transistor (FET) was first proposed and patented in the 1930s; however, it was the bipolar transistor that made it first to commercial products. Shockley resurrected the idea of the FET in the early 1950s, but it took until 1962 before a working FET was fabricated. These new FETs proved to be more compatible with both IC and Si technology.

Integrated circuits made not only third-generation computers possible but also cameras, clocks, PDAs, RF-IDs, etc. The National Academy of Sciences declared ICs the progenitor of the "Second Industrial Revolution," and Jerry Sanders, founder of Advanced Microdevices, Inc., called ICs the crude oil of the 1980s. A very well-written popular account of the invention of the IC is T.R. Reid's *The Chip: How Two Americans Invented the Microchip and Launched a Revolution*.[4]

Robert Noyce, Gordon Moore, and Andrew Grove left Fairchild to start Intel in 1968 with the aim of developing random access memory (RAM) chips. The question these inventors wanted answered was this: since transistors, capacitors, and resistors can be put on a chip, would it be possible to put a computer's central processor unit (CPU) on one? The answer came swiftly; by 1969 Ted Hoff had designed the Intel 4004, the first general-purpose 4-bit microprocessor. The Intel 4004 microprocessor was a 3-chip set with a 2-kbit read-only memory (ROM) IC, a 320-bit RAM

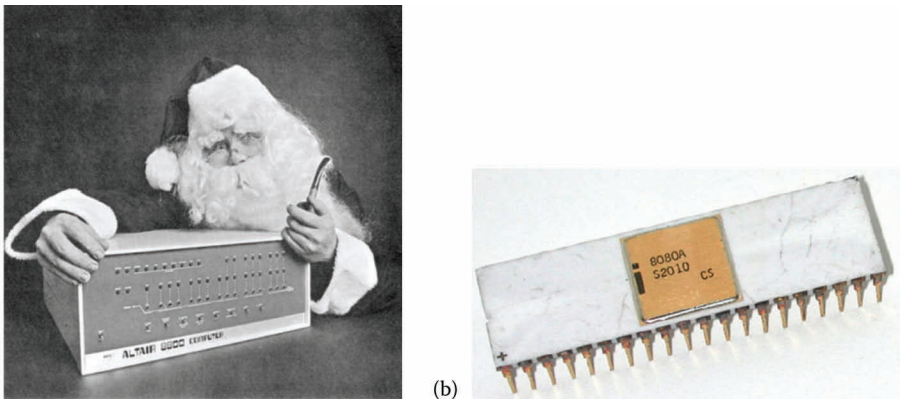


FIGURE 1.13 (a) The Altair 8800 computer and (b) the Intel 8080 microprocessor.

IC, and a 4-bit processor, each housed in a 16-pin dual in-line package (DIP). The processor, made in a 10-µm silicon gate pMOS process, contained 2,250 transistors and could execute 60,000 operations per second on a die size of 13.5 mm^2. It came on the market in 1971, giving rise to the fourth generation of computers based on microprocessors and the first personal computer (PC). The era of a computer in every home—a favorite topic among science fiction writers—had arrived!

The first desktop-size PC appeared in 1975, offered by Micro Instrumentation Telemetry Systems (MITS) as a mail-order computer kit. The computer, the Altair 8800, named after a planet on a *Star Trek* episode, retailed for $397. It had an Intel 8080 microprocessor, 256 bytes of memory (not 256K), no keyboard, no display, and no auxiliary storage device, but its success was phenomenal, and the demand for the microcomputer kit was overwhelming (Figure 1.13). Scores of small entrepreneurial companies responded to this demand by producing computers for the new market. In 1976, Bill Gates, Paul Allen, and Monte Davidoff wrote their first software program for the Altair—a BASIC (Beginners All-purpose Symbolic Instruction Code) interpreter (a high-level language translator that converts individual high-level computer language program instructions [source code] into machine instructions).

The first major electronics firm to manufacture and sell personal computers, Tandy Corporation (Radio Shack), introduced its computer model (TRS-80) in 1977. It quickly dominated the field because of the combination of two attractive features: a keyboard and a cathode ray display terminal. It was also popular because it could be programmed, and the user was able to store information by means of a cassette tape. In 1976, Steve Wozniak, who could not afford an Altair, built his own computer using a cheaper microprocessor and adding several memory chips. As a circuit board alone, it could do more than the Altair. Wozniak and Steve Jobs called it Apple I, and Jobs took on the task of marketing it while Wozniak continued with improvements (Figure 1.14). By 1977, Wozniak had built the Apple II and quit his day job. The Apple II had 16–64K RAM and secondary memory storage in the shape of a cassette tape or a 5.25-in. floppy disk drive and cost $1,300. At that time, Wozniak and Jobs formed Apple Computer,

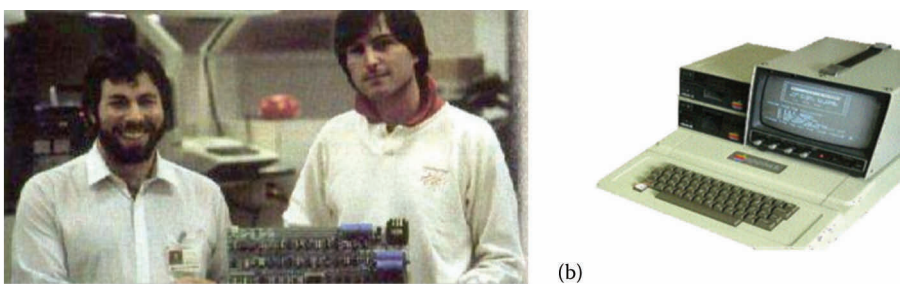


FIGURE 1.14 (a) Jobs and Wozniak with the board for Apple I and (b) the Apple II.

Inc. When it went public in 1980, its stock value was $117 million; three years later it was worth $985 million.

Vacuum tubes coexisted with their progeny, the transistor, and even with ICs for a short while. Although solid-state technology overwhelmingly dominates today's world of electronics, vacuum tubes are still holding out in some areas. You might, for example, still have a CRT (cathode ray tube) as your television or computer screen. Tubes also remain in two small but vibrant areas for entirely different reasons. The first involves microwave technology, which still relies on vacuum tubes for their power-handling capability at high frequencies. The other—the creation and reproduction of music—is a more complicated story. Tubes distort signals differently than transistors when overdriven, and this distortion is regarded as being more "pleasant" by much of the music community.

Extrapolating back to 1961, Gordon Moore in 1965 (Figure 1.15), while at Fairchild, predicted that transistors would continue to shrink, doubling in density on an IC every 18–24 months, while the price would continue to come down—this prediction we know today as Moore's Law. History has proven Moore right as evidenced by past and projected feature sizes of ICs in the International Technology Roadmap for Semiconductors (ITRS) shown in Appendix 1A and on the Internet at http://public.itrs.net (updated in 2007).[5] In this International Technology Roadmap for Semiconductors, technology modes have been defined. These modes are the feature sizes that have to be in volume manufacturing at a fixed date (year of production). The feature size is defined as half-pitch, i.e., half of a dense pair of lines and spaces (see figure in Appendix 1A).

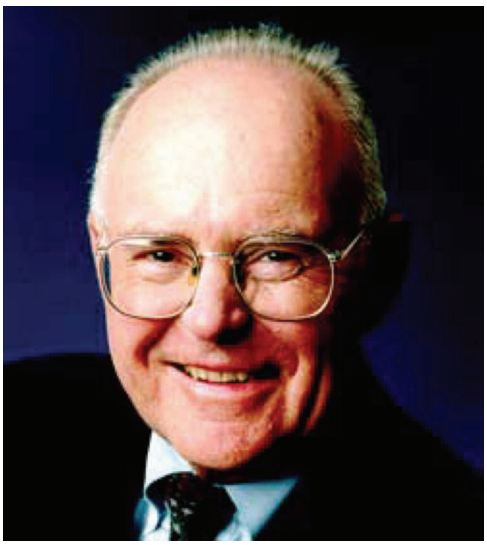


FIGURE 1.15 Gordon Moore, cofounder of Intel.

TABLE 1.1 Integration Scale and Circuit Density

IC Evolution	Acronym	Number of Logic Gates	Year of Introduction
Zero-scale integration	ZSI	1	1950
Small-scale integration	SSI	2–30	1965
Medium-scale integration	MSI	30–10^3	1970
Large-scale integration	LSI	10^3–10^5	1980
Very large-scale integration	VLSI	10^5–10^7	1985
Ultra-large-scale integration	ULSI	10^7–10^9	1990
Giga-scale integration	GSI	10^9–10^{11}	2005
Tera-scale integration	TSI	10^{11}–10^{13}	2020

In Table 1.1 the increasing numbers of devices integrated on an IC are tabulated. As we will learn in Chapters 1 and 2 on lithography in Volume II, new lithography techniques, novel device structures, and the use of new materials drive Moore's Law.

The state of the art in ICs today is a 2-GB DRAM* with 60-nm features (Samsung). Intel introduced the Core 2 Quad "Kentsfield" chip in January 2007, a chip featuring a 65-nanometer technology mode.

The 32-nm node should be achieved in 2009. The first Moore's Law is the good news, but there is a second Moore's Law that is a bit problematic; this second law states that the cost of building a chip factory doubles with every other chip generation, i.e., every 36 months. Today's technology involves Si wafers with a 12-in. diameter and factories that cost $3–4 billion to construct. With this type of start-up costs, few countries can afford to enter the IC market, and the search is on for alternative, less-expensive

* Dynamic random access memory. A type of memory component used to store information in a computer system. "Dynamic" means the DRAMs need a constant "refresh" (pulse of current through all the memory cells) to keep the stored information.

bottom-up NEMS techniques (below). From Appendix 1B, the IC market for 2003 was $166 billion worldwide, with 2004 projected at $241 billion (data by the World Semiconductor Trade Statistics [WSTS]; http://www.wsts.org).[6] Also from Appendix 1B we learn that the IC business is feeding a trillion-dollar electronic equipment business. It is worth pointing out that China is expected to control 5% of the IC market by 2010.

MEMS

Single-crystal silicon is not only an excellent electronic material but it also exhibits superior mechanical properties; the latter gave birth to the **m**icro**e**lectro**m**echanical **s**ystems (MEMS) field on the coattails of the IC industry. Originally MEMS constituted mostly mechanical types of devices based on single-crystal silicon with at least one or more of their dimensions in the micrometer range. As MEMS applications broadened, in Europe the acronym MST for ***m**icrosystem **t**echnology* became more popular. In Japan one refers to *micromachining*, and *microengineering* is popular in the United Kingdom. Development of single-crystal silicon mechanical MEMS involves the fabrication of micromechanical parts, e.g., a thin membrane in the case of a pressure sensor or a cantilever beam for an accelerometer. These micromechanical parts are fabricated by selectively etching areas of a Si substrate away to leave behind the desired geometries. The terms *MEMS* and *micromachining* came into use in 1982 to name these fabrication processes. Around the same time references to "bulk" micromachining techniques also appeared. Richard Feynman's December 26, 1959 presentation "There's Plenty of Room at the Bottom" is considered by many to be the starting bongo for MEMS (Figure 1.16) (http://www.its.caltech.edu/~feynman/plenty.html),[7] but in a practical sense, it was the invention of the transistor and the processes developed to fabricate transistors, six years earlier, that enabled MEMS.

An early milestone for the use of single-crystal silicon in MEMS was the 1956 discovery of porous Si by Uhlir.[8] His discovery eventually led to all types of interesting new, single-crystal Si-based devices, from reference electrodes for electrochemical sensors, biosensors, quantum structures, and permeable membranes to photonic crystals and photoluminescent and electroluminescent devices.

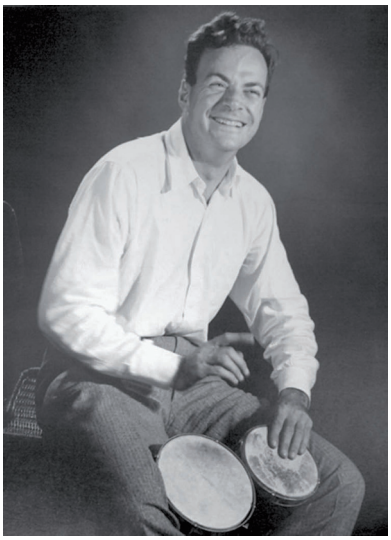


FIGURE 1.16 Richard Feynman on the bongo drums.

The first impetus for the use of single-crystal silicon as a micromechanical element in MEMS can be traced to the discovery of its large piezoresistance. Piezoresistance is the change in the resistivity of certain materials as a result of an applied mechanical strain. Charles Smith, of the Case Institute of Technology (now part of the Case Western Reserve University), during a sabbatical leave at Bell Labs in 1953, studied the piezoresistivity of semiconductors and published the first paper on the piezoresistive effect in Si and Ge in 1954.[9] The piezoresistive coefficients Smith measured demonstrated that the gauge factor* of Si and Ge strain gauges (see Figure 1.17) was 10–20 times larger than the gauge factor of metal film strain gauges, and, therefore, semiconductor gauges were expected to be much more sensitive.

Motivated by these results, companies such as Kulite and Honeywell started developing Si strain gauges commercially from 1958 onward. Pfann and colleagues, in 1961, proposed a dopant diffusion technique for the fabrication of silicon piezoresistive sensors for the measurement of stress, strain, and pressure.[10] Based on this idea, Kulite integrated

* A strain gauge is a device used to measure deformation (strain) of an object. The gauge factor of a strain gauge relates strain to change in electrical resistance.

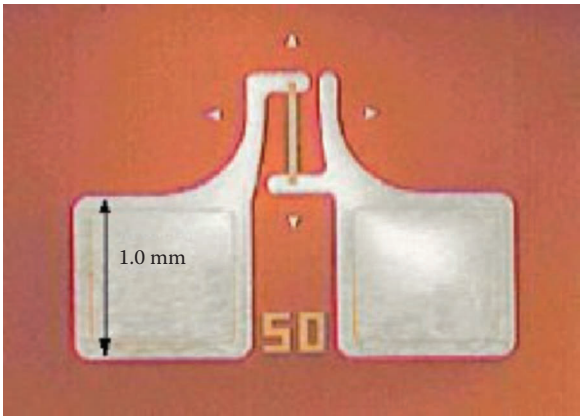


FIGURE 1.17 The Si single-crystal gauge element can be seen as the vertical bar centered between the two solder pads. The single-crystal silicon strain gauge offers sensitivities 20–50 times greater than metal foil gauges. A microphotograph of the LN-100. (BF Goodrich Advanced Micro Machines.)

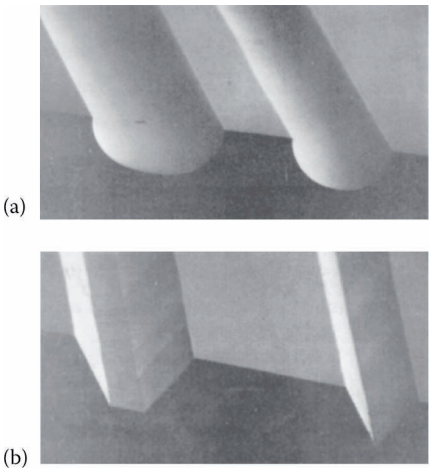


FIGURE 1.19 Isotropic and anisotropic etching profiles in single-crystal Si. Isotropic etching of grooves in (100) Si (a) and anisotropic etching of grooves in (100) Si (b) using rectangular mask openings. Features are in the 100-μm range.

Si strain gauges on a thin Si substrate with diffused resistors in 1961. As early as 1962, Tufte and coworkers at Honeywell, using a combination of wet isotropic etching, dry etching (using a plasma instead of a solution), and oxidation, made the first thin Si piezoresistive diaphragms for pressure sensors.[11] Isotropic etching of Si had been developed earlier for transistor fabrication. In the mid-1960s, Bell Labs started work on single-crystal silicon etchants with directional preferences, i.e., anisotropic etchants, such as mixtures of, at first, KOH, water, and alcohol and later KOH and water.[12] Both chemical and electrochemical anisotropic etching methods were pursued. The aspect ratio (height-to-width ratio) of features in MEMS is typically much higher than in ICs (Figure 1.18). The first high-aspect-ratio cuts in silicon were used in the fabrication of dielectrically isolated structures in ICs such as those for beam leads. In the mid-1970s, a surge of activity in anisotropic etching was associated with the work on V-groove and U-groove transistors. Isotropic and anisotropic etching profiles are compared in Figure 1.19. Figure 1.19a shows the isotropic etching of grooves in (100) Si, and Figure 1.19b shows the anisotropic etching of grooves in (100) Si. In both cases, rectangular mask openings are used.

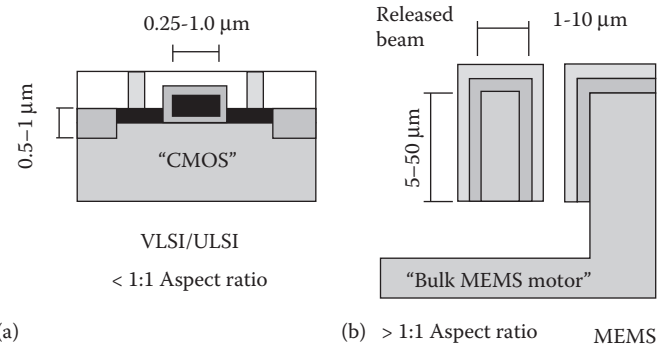


FIGURE 1.18 Aspect ratio (height-to-width ratio) typical in (a) fabrication of integrated circuits and (b) microfabricated component.

Most single-crystal silicon MEMS devices feature bonding of one Si wafer to another or to a differing substrate, say a glass pedestal, and some MEMS involve cavity-sealing techniques, perhaps for a vacuum reference or to accommodate a deflecting cantilever beam. The most prominent techniques developed to achieve these features are field-assisted bonding, invented by Wallis and Pomerantz in 1969,[13] and Si fusion bonding (SFB) by Shimbo in 1986.[14] Field-assisted thermal bonding, as shown in Figure 1.20, also known as *anodic* bonding, *electrostatic* bonding, or the *Mallory* process, is commonly used for joining glass to silicon at high temperatures (e.g., 400°C) and high voltages (e.g., 600 V). The ability to bond two Si wafers directly, at high temperatures (>800°C) in an oxidizing environment, without intermediate layers or applying an electric field, simplified the fabrication of many devices in silicon fusion bonding (SFB).

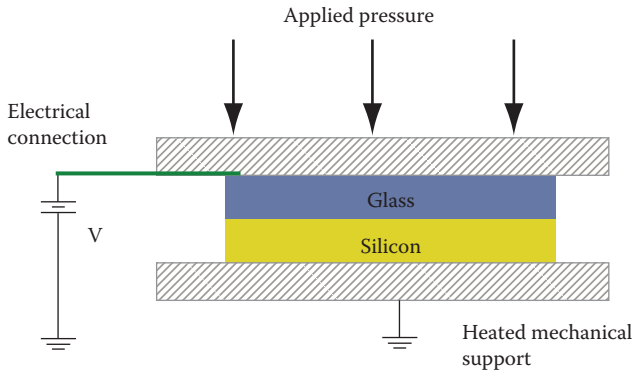


FIGURE 1.20 During anodic bonding, the negative potential applied to the borosilicate glass plate, which has been heated to 500°C, allows the migration of positive ions (mostly Na^+) away from the wafer's interface, creating a strong electric field between the glass and the Si wafer.

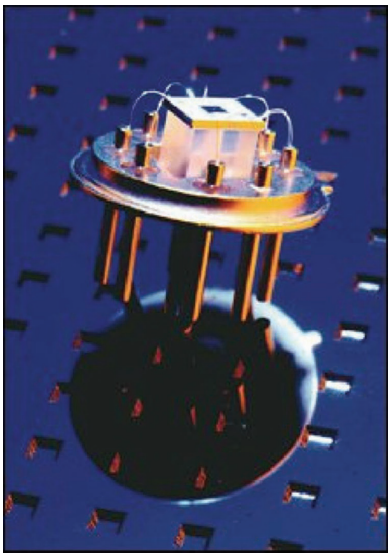


FIGURE 1.21 Piezoresistive pressure sensor featuring a Si/glass bond achieved by anodic bonding. The Bosch engine control manifold absolute pressure (MAP) sensor is used in automobile fuel injection systems. By measuring the manifold pressure, the amount of fuel injected into the engine cylinders can be calculated. Micromachined silicon piezoresistive pressure sensors are bonded at the wafer level to a glass wafer using anodic bonding before dicing. The glass pedestal that is created by this process provides stress isolation for the silicon sensor from package-induced thermal stresses. (Photo courtesy of Robert Bosch GmbH, Germany.)

The first Si accelerometer was demonstrated in 1970 at Kulite. In 1972, Sensym became the first company to make stand-alone Si sensor products. By 1974, National Semiconductor Corporation, in California, carried an extensive line of Si pressure transducers as part of the first complete silicon pressure transducer catalog.[15] Other early commercial suppliers of micromachined pressure sensor products were Foxboro/ICT, Endevco, Kulite, and Honeywell's Microswitch. To achieve better sensitivity and stability than possible with piezoresistive pressure sensors, capacitive pressure sensors were first developed and demonstrated by Dr. James Angell at Stanford University around 1977.[16] In Figure 1.21 we show a typical piezoresistive single-crystal silicon pressure sensor, with the silicon sensor anodically bonded to a glass substrate.

In many cases, it is desirable to stop the etching process when a certain cavity depth or a certain membrane thickness is reached. High-resolution silicon micromachining relies on the availability of effective etch-stop layers rather than the use of a stopwatch to control the etch depth. It was the discovery/development of impurity-based etch stops in silicon that allowed micromachining to become a high-yield commercial production process. The most widely used etch-stop technique is based on the fact that anisotropic etchants do not attack heavily boron-doped (p++) silicon layers. Selective p++ doping is typically implemented using gaseous or solid boron diffusion sources with a mask (such as silicon dioxide). The boron etch-stop effect was first noticed by Greenwood in 1969,[17] and Bohg in 1971[18] found that an impurity concentration of about $7 \times 10^{19}/cm^3$ resulted in the anisotropic etch rate of Si decreasing sharply.

Innovative, micromachined structures, different from the now mundane pressure sensors, accelerometers, and strain gauges, began to be explored by the mid- to late 1970s. Texas Instruments produced a thermal print head in 1977,[19] and IBM produced inkjet nozzle arrays the same year.[20] In 1980, Hewlett Packard made thermally isolated diode detectors,[21] and fiberoptic alignment structures were manufactured at Western Electric. Chemists worldwide took notice when Terry, Jerman, and Angell, from Stanford University, integrated a gas chromatograph on a Si wafer in 1979 as shown in Figure 1.22.[22,23] This first analytical chemistry application would eventually lead to the concept of total analytical systems on a chip, or μ-TAS. An important milestone in the MEMS world was the founding of NovaSensor,

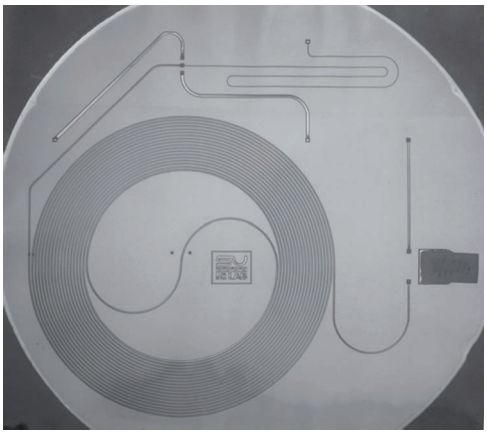


FIGURE 1.22 Gas chromatograph on a Si wafer. (Courtesy Hall Jerman.)

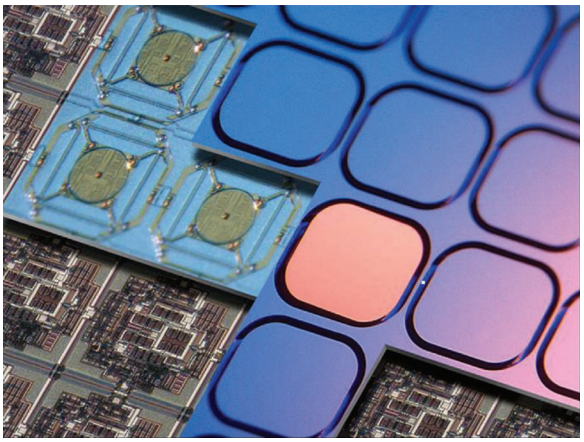


FIGURE 1.24 Integrated photonic mirror array from Transparent Networks. MEMS-VLSI integration achieved through wafer bonding. There are 1200 3D mirrors on the chip; each is 1 × 1 mm, with a ±10° tilt in two axes. (Courtesy of Janusz Bryzek.)

in 1985, by Kurt Petersen, Janusz Bryzek, and Joe Mallon (Figure 1.23). This was the first company totally dedicated to the design, manufacture, and marketing of MEMS sensors.

Kurt Petersen developed the first torsional, scanning micromirror in 1980 at IBM.[24] A more recent version of a movable mirror array is shown in Figure 1.24. Mirror arrays of this type led to the infamous stock market optical MEMS bubble of 2000, one of the bigger disappointments befalling the MEMS community.

The first disposable blood pressure transducer became available in 1982 from Foxboro/ICT, Honeywell for $40. Active on-chip signal conditioning also came of age around 1982. European and Japanese companies followed the U.S. lead more than a decade later; for example, Druck Ltd., in the United Kingdom, started exploiting Greenwood's micromachined pressure sensor in the mid-1980s. Petersen's 1982 paper extolling the excellent mechanical properties of single-crystalline silicon helped galvanize academia's involvement in Si micromachining in a major way.[25]

FIGURE 1.23 NovaSensor founders in 2003 at the Boston Transducer Meeting. Left to right: Joe Mallon, Kurt Petersen, and Janusz Bryzek.

In MEMS the need sometimes arises to build structures on both sides of a Si wafer; in this case, a double-sided alignment system is required. These systems started proliferating after the EV Group (formally known as Electronic Visions) created the world's first double-side mask aligner with bottom side microscope in 1985 (http://www.evgroup.com). In the mid-1990s, new high-density plasma etching equipment became available, enabling directional deep dry reactive ion etching (DRIE) of silicon. Dry plasma etching was now as fast as wet anisotropic etching, and as a consequence the MEMS field underwent a growth spurt.

U.S. government agencies started large MEMS programs beginning in 1993. Older MEMS researchers remember the idealistic and inspired leadership of Dr. Kaigham (Ken) Gabriel (Figure 1.25) at Defense Advanced Research Projects Agency (DARPA). Gabriel got many important new MEMS products launched.

When the first polysilicon MEMS devices, made in a process called surface micromachining pioneered at University of California, Berkeley by Muller

FIGURE 1.25 Dr. Kaigham Gabriel: Early champion of MEMS work at DARPA.

and Howe, appeared in 1983,[26] bulk single-crystal Si micromachining started to get some stiff competition. This was exacerbated when, from the mid-1990s onward, MEMS applications became biomedical and biotechnology oriented. The latter applications may involve inexpensive disposables or implants, and Si is not a preferred inexpensive substrate nor is it biocompatible. Glass and polymers became very important substrates in microfluidics, and many researchers started using a flexible rubber (polydimethylsiloxane or PDMS) as a building material in a process called soft lithography. The latter manufacturing method, which dramatically shortened the time between novel fluidic designs and their testing, was invented in the late 1990s by Harvard's Whitesides.[27]

In the early years of MEMS, it was often projected that the overall MEMS market would grow larger than that for ICs. This notion was based on the expectation of many more applications for the former than the latter. This market projection has not been fulfilled. Including nonsilicon devices such as read-write heads, one can claim a MEMS market of about 10% of the total IC market today. From Appendix 1B we learn that Si sensors and actuators amount to only 4% of IC sales.

MEMS never really constituted a paradigm shift away from the IC concept but rather a broadening of it: incorporating more diverse materials, higher aspect ratio structures, and a wider variety of uses in smaller and more fragmented applications. In almost all respects MEMS remained IC's poor cousin: using second-hand IC equipment, with less than 5% of IC sales, and no Nobel laureates to trumpet breakthrough new concepts. Today, the Si MEMS market prospects are looking much better as MEMS are finally penetrating mass consumer products from projectors, to game controllers, to portable computers to cameras, mobile phones, and iPods with MEMS digital micromirror devices (DMDs), oscillators, accelerometers, gyros, etc. Even MEMS foundries are now thriving inside and outside the United States. This new generation of MEMS products fits high-throughput production lines on large Si substrates and is succeeding in the marketplace. The IC world has started to absorb the Si-MEMS world.

Over the years, the many MEMS applications did lead to a plethora of MEMS acronyms, some of them perhaps coined by assistant professors trying to get tenure faster. Here are some attempts at 15 minutes of fame:

- BioMEMS = MEMS applied to the medical and biotechnology field
- Optical MEMS = mechanical objects + optical sources/detectors
- Power-MEMS
- C-MEMS (carbon MEMS for this author but ceramic MEMS for others)
- HI-MEMS = hybrid insect-microelectromechanical systems
- RF-MEMS = radiofrequency MEMS
- Cif-MEMS = CMOS IC Foundry MEMS
- COTS MEMS = commercial off-the-shelf MEMS
- MOEMS = microoptical electromechanical systems
- P-MEMS = polymer MEMS
- CEMS = cellular engineering microsystems
- HARMEMS = high-aspect-ratio MEMS

We do expect that there are many more MEMS applications yet to be realized and that MEMS will facilitate the handshake between the macro and nano world in nanoelectromechanical systems (NEMS).

In Table 1.2 we sketch our attempt at a Si/MEMS history line. Many more MEMS milestones than listed in the preceding text are captured here.

Table 1.2 MEMS History

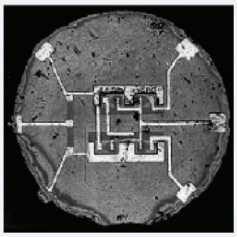
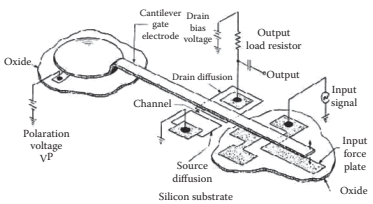
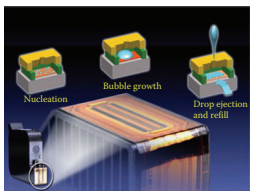
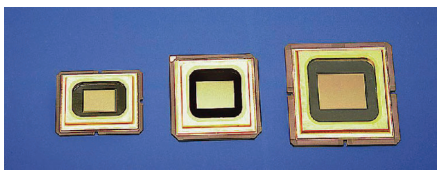
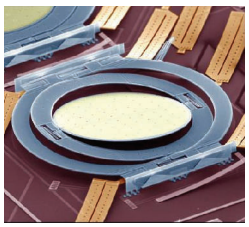
Year	Fact
1824	Berzelius discovers Si
1910	First patent on the MOS transistor concept
1927	Field effect transistor patented (Lilienfield)
1939	First pn junction transistor (J. Bardeen, W.H. Brattain, W. Shockley)
1947 (23 December)	Invention of the transistor made from germanium at Bell Telephone Laboratories
1954	Evidence of piezoresistive effect in Si and Ge by Smith[9]
1956	An early milestone for the use of single-crystal silicon in MEMS was the discovery of porous Si by Uhlir[8]
1958  1960s IC (1st Monolithic BJT IC, 4BJT)	Jack Kilby of Texas Instruments invents the IC, using GE devices. A patent was issued to Kilby in 1959. A few months later, Robert Noyce of Fairchild Semiconductor announced the development of a planar Si IC
1958	Silicon strain gauges commercially available
1958	First IC (oscillator)
1959	R. Feynman famous talk: "There's Plenty of Room at the Bottom"[7]
1961	Fabrication of the first piezoresistive sensor, pressure (Kulite)
1967	Anisotropic deep silicon etching (H.A. Waggener[12])
1967 	First surface micromachining process (H. Nathanson[28]): resonant gate before it was called MEMS
1969–1970	Anodic bonding of glass to Si[13]
1972	National Semiconductor: commercialize a Si MEMS pressure sensor
1975	Gas chromatograph on a Si wafer by S.C. Terry, J.H. Jerman, and J.B. Angell at Stanford University[23]
1977	First capacitive pressure sensor (Stanford)[16]
1977 	IBM–HP: micromachined ink-jet nozzle[24]
1980	K.E. Petersen, silicon torsional scanning mirror[24]

Table 1.2 MEMS History (Continued)

Year	Fact
1982	Review paper "Silicon as a mechanical material" published by K.E. Petersen[25]
1982	Disposable blood pressure transducer (Foxboro/ICT, Honeywell, $40)
1982	The use of x-ray lithography in combination with electroplating and molding (or LIGA), introduced by Ehrfeld and his colleagues[29]
1983	Integrated pressure sensor (Honeywell)
1983	"First" polysilicon MEMS device (Howe, Muller UCB[26]); see also Nathanson in 1967[28]
1986	Silicon to silicon wafer bonding (M. Shimbo[14])
1987	Texas Instrument's Larry Hornbeck invents the digital micromirror devices (DMDs)
1988	Rotary electrostatic side drive motors (Fan, Tai, Muller[30]) Electrostatic micromotor (UC-Berkeley BSAC)
1988	First MEMS conference (first transducers conference held in 1987)
1989 	Lateral comb drive (Tang, Nguyen, Howe[31])
1990	The concept of miniaturized total chemical analysis system or μ-TAS is introduced by Manz et al.[32] This may be seen as the beginning of BIOMEMS
1992	Grating light modulator (Solgaard, Sandejas, Bloom)
1992	First MUMPS process (MCNC) (with support of DARPA). Now owned by MEMSCAP
1992	First MEMS CAD tools: MIT, S.D. Senturia, MEMCAD 1.0 Michigan, Selden Crary, CAEMEMS 1.0
1992	Single-crystal reactive etching and metallization (SCREAM) developed at Process (Cornell)
1993	Analog devices: commercialize multiaxis accelerometer integrating electronics (ADXL50)
1995	Intellisense Inc. introduces MEMS CAD IntelliSuite. MEMCAD 2.0 is launched, and ISE introduces SOLIDIS and ICMAT
1996  TI's VGA (640 × 480), the SVGA (800 × 600), and the XGA (1024 × 768)	The first digital mirror device (DMD)–based products (Texas Instruments) appear on the market
1996	DRIE (Bosch Process)
1997	Printing meets lithography when George M. Whitesides et al. at Harvard discover soft lithography[27]
1998	First PCR-microchips
1998	Sandia's ultraplanar multilevel technology SUMMiT-IV and -V technologies. Four- and five-level poly-Si processes

(*continued*)

Table 1.2 MEMS History (Continued)

Year	Fact
1999 Affymetrix genechip	DNA microarray techniques
1999	Electrokinetic platforms (Caliper, Aclara, and Agilent)
2000	Nortel buys Xros for $3.25 billion
2002  Lucent 3D optical switch	The telecom recession puts many things on standby
2004	MEMS rebuilds. First application of accelerometer in consumer electronics (CE) to hard drive protection in notebooks. IBM puts dual-axis accelerometer in the notebook (now Lenovo)
2006	Sony (PS3) and Nintendo (Wii) introduce motion-based game controllers
2007	Apple announces the iPhone with motion-based features

NEMS

The criteria we use in this book for classifying something as a nanoelectromechanical system (NEMS) are not only that the miniaturized structures have at least one dimension that is smaller than 100 nanometers, but also that they are crafted with a novel technique (so beer making is out) or have been intentionally designed with a specific nanofeature in mind (so medieval church stained glass is out). This definition fits well within the one adopted by the National Nanotechnology Institute (NNI; http://www.nano.gov/html/facts/whatisnano.html):

1. Nanotechnology involves R&D at the 1- to 100-nm range.
2. Nanotechnology creates and uses structures that have novel size-based properties.
3. Nanotechnology takes advantage of the ability to control or manipulate at the atomic scale.

Paul Davis (http://cosmos.asu.edu), a theoretical physicist and well-known science popularizer, said, "The nineteenth century was known as the machine age, the twentieth century will go down in history as the information age and I believe the twenty-first century will be the quantum age." We believe that the current nanotechnology revolution, underpinned by quantum mechanics, is already leading the way toward that reality.

The manufacture of devices with dimensions between 1 and 100 nanometers is either based on top-down manufacturing methods (starting from bigger building blocks, say a whole Si wafer, and chiseling them into smaller and smaller pieces by cutting, etching, and slicing), or it is based on bottom-up manufacturing methods (in which small particles such as atoms, molecules, and atom clusters are added for the construction of bigger functional constructs). The top-down approach to nanotechnology we call *nanofabrication* or *nanomachining*, an extension of

the MEMS approach. The bottom-up approach we like to refer to as *nanochemistry*. An example of this second approach is the self-assembly of a monolayer (SAM) from individual molecules on a gold surface. Bottom-up methods are nature's way of growing materials and organisms, and in biomimetics one studies how nature, through eons of time, developed manufacturing methods, materials, structures, and intelligence and tries to mimic or replicate what nature does in the laboratory to produce MEMS or NEMS structures.

A history-line with the most important NEMS milestones on it is difficult to put together as so many authors of such charts automatically include themselves or their institution on it first (one author of an early MEMS/NEMS timeline puts himself on it three times and his institution four times!). It sometimes seems that science and engineering are starting to resemble FOX News more by the day. What follows are some milestones toward nanotechnology that many scientists/engineers might agree to.

Norio Taniguchi introduced the term *nanotechnology* in 1974, in the context of traditional machining with tolerances below 1 micron. The 1959 Feynman lecture "There's Plenty of Room at the Bottom," which helped launch the MEMS field (see above), was geared more toward NEMS than MEMS (http://www.its.caltech.edu/~feynman/plenty.html).[7] Feynman proclaimed that he knew of no principles of physics that would prevent the direct manipulating of individual atoms. In his top-down *gedanken* experiment, he envisioned a series of machines each an exact duplicate, only smaller and smaller, with the smallest in the series being able to manipulate individual atoms (see Figure 1.26).

In 1981, Gerd Binning and Heinrich Rohrer of IBM Zurich invented the scanning tunneling microscope (STM), enabling scientists to see and move individual atoms. Such a microscope, shown in Figure 1.27, measures the amount of electrical current flowing between a scanning tip and the conductive surface that is being measured. This unexpectedly simple instrument allowed for the imaging of micro- and nanostructures, catapulted nanotechnology onto the world stage, and got its inventors the 1986 Nobel Prize (http://www.zurich.ibm.com/imagegallery/st/nobelprizes). Just as 350 years before the microscope changed the way we

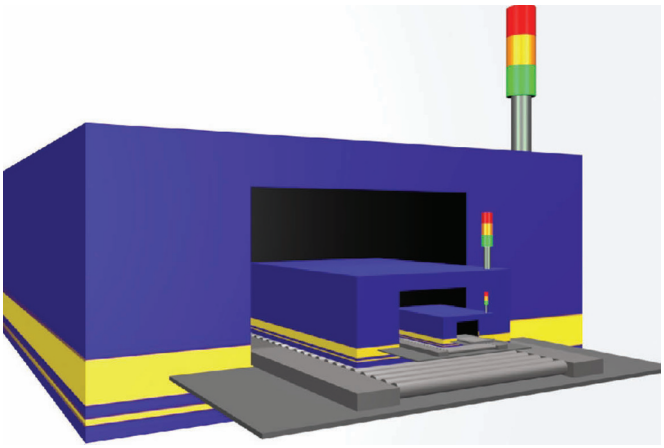


FIGURE 1.26 Master and slave hands on a set of Feynman machines.

viewed the world, the STM impacts our current view of biology, chemistry, and physics.

In fast succession, a series of similar instruments, all called scanning proximal probes, followed the introduction of the STM. For example, Binnig, Quate (Stanford), and Gerber (IBM) developed the atomic force microscope (AFM) in 1986. An AFM, in contact mode, measures the repulsive force interaction between the electron clouds on the probe tip atoms and those on the sample—making it possible to image both insulating and conducting surfaces. This results in the visualization of the interactions

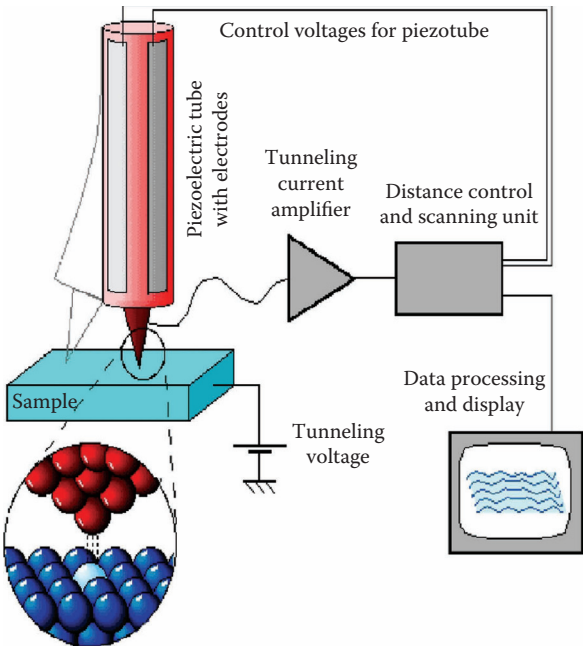


FIGURE 1.27 Scanning tunneling microscope (STM). Operational principle of an STM. (Courtesy Michael Schmidt, TU Wien.)

between molecules at the nanoscale, thus increasing our ability to better understand the mechanism of molecular and biological processes. Other forms of scanning probe microscopes—those that do not depend on tunneling or forces between a probe tip and a sample surface—have also been demonstrated. Examples include the scanning thermal microscope, which responds to local thermal properties of surfaces, the scanning capacitance microscope for dopant profiling, and the near-field scanning optical microscope (NSOM, also known as SNOM). In the latter instrument,[33] the wavelength limitation of the usual far-field optics of a light microscope is avoided by mounting the light detector (say an optical fiber) on an AFM tip, at a distance from the sample that is a fraction of the wavelength used; this way it is possible to increase the resolution of a light microscope considerably. These new tools were an important catalyst behind the surge in nanotechnology activities worldwide, and this illustrates that progress in science is inextricably linked to the development of new measurement tools.

The discovery, in the early 1980s at Bell Labs by David L. Allara (now at Pennsylvania State University) and Ralph G. Nuzzo (now at University of Illinois, Urbana-Champaign) of the self-assembly of disulfide and, soon thereafter, of alkanethiol monolayers (SAMs) on metal surfaces coincided with the maturation of STM technology.[34,35] SAMs, especially on Au, turned out to be a valuable type of sample for STM investigation, showing these films to spontaneously assemble into stable and highly organized molecular layers, bonding with the sulfur atoms onto the gold and resulting in a new surface with properties determined by the alkane head group.

Like SAMs, dendrimers, which are branching polymers sprouting successive generation of branches off like a tree, are an important tool for bottom-up nanotechnologists. Dendrimers (from the word *dendron*, Greek for tree) were invented, named, and patented by Dr. Donald Tomalia (now CTO at Dentritic NanoTechnologies, Inc.) in 1980 while at Dow Chemical.[36]

In 1970, Arthur Ashkin was the first to report on the detection of optical scattering and gradient forces on micron-sized particles.[37] In 1986, Ashkin and colleagues reported the first observation of what is now commonly referred to as an optical trap, i.e., a tightly focused beam of light capable of holding microscopic particles stable in three dimensions.[38] One of the authors of this seminal 1986 paper, Steven Chu (Figure 1.28), would go on to make use of optical tweezing techniques in his work on cooling and trapping of atoms. Where Ashkin was able to trap larger particles (10 to 10,000 nanometers in diameter), Chu extended these techniques to the trapping of individual atoms (0.1 nanometer in diameter). This research earned him the 1997 Nobel Prize in Physics (with Claude-Cohen Tannoudji and William D. Phillips). In another heralded experiment, Steven Chu was also the one who demonstrated that by attaching polystyrene beads to the ends of DNA one can pull on the beads with laser tweezers to stretch the DNA molecule (http://www.stanford.edu/group/chugroup/steve_personal.html).

FIGURE 1.28 Steven Chu (Stanford University), recipient of the 1997 Nobel Prize in Physics for his work on cooling and trapping of atoms.

In 1985, Robert F. Curl Jr., Harold W. Kroto, and the late Richard E. Smalley serendipitously (while investigating the outer atmosphere of stars) discovered a new form of carbon: buckminsterfullerene, also known as buckyball or C60, shown in Figure 1.29.[39] They were awarded the Nobel Prize in 1996.

Perhaps a more important discovery, because of its generality and broader applicability, is the one by NEC's Sumio Iijima, who, in 1991, discovered

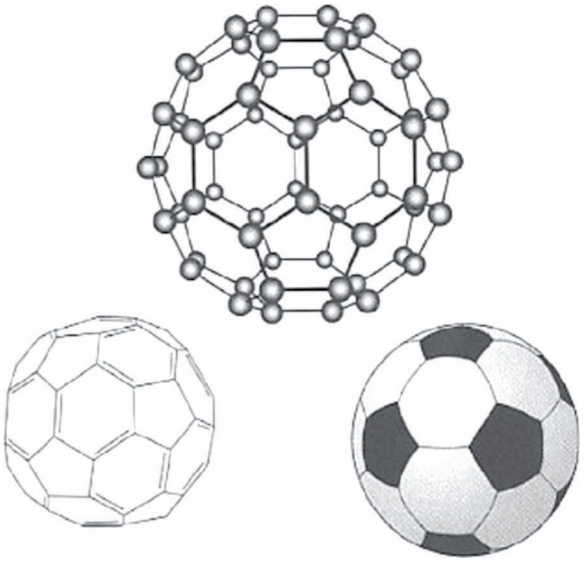


FIGURE 1.29 Buckminsterfullerene C60 or buckyball with 60 atoms of carbon; each is bound to three other carbons in an alternating arrangement of pentagons and hexagons.

carbon nanotubes, with an electrical conductivity that is up to six orders of magnitude higher than that of copper (http://www.nec.co.jp/rd/Eng/innovative/E1/myself.html). Like buckyballs, cylindrical nanotubes each constitute a lattice of carbon atoms, and each atom is again covalently bonded to three other carbons.

Carbon nanotubes exist as single-walled (SWNT) and multiwalled (MWNT); the ones depicted in Figure 1.30a are multiwall nanotubes. Unique among the elements, carbon can bond to itself to form extremely strong two-dimensional sheets, as it does in graphite, as well as buckyballs and nanotubes.

Cees Dekker demonstrated the first carbon nanotube transistor in 1998 at the Delft University of Technology[40] (see Figure 1.30b). In this device, a semiconducting carbon nanotube of only about 1 nm in diameter bridges two closely separated metal electrodes (400 nm apart) atop a silicon surface coated with silicon dioxide. Applying an electric field to the silicon (via a gate electrode) turns on and off the flow of current across the nanotube by controlling the movement of charge carriers in it. By carefully controlling the formation of metal gate electrodes, Dekker's group (http://www.ceesdekker.net) was able to create transistors with an output signal 10 times stronger than the input. At around the same time, the first nanotransistor was built at Lucent Technologies (1997). The MOS semiconductor transistor was 60 nm wide, including the source, drain, and gate; its thickness was only 1.2 nm. Other companies have since built smaller nanotransistors.

At one point, in the late 1990s, it was—with just a bit of exaggeration—hard to find a proposal to a government agency that did not involve carbon nanotubes. But perhaps more real progress was mapped in the meantime in the area of nanocrystals or quantum dots (QD).

Quantum dots possess atom-like energy states. The behavior of such small particles was beginning to be understood with work by the Russians Ekimov and Efros from 1980 to 1982.[41,42] They recognized that nanometer-sized particles of CdSe, with their very high surface-to-volume ratio, could

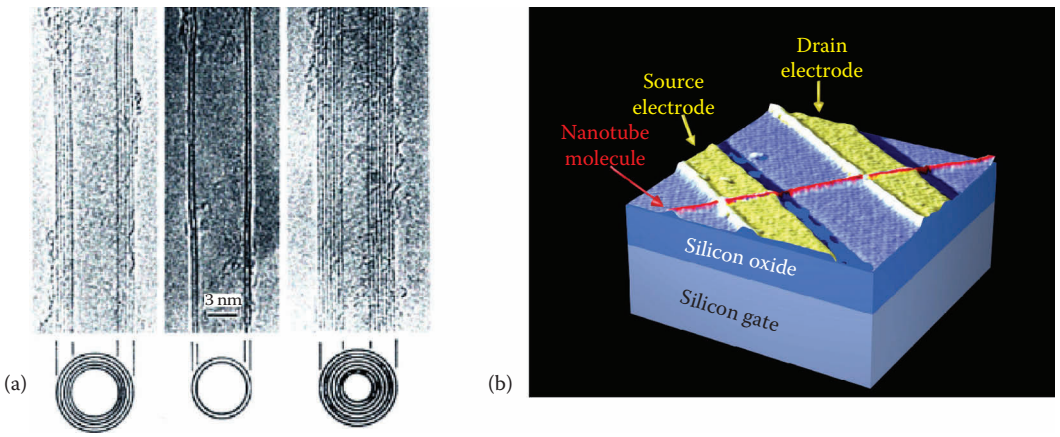


FIGURE 1.30 (a) Multiwall nanotubes with Russian inset doll structure; several inner shells are shown a typical radius of the outermost shell >10 nm. (From S. Iijima, *Nature* 354, 54, 1991.) (b) First nanotube transistor. This three-terminal device consists of an individual semiconducting nanotube on two metal nanoelectrodes with the substrate as a gate electrode.

trap electrons and that these trapped electrons might affect the crystal's response to electromagnetic fields, that is, absorption, reflection, refraction, and emission of light. Louis E. Brus, a physical chemist at Bell Laboratories at the time, and now at Columbia University, put this to practice when he learned to grow CdSe nanocrystals in a controlled manner.[43,44] Murray, Norris, and Bawendi synthesized the first high-quality quantum dots in 1993.[45] Crystallites from –12 to –115 Å in diameter with consistent crystal structure, surface derivatization, and a high degree of monodispersity were prepared in a single reaction based on the pyrolysis of organometallic reagents by injection into a hot coordinating solvent. The confinement of the wave-functions in a nanocrystal or quantum dot lead to a blue energy shift, and by varying the particle size one can produce any color in the visible spectrum, from deep (almost infra-) reds to screaming (almost ultra-) violet as illustrated in Figure 1.31. Today, quantum dots form an important alternative to organic dye molecules. Unlike fluorescent dyes, which tend to decompose and lose their ability to fluoresce, quantum dots maintain their integrity, withstanding many more cycles of excitation and light emission (they do not bleach as easily!). Combining a number of quantum dots in a bead conjugated to a biomolecule is used as a spectroscopic signature—like a barcode on a commercial product—for tagging those biomolecules.

Carbon nanotubes are only one type of nanowire. In terms of investigating and exploiting quantum confinement effects, semiconductor wires, with diameters in the 10s of nanometers, often single crystalline, represent the smallest dimension for efficient transport of electrons and excitons and are the logical interconnects and critical devices for nanoelectronics and nanooptoelectronics of the future. Over the past decade, there has been major progress in chemical synthesis technologies for growing these nanoscale semiconductor wires. As originally proposed by R.S. Wagner and W.C. Ellis from Bell Labs for the Au-catalyzed Si whisker growth, a vapor-liquid-solid mechanism is still mostly used.[46] But the field got a shot in the arm (a rebirth so to speak) with efforts by Charles Lieber (Harvard), Peidong Yang (http://www.cchem.berkeley.edu/pdygrp/main.html), James Heath (http://www.its.caltech.edu/~heathgrp), and Hongkun Park (http://www.people.fas.harvard.edu/~hpark). Lieber's group at Harvard (http://cmliris.harvard.edu) reported arranging indium phosphide semiconducting nanowires into a simple configuration that resembled the lines in a tick-tack-toe board. The team used electron beam lithography to place electrical contacts at the ends of the nanowires to show that the array was electronically active. The tiny arrangement was not a circuit yet, but it was the first step, showing that separate nanowires could communicate with one another.

Molecules are 30,000 times smaller than a transistor (180 nm on a side), so obviously it is of some use to investigate whether molecules can act as switches. Mark Ratner and Ari Aviram had suggested this as far back as 1974.[47] The suggestion remained a pipe dream until the advent of scanning probe microscopes in the 1980s, which gave researchers finally

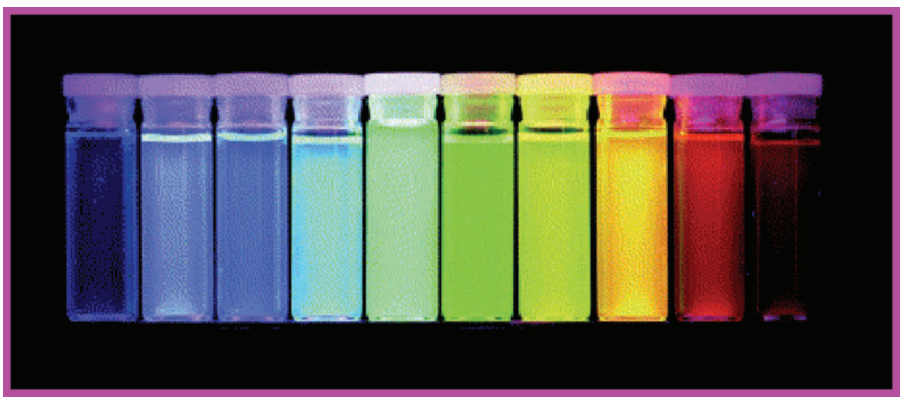


FIGURE 1.31 Different-sized quantum dots in response to near-UV light. Also composition of core affects wavelength. Red: bigger dots! Blue: smaller dots!

the tools to probe and move individual molecules around. This led to a large number of studies in the late 1990s that demonstrated that individual molecules can conduct electricity just like metal wires, and turning individual molecules into switches came not far behind. In 1997, groups led by Robert Metzger (http://bama.ua.edu/~rmmgroup) of the University of Alabama, Tuscaloosa, and Mark Reed of Yale University (http://www.eng.yale.edu/reedlab) created molecular diodes. In July 1999, another group headed by James Heath and Fraser Stoddart of the University of California, Los Angeles (UCLA) (http://stoddart.chem.ucla.edu) also created a rudimentary molecular switch, a molecular structure that carries current but, when hit with the right voltage, alters its molecular shape and stops conducting. Heath's team placed molecules called rotaxanes, which function as molecular switches, at each junction of a circuit. By controlling the input voltages the scientists showed that they could make 16-bit memory circuits work. The field of moletronics was born.

As shown in Figure 1.32, rotaxanes are "mechanically linked" molecules that consist of a dumbbell-shaped molecule, with a cyclic molecule linked around it between the two ends. The two ends of the dumbbell molecule are very big and prevent the cyclic molecule from slipping off the end. A number of factors (e.g., charge, light, pH) can influence the position of the cyclic molecule on the dumbbell.

The use of x-ray lithography in combination with electroplating and molding (or LIGA), introduced by Ehrfeld and his colleagues in 1982,[29] demonstrated to the world that lithography may be merged with more traditional manufacturing processes to make master molds of unprecedented aspect ratios and tolerances to replicate microstructures in ceramics, plastics, and metals. The hard x-rays used enable nano-sized patterns to be printed.

At around 1997, Whitesides et al. introduced soft lithography, including the use of pattern transfer of self-assembled monolayers (SAMs) by elastomeric stamping.[27] This technique formed a bridge between top-down and bottom-up machining; a master mold is made based on "traditional" lithography, and the stamp generated from this master is inked with SAMs to print (stamp) substrates with nano-sized patterns.

Imposing boundaries on photons, by making them move in a material with a periodic dielectric constant in one, two, or three directions, leads to photonic crystals. Photonic crystals were first studied by Lord Rayleigh in 1887, in connection with the peculiar reflective properties of a crystalline mineral with periodic "twinning" planes.[48] He identified a narrow bandgap prohibiting light propagation through the planes. This bandgap was angle-dependent because of the differing periodicities experienced by light propagating at non-normal incidences, producing a reflected color that varies sharply with angle. A similar effect is seen in nature, such as in butterfly wings (Figure 1.33) and abalone shells.

A one-dimensional periodic structure, such as a multilayer film (a Bragg mirror), is the simplest type of photonic crystal, and Lord Rayleigh showed that any such one-dimensional system has a bandgap. The possibility of two- and three-dimensionally periodic crystals with corresponding two- and three-dimensional bandgaps was suggested 100

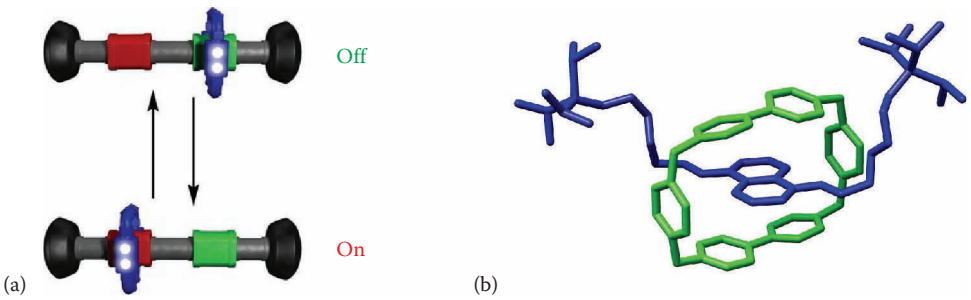


FIGURE 1.32 (a) A diagram of rotaxane. The usefulness of rotaxanes is because there are a number of positions along the dumbbell molecule that the cyclic molecule can attach to temporarily. The dumbbell can be thought of as a train track, with the positions on the dumbbell molecule as stations and the cyclic molecule as the train. (b) Crystal structure of rotaxane with a cyclobis(paraquat-p-phenylene) macrocycle.

FIGURE 1.33 A full-grown *Morpho rhetenor* butterfly, a native to South America. (From University of Southhampton. Color by nanostructure instead of dyes.)

years after Rayleigh by Eli Yablonovitch[49] (http://www.ee.ucla.edu/labs/photon/homepage.html) and Sajeev John[50] (http://www.physics.utoronto.ca/~john) in 1987. Yablonovitch (in 1991)[51] demonstrated the first microwave photonic bandgap (PBG) structure experimentally with 1-mm holes drilled in a dielectric material as illustrated in Figure 1.34 and known today as *yablonovite*. Since then, several research groups verified this prediction, which has ignited a worldwide rush to build tiny "chips" that control light beams instead of electron streams. In photonic crystals the repeat unit in the lattice is of the same size as the incoming wavelength, so homogeneous (effective) media theory cannot be applied. Photonic crystals feature lattice spacings ranging from the macroscopic (say 1 mm, for operating in the microwave domain-like yablonovite) to the 100s of nanometer range (to operate in the visible range). We cover photonic crystals here under NEMS, although only photonic crystals for the visible range qualify as nanotechnology. The potential applications of photonics are limitless, not only as a tool for controlling quantum optical systems but also in more efficient miniature lasers for displays and telecommunications, in solar cells, LEDs, optical fibers, nanoscopic lasers, ultrawhite pigments, radiofrequency antennas and reflectors, photonic integrated circuits, etc.

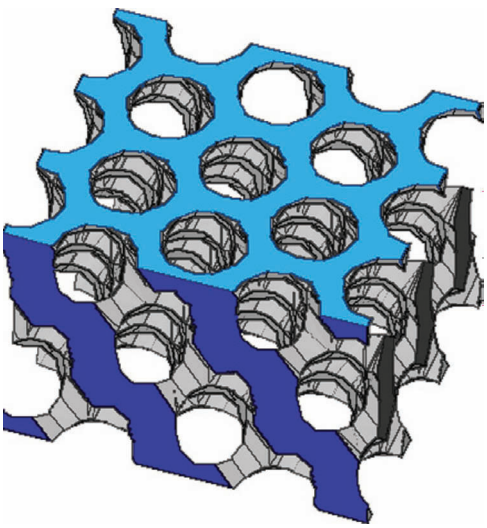


FIGURE 1.34 Holes drilled in dielectric: known now as *yablonovite*, after Yablonivitch (http://www.ee.ucla.edu/~pbmuri). The holes are 1 mm in size, and this photonic crystal is meant to operate in the microwave range.

In 1967 Victor Veselago, a Russian physicist, predicted that composite metamaterials might be engineered with negative magnetic permeability and negative permittivity.[52] Metamaterials are artificially engineered materials possessing properties that are not encountered in nature. Whereas photonic materials do exist in nature, metamaterials do not; moreover, in the case of metamaterials, the building blocks are small compared with the incoming wavelength so that effective media theory can be applied. In conventional materials the plane of the electrical field, the plane of the magnetic field, and the direction in which light travels are all oriented at right angles to each other and obey the right-hand rule. In Veselago's imaginary metamaterial, the above quantities obey a left-hand rule (as if they were reflected in a mirror). These materials would interact with their environment in exactly the opposite way from natural materials (see negative refractive index water in Figure 1.35). One intriguing prediction was that the left-hand rule would allow for a flat superlens to focus light to a point and that could image with a resolution far beyond the diffraction limit associated with far-field illumination. Veselago's prediction that such perfect lenses could be made from metamaterials lay dormant until 1996–2000, when the remarkable John Pendry, a physicist at Imperial College in London, showed that certain metals could be engineered to respond to electric fields as though the field parameters were negative.[53–58] In 2001, researchers at Imperial College and Marconi Caswell Ltd. (London) announced a magnetic resonance

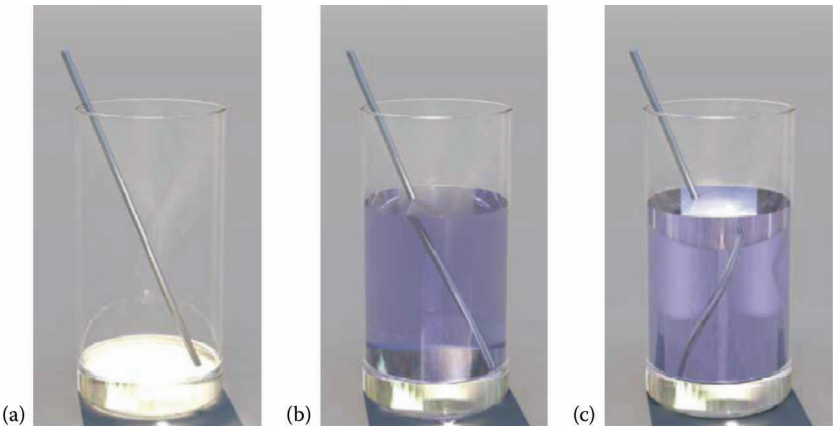


FIGURE 1.35 Refraction illustrated (a) empty glass: no refraction; (b) typical refraction with pencil in water with $n = 1.3$; (c) what would happen if the refractive index were negative with $n = -1.3$ (see metamaterials). (From Gennady Shvets, The University of Texas at Austin; http://www.ph.utexas.edu/~shvetsgr/lens.html.)

imaging system using a magnetic metamaterial based on Pendry's design.[59] Physicist Richard Shelby's group at the University of California, San Diego demonstrated a left-handed composite metamaterial that exhibited a negative index of refraction for microwaves.[60] The simple arrangement consisted of a planar pattern of copper split-ring resonators (SRRs) and wires on a thin fiberglass circuit board. Operating in the microwave range these metal patterns are large (5 mm repeat unit) but progress toward metamaterials operating in the visible was very swift. By 2005, Zhang's group at University of California, Berkeley made a 35-nm thick Ag superlens and imaged objects as small as 40 nm with 365 nm light, clearly breaking the diffraction limit of far-field imaging[61] (find Zhang's group at http://xlab.me.berkeley.edu). By 2007, a left-handed material operating in the visible range (780 nm) was demonstrated[62] by Soukoulis (http://cmpweb.ameslab.gov/personnel/soukoulis) at the U.S. Department of Energy's Ames Laboratory on the Iowa State University campus and Wegener's group from the University of Karlsruhe (http://www.aph.uni-karlsruhe.de/wegener), Germany.

As in the case of photonic crystals, only the metamaterials operating in the visible qualify as nanotechnology, but for comprehensiveness sake we cover all of them together here.

In 2000, IBM scientists placed a magnetic cobalt atom inside an elliptical coral of atoms. They observed the Kondo effect, i.e., electrons near the atom align with the atom's magnetic moment, effectively canceling it out. When the atom was placed at one focus of the elliptical coral, a second Kondo effect was observed at the other focus, even though no atom was there (see Figure 1.36). Hence some of the properties (info) carried by an atom are transferred to the other focus (www.research.ibm.com). This quantum mirage effect "reflects" information using the wave nature of the electrons rather than transmission of info using electrons in a wire. It has the potential to be able to transfer data within future nanoscale electronic circuits where wires would not work. This would allow miniaturization of circuits well below what is envisioned today.

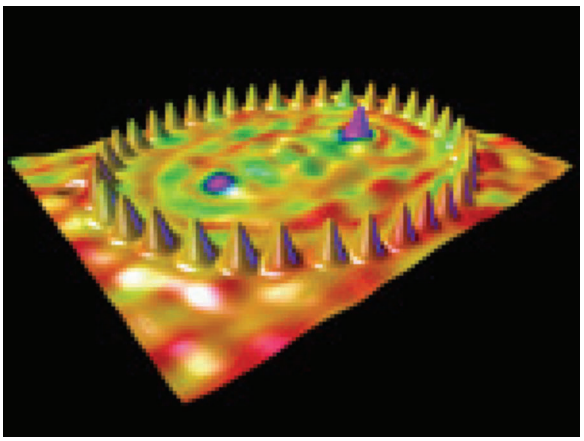


FIGURE 1.36 Quantum mirage phenomenon (http://domino.research.ibm.com/comm/pr.nsf/pages/rsc.quantummirage.html).

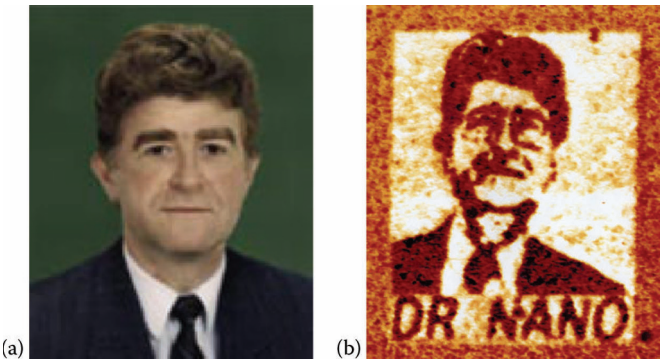


FIGURE 1.37 NSF's Dr. Mike Roco, a photo (a) and a nanograph (b). The nanograph of Dr. M. Roco was recorded at Oak Ridge National Laboratory using piezoresponse force microscopy, one of the members of the family of techniques known as scanning probe microscopy, which can image and manipulate materials on the nanoscale. Each picture element is approximately 50 nanometers in diameter; the distance from chin to eyebrow is approximately 2.5 microns. (Courtesy Dr. Roco.)

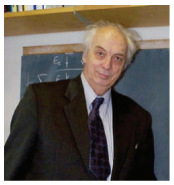
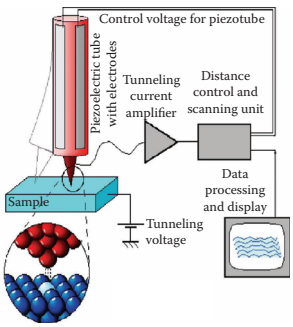
In 1999 President Clinton announced the National Nanotechnology Initiative (NNI); this first formal government program for nanotechnology accelerated the pace of nano research (the program had been around unofficially since 1996). In December 2003, George W. Bush signed the 21st Century Nanotechnology Research and Development Act. In this government NEMS program, Dr. Mike Roco (Figure 1.37) played a similar catalyzing role that Dr. Ken Gabriel played earlier in the MEMS field (see above).

Two other important nanotech promoters are Ray Kurzweil (http://www.kurzweilai.net/index.html) and K. Eric Drexel (http://www.foresight.org). Whereas Feynman continues to receive almost universal praise for his inspiring 1959 speculative talk, K. Eric Drexel, who in 1981 described mechanochemistry in his speculative paper "Molecular Engineering: An Approach to the Development of General Capabilities for Molecular Manipulation," continues to receive mostly harsh criticism—sometimes bordering on derision. In this paper and in two books,[63,64] Drexel builds nanotechnology, bottom-up, atom by atom, rather than whittling down materials as Feynman had suggested. Drexel also makes more of the fact that nature and molecular biology are proof of concept for this type of molecular technology. Drexel's early warnings about "gray goo," his emphasis on assemblers—small machines that would guide chemical bonding operations by manipulating reactive molecules—and building nanotechnology in a dry environment probably explain most of the hostility toward his work (even by those who do not even make an attempt to understand it; see for example Atkinson[65] in Nanocosm—we might as well listen to Newt Gingrich talk about nanotechnology). Drexel, unfortunately, has been associated too much with the nano pop culture. *Nano!* by Ed Regisis is an engaging and entertaining book that describes some of the researchers involved in nanotechnology; he is uncharacteristically positive about Drexler.[66]

Market projections on NEMS are today even wilder than the early ones on MEMS. Overall, though, this author believes that it is in nanotechnology, especially when considering bottom-up manufacturing, that a paradigm shift away from IC-type manufacturing is taking shape, and that it is nanotechnology that holds the potential of having a much larger impact on society than IC technology ever did. Indeed, the nanoscale is unique because at this length scale important material properties, such as electronic conductivity, hardness, or melting point, start depending on the size of the chunk of material in a way they do not at any other scale. Moreover, in biotechnology, molecular engineering already has made major progress, and the confluence of miniaturization science and molecular engineering is perhaps the most powerful new avenue for progress of humankind in general. In this regard, NEMS must be seen as a support technology to extract yet more benefits from the ongoing biotechnology revolution.

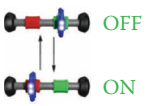
In Table 1.3 we show a milestone chart in nanotechnology. It can be argued that molecular scientists and genetic engineers were practicing nanotechnology long before the name became popular with electrical and mechanical engineers. Molecular biology or "wet nanotechnology" has been called "nanotechnology that works." But adding breakthroughs pertaining to molecular biology to this table would make it much too long. For the same reason we also did not list any IC-related milestones.

Table 1.3 A Milestone Chart in Nanotechnology

3.5 billion years ago	The first living cells emerge. Cells house nanoscale biomachines that perform such tasks as manipulating genetic material and supplying energy
400 BC	Democritus coins the word "atom," which means "not cleavable" in Greek
1857	Michael Faraday introduces "colloidal gold" to the Royal Society
1887	Photonic crystals are studied by Lord Rayleigh, in connection with the peculiar reflective properties of a crystalline mineral with periodic "twinning" planes[48]
1905	Albert Einstein publishes a paper that estimates the diameter of a sugar molecule as 1 nm. Jean-Baptist Perrin confirmed these results experimentally and was awarded the 1926 Nobel Prize for this work
1931	Max Knoll and Ernst Ruska develop the electron microscope, which enables nanometer imaging
1932	Langmuir establishes the existence of monolayers (Nobel Prize in 1932)
1959	Richard Feynman gives his famed talk "There's Plenty of Room at the Bottom," on the prospects for miniaturization[7]
1967  Victor Veselago	Victor Veselago, a Russian physicist, predicted that composite metamaterials might be engineered with negative magnetic permeability and negative permittivity[52]
1968	Alfred Y. Cho and John Arthur of Bell Laboratories and their colleagues invent molecular-beam epitaxy, a technique that can deposit single atomic layers on a surface
Early 1970s	Groups at Bell Laboratories and IBM fabricate the first two-dimensional quantum wells
1974	Norio Taniguchi conceives the word *nanotechnology* to signify machining with tolerances of less than a micron
1974	Mark Ratner and Ari Aviram suggest using molecules as switches[47]
1980	The behavior of quantum dots began to be understood with work by the Russians Ekimov and Efros in 1980–1982.[41,42] Louis E. Brus learned to grow CdSe nanocrystals in a controlled manner[43,44]
1980	Dendrimers (from the word *dendron*, Greek for tree) were invented, named, and patented by Dr. Donald Tomalia[36]
1981 	Gerd Binnig and Heinrich Rohrer create the scanning tunneling microscope, which can image individual atoms
Early 1980s	The discovery, in the early 1980s, by David L. Allara and Ralph G. Nuzzo of the self-assembly of disulfide and, soon thereafter, of alkanethiol monolayers (SAMs) on metal surfaces
1982	The use of x-ray lithography in combination with electroplating and molding (or LIGA) is introduced by Ehrfeld and his colleagues[29]
1984	Pohl develops near-field scanning optical microscope (NSOM, also known as SNOM)[33]
1985	Robert F. Curl, Jr., Harold W. Kroto, and Richard E. Smalley discover buckminsterfullerenes, also known as buckyballs, which measure about a nanometer in diameter[39]

(*continued*)

Table 1.3 A Milestone Chart in Nanotechnology (Continued)

1986	Ashkin and colleagues reported the first observation of what is now commonly referred to as an optical trap, i.e., a tightly focused beam of light capable of holding microscopic particles stable in three dimensions[38]
1986	K. Eric Drexel publishes *Engines of Creation*, a futuristic book that popularizes nanotechnology
1987	The possibility of two- and three-dimensionally periodic crystals with corresponding two- and three-dimensional bandgaps was suggested 100 years after Rayleigh, by Eli Yablonovitch[49] and Sajeev John[50]
1989	Donald M. Eigler of IBM writes the letters of his company's name using 35 individual xenon atoms on a nickel surface (in high vacuum and at liquid helium temperatures)
1991	Yablonovitch[52] demonstrates the first microwave photonic bandgap (PBG) structure experimentally (holes drilled in dielectric), known now as *yablonovite*
1991	Sumio Iijima of NEC in Tsukuba, Japan, discovers carbon nanotubes. The first single-walled nanotubes (SWNT) were produced in 1993
1993	The first high-quality quantum dots are synthesized by Murray, Norris, and Bawendi[45,67]
1993	Warren Robinett of the University of North Carolina and R. Stanley Williams of UCLA devise a virtual reality system connected to an STM that lets users see and touch atoms
1997	The first complete metal oxide semiconductor transistor (60 nm wide) is invented by Lucent Technologies. The key breakthrough was the 1.2-nm-thick gate oxide
1997	Whitesides et al.[27] introduced soft lithography, including the use of pattern transfer of self-assembled monolayers (SAMs) by elastomeric stamping
1998	Cees Dekker's group at the Delft University of Technology in the Netherlands creates a transistor from a carbon nanotube[40]
1999 	James Heath and Fraser Stoddart of UCLA (http://stoddart.chem.ucla.edu/) create rudimentary molecular switches, with molecules called rotaxanes, which function as molecular switches
1999	James M. Tour, now at Rice University, and Mark A. Reed of Yale University demonstrate that single molecules can act as molecular switches[68]
1999	The Clinton administration announces the National Nanotechnology Initiative, which provides a big boost in funding and gives the field greater visibility
1999	Thermomechanical memory device, unofficially known as "Millipede," first demonstrated at IBM Zurich
2000	Eigler and other IBM scientists devise a quantum mirage—placing a magnetic atom at the focus of an elliptical ring of atoms creates a mirage atom at the other focus—transmitting info without wires
2000	John Pendry, a physicist at Imperial College in London, showed that certain metals could be engineered to respond to electric fields as though the field parameters were negative[55]
2001	Researchers at Imperial College and Marconi Caswell (London) announced a magnetic resonance imaging system using a magnetic metamaterial based on Pendry's design[59]
2004	Physicists at the University of Manchester make graphene sheets[69]
2005	Zhang et al. demonstrate the near-field superlens—imaging objects in the tens of nanometer range with 365 nm light[61]
2007	The first left-handed material in the visible range[62]

Acknowledgments

Special thanks to Xavier Casadevall i Solvas and Drs. Sylvia Daunert and Benjamin Park.

Appendix 1A: International Technology Roadmap for Semiconductors (ITRS)

The complete 2003 ITRS and past editions of the ITRS editions are available for viewing and printing as an electronic document at http://public.itrs.net. The International Technology Roadmap for Semiconductors (ITRS) predicts the main trends in the semiconductor industry spanning across 15 years into the future. The participation of experts from Europe, Japan, Korea, and Taiwan as well as the United States ensures that the ITRS is a valid source of guidance for the semiconductor industry as it strives to extend the historical advancement of semiconductor technology and the worldwide integrated circuit (IC) market. The 2003 ITRS edition, used as the source for the tables below, extends to the year 2018. The 2003 ITRS does not predict a further acceleration in the timing of introduction of new technologies as the industry struggles through the worst recession of its history during the past couple of years. As projected, though, the half-pitch of 90 nm (hp90 nm) for DRAMs was introduced in 2004 (Intel's Prescott Pentium IV). Traditionally, the ITRS has focused on the continued scaling of CMOS (complementary metal-oxide-silicon) technology. By 2001, the horizon of the Roadmap started challenging the most optimistic projections for continued scaling of CMOS (e.g., MOSFET channel lengths below 9 nm). By that time it also became difficult for most people in the semiconductor industry to imagine how one could continue to afford the historic trends of increase in process equipment and factory costs for another 15 years! Thus, the ITRS started addressing post-CMOS devices. The Roadmap became necessarily more diverse for these devices, ranging from more familiar nonplanar CMOS devices to exotic new devices such as spintronics. Whether extensions of CMOS or radical new approaches, post-CMOS technologies must further reduce the cost per function and increase the performance of integrated circuits. Thus, new technologies may involve not only new devices but also new manufacturing paradigms.

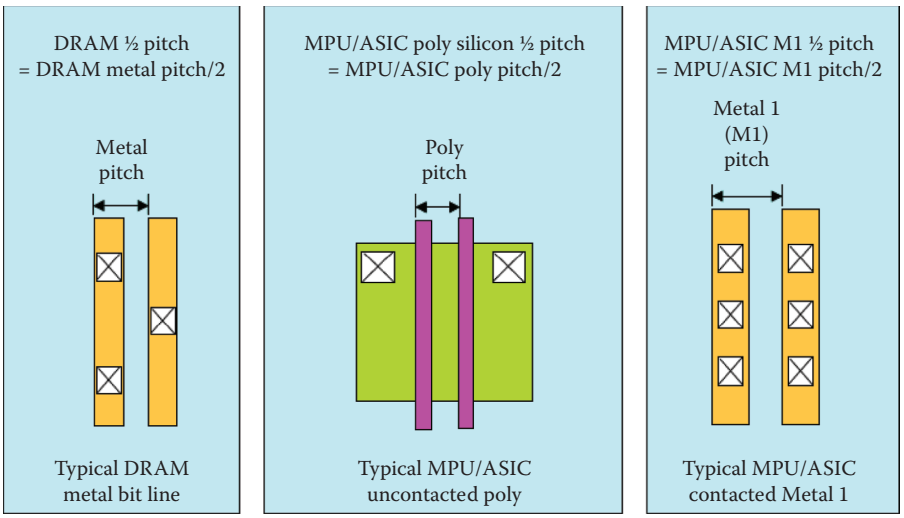
The ITRS technology nodes in the table below are defined as the minimum metal pitch used on any product, for example, either DRAM half-pitch or Metal 1 (M1) half-pitch in Logic/MPU (see also figure below the tables). In 2003, DRAMs continue to have the smallest metal half-pitch; thus, it continues to represent the technology node. Commercially used numbers for the technology generations typically differ from the ITRS technology node numbers. However, the most reliable technology standard in the semiconductor industry is provided by the above definition, which is quite clear in that the patterning and processing (e.g., etching) capabilities of the technology are represented as the pitch of the minimum metal lines. The above definition is maintained not only for the 2003 version but also as a continuation from previous ITRS editions. Therefore, the official 2003 ITRS metal hpXX node indicator has been added to differentiate the ITRS definition from commercial technology generation numbers. Interim shrink-level node trend numbers are calculated and included for convenience of monitoring the internode progress of the industry.

Near-Term Years

Year of Production	2003	2004	2005	2006	2007	2008	2009
Technology node		hp90			hp65		
DRAM half-pitch (nm)	100	90	80	70	65	57	50
MPU/ASIC MI half-pitch (nm)	120	107	95	85	75	67	60
MPU/ASIC Poly Si half-pitch (nm)	107	90	80	70	65	57	50
MPU printed gate length (nm)	65	53	45	40	35	32	28
MPU physical gate length (nm)	45	37	32	28	25	22	20

Long-Term Years

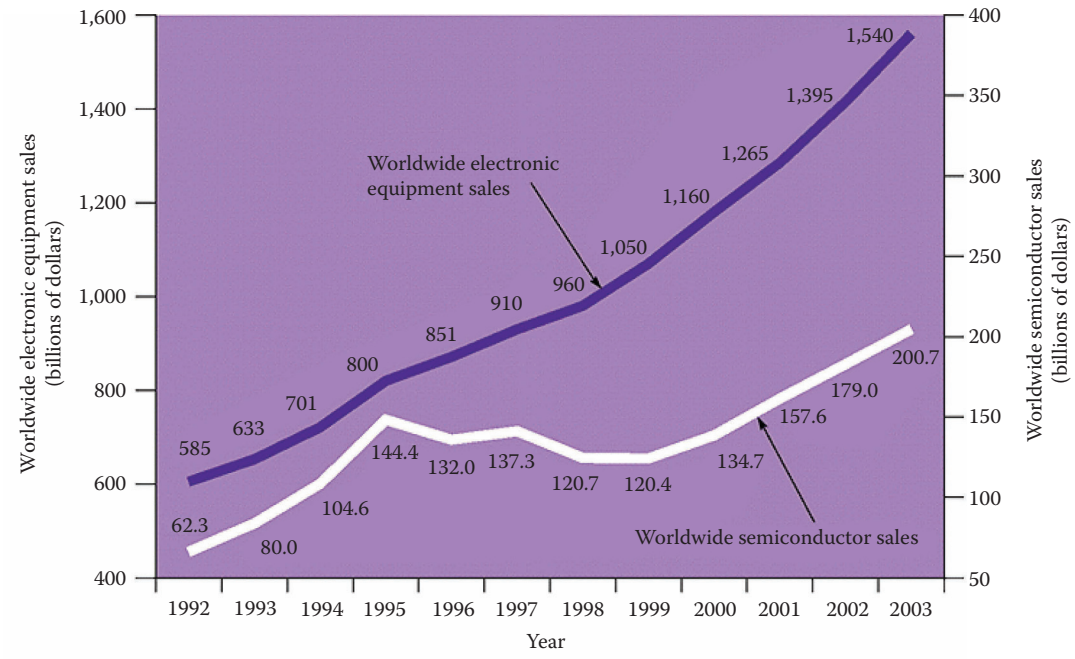
Year of Production	2010	2012	2013	2015	2016	2018
Technology node	hp45		hp32		hp22	
DRAM half-pitch (nm)	45	35	32	25	22	18
MPU/ASIC MI half-pitch (nm)	54	42	38	30	27	21
MPU/ASIC Poly Si half-pitch (nm)	45	35	32	25	22	18
MPU printed gate length (nm)	25	20	18	14	13	10
MPU physical gate length (nm)	18	14	13	10	9	7



Appendix 1B: Worldwide IC and Electronic Equipment Sales

	Amounts in US $M				Year on Year Growth in %			
	2003	2004	2005	2006	2003	2004	2005	2006
Americas	**32,330.7**	**39,514.2**	**41,734.6**	**40,089.1**	3.4	22.2	5.6	–3.9
Europe	**32,310.0**	**40,537.5**	**43,693.5**	**43,082.1**	16.3	22.5	7.8	–1.4
Japan	**38,942.2**	**47,822.9**	**51,066.8**	**50,306.9**	27.7	22.8	6.8	–1.5
Asia Pacific	**62,842.6**	**85,756.0**	**95,253.0**	**96,546.2**	22.8	36.5	11.1	1.4
Total world*	**166,425.5**	**213,630.6**	**231,748.0**	**230,024.4**	18.3	28.4	8.5	–0.7
Discrete semiconductors	**13,347.0**	**16,043.4**	**17,036.5**	**16,689.1**	8.1	20.2	6.2	–2.0
Optoelectronics	**9,544.7**	**13,100.8**	**14,851.7**	**15,281.2**	40.6	37.3	13.4	2.9
Sensors and actuators	**3,569.2**	**4,827.6**	**5,739.0**	**6,262.1**	†	35.3	18.9	9.1
Integrated circuits	**139,964.7**	**179,658.8**	**194,120.8**	**191,792.0**	16.1	28.4	8.0	–1.2
Bipolar	216.8	239.7	200.8	150.6	–4.2	10.6	–16.3	–25.0
Analog	26,793.9	33,652.2	36,971.8	36,952.4	12.0	25.6	9.9	–0.1
Micro	43,526.1	52,412.0	57,218.6	57,564.8	14.3	20.4	9.2	0.6
Logic	36,921.9	46,421.3	50,631.8	49,571.9	18.1	25.7	9.1	–2.1
Memory	32,506.0	46,933.6	49,097.9	47,552.3	20.2	44.4	4.6	–3.1
Total products*	**166,425.5**	**213,630.6**	**231,748.0**	**230,024.4**	18.3	28.4	8.5	–0.7

* All numbers are displayed as rounded to one decimal place, but totals are calculated to three decimal places precision.
† WSTS included actuators in this category from 2003. Before only sensors were reported. Therefore, a growth rate is not meaningful to show.



Percent semiconductor	10.7	12.6	14.9	18.1	15.5	15.1	12.6	11.5	11.6	12.5	12.8	13.0

Questions

1.1: Why is silicon so important to MEMS and NEMS?

1.2: Compare the pros and cons of transistors and vacuum tubes.

1.3: Why was the Si-MEMS market at one point in time expected to be much larger than the IC market?

1.4: Can you list some of the current technological and economic barriers that restrict the wider commercialization of Si-MEMS?

1.5: (a) State Moore's first law (we are talking about Moore, Intel's cofounder). (b) What is Moore's second law?

1.6: Why did surface micromachining catch on so fast with the IC industry?

1.7: Why are MEMS market forecasts so difficult to prepare? How would you go about making a better MEMS market forecast?

1.8: How does radar work? How is it useful?

1.9: What is the biggest advantage Ge has over Si IC circuits?

1.10: What is a strain gauge and what is its gauge factor?

1.11: What is the definition of nanotechnology?

1.12: List at least five commercial products that incorporate nanotechnology.

1.13: What year was the word "nanotechnology" first used?

1.14: What was Feynman's role in catalyzing the genesis of MEMS and NEMS?

1.15: Why was the honeymoon with the transistor over so quickly? What technology took over very fast?

1.16: What does ITRS stand for? What does it mean?

1.17: List a number of nanostructures that have been fabricated with bottom-up methodologies.

1.18: What is a photonic crystal?

1.19: What is a metamaterial?

1.20: What are the important differences between typical devices made in the IC industry and MEMS?

References

We took advantage of Google and Wikipedia on numerous occasions and also relied on the following references:

1. Ohl, R. S. 1946. Light-sensitive electric device, Bell Labs. US Patent 2402662.
2. Bardeen, J., and W. Brattain. 1950. Three electrode circuit element utilizing semiconductive materials, Bell Labs. US Patent 2524035.

3. Riordan, M., and L. Hoddeson. 1998. *Crystal fire: the birth of the information age.* New York: W. W. Norton & Company.
4. Reid, T. R. 2001. *The chip: how two Americans invented the microchip and launched a revolution*. New York: Random House Publishing House.
5. ITRS. 2007. International Technology Roadmap for Semiconductors. *Presentations from the 2007 ITRS Conference*, December 5, 2007, Makuhari Messe, Japan. http://www.itrs.net/Links/2007Winter/2007_Winter_Presentations/Presentations.html.
6. World Semiconductor Trade Statistics (WSTS). 2004. WSTS Semiconductor Market Forecast Spring 2004. Press release. http://www.wsts.org/plain/content/view/full/869.
7. Feynmann, R. 1959. There is plenty of room at the bottom. http://www.its.caltech.edu/~feynman/plenty.html.
8. Uhlir, A. 1956. Electrolytic shaping of germanium and silicon. *Bell Syst Tech J* 35:333–47.
9. Smith, C. S. 1954. Piezoresistance effect in germanium and silicon. *Phys Rev* 94:42–49.
10. Pfann, W. G. 1961. Improvement of semiconducting devices by elastic strain. *Solid State Electron* 3:261–67.
11. Tufte, O. N., P. W. Chapman, and D. Long. 1962. Silicon diffused-element piezoresistive diaphragms. *J Appl Phys* 33:3322–27.
12. Waggener, H. A., R. C. Kragness, and A. L. Taylor. 1967. *International Electron Devices Meeting, IEDM '67 68.* Washington, DC, IEEE.
13. Wallis, P. R., and D. I. Pomeranz. 1969. Field assisted glass-metal sealing. *J Appl Phys* 40:3946–49.
14. Shimbo, M., K. Furukawa, K. Fukuda, and K. Tanzawa. 1986. Silicon-to-silicon direct bonding method. *J Appl Phys* 60:2987–89.
15. National Semiconductor. 1974. *Transducers, Pressure, and Temperature* (catalog). Sunnyvale, CA: Author.
16. Angell, J. B., S. C. Tery, and P. W. Barth. 1983. Silicon micromechanical devices. *Scientific American* 248:44–55.
17. Greenwood, J. C. 1969. Ethylene diamine-cathechol-water mixture shows preferential etching of a p-n junction. *J Electrochem Soc* 116:1325–26.
18. Bohg, A. 1971. Ethylene diamine-pyrocatechol-water mixture shows etching anomaly in boron-doped silicon. *J Electrochem Soc* 118:401–02.
19. Texas Instruments. 1977. Texas Instruments Thermal Character Print Head. Austin, TX: Author.
20. Bassous, E., H. H. Taub, and L. Kuhn. 1977. Ink jet printing nozzle arrays etched in silicon. *Appl Phys Lett* 31:135–37.
21. O'Neil, P. 1980. A monolithic thermal converter. *Hewlett-Packard J*, 12.
22. Terry, S. C., J. H. Jerman, and J. B. Angell. 1979. A gas chromatograph air analyzer fabricated on a silicon wafer. *IEEE Trans Electron Devices* 26:1880–86.
23. Terry, S. C. 1975. *A gas chromatography system fabricated on a silicon wafer using integrated circuit technology.* PhD diss., Stanford University.
24. Petersen, K. E. 1980. Silicon torsional scanning mirror. *IBM J Res Dev* 24:631–37.
25. Petersen, K. E. 1982. Silicon as a mechanical material. *Proc IEEE* 70:420–57.
26. Howe, R. T., and R. S. Muller. 1983. Polycrystalline silicon micromechanical beams. *J Electrochem Soc* 130:1420–23.
27. Xia, Y., and G. M. Whitesides. 1998. Soft lithography. *Angew Chem Int Ed Engl* 37:551–75.
28. Nathanson, H. C., W. E. Newell, R. A. Wickstrom, and J. R. Davis. 1967. The resonant gate transistor. *IEEE Trans Electron Devices* ED-14:117–33.
29. Becker, E. W., W. Ehrfeld, D. Munchmeyer, H. Betz, A. Heuberger, S. Pongratz, W. Glashauser, H. J. Michel, and V. R. Siemens. 1982. Production of separation nozzle systems for uranium enrichment by a combination of x-ray lithography and galvanoplastics. *Naturwissenschaften* 69:520–23.
30. Fan, L. S., Y. C. Tai, and R. S. Muler. 1989. IC-processed electrostatic micro-motors. *Sensors and Actuators* A20:41–48.
31. Tang, W. C., T. C. Nguyen, and R. T. Howe. 1989. Laterally driven polysilicon resonant microstructures. *IEEE Micro Electro Mechanical Systems* 20:25–32.
32. Manz, A., N. Graber, and H. M. Widmer. 1990. Miniaturized total chemical analysis systems: A novel concept for chemical sensing. *Sens Actuators* B1:244–248.
33. Pohl, D. W., W. Denk, and M. Lanz. 1984. Optical stethoscopy: image recording with resolution l/20. *Appl Phys Lett* 44:651–53.
34. Allara, D. L., and R. G. Nuzzo. 1985. Spontaneously organized molecular assemblies; II. Quantitative infrared spectroscopic determination of equilibrium structures of solution adsorbed n-alkanoic acids on an oxidized aluminum surface. *Langmuir* 1:52–66.
35. Allara, D. L., and R. G. Nuzzo. 1985. Spontaneously organized molecular assemblies; I. Formation, dynamics and physical properties of n-alkanoic acids adsorbed from solution on an oxidized aluminum surface. *Langmuir* 1:45–52.
36. Tomalia, D. A., H. Baker, J. R. Dewald, M. Hall, G. Kallos, S. Martin, J. Roeck, J. Ryder, and P. Smith. 1985. A new class of polymers: starburst-dendritic macromolecules. *Polym J* 17:117–32.
37. Ashkin, A. 1970. Acceleration and trapping of particles by radiation pressure. *Phys Rev Lett* 24:156–59.
38. Ashkin, A., J. M. Dziedzic, J. E. Bjorkholm, and S. Chu. 1986. Observation of a single-beam gradient force optical trap for dielectric particles. *Optics Letters* 11:288–90.
39. Kroto, H. W., J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley. 1985. C60: buckminsterfullerene. *Nature* 318:162–63.
40. Tans, S. J., A. R. M. Verschueren, and C. Dekker. 1998. Room-temperature transistor based on a single carbon nanotube. *Nature* 393:49–52.
41. Ekimov, A. I., A. L. Efros, and A. A. Onushchenko. 1985. Quantum size effect in semiconductor microcrystals. *Solid State Commun* 56:921–24.
42. Efros, A. L., and A. L. Efros. 1982. Interband absorption of light in a semiconductor sphere. *Sov Phys Semicond* 16:772–74.
43. Brus, L. E. 1984. On the development of bulk optical properties in small semiconducting crystallites. *J Luminescence* 31/32: 381.
44. Brus, L. E. 1984. Electron-electron and electron-hole interactions in small semiconductor crystallites: the size dependence of the lowest excited electronic state. *J Chem Phys* 80:4403–07.
45. Murray, C. B., D. J. Norris, and M. G. Bawendi. 1993. Synthesis and characterization of nearly monodisperse CdE (E=S, Se, Te) semiconductor nanocrystallites. *J Am Chem Soc* 115:8706–15.
46. Wagner, R. S., and W. C. Ellis. 1964. Vapour-liquid-solid mechanism of single crystal growth. *Appl Phys Lett* 4:89–90.

47. Aviram, A., and M. A. Ratner. 1974. Molecular rectifiers. *Chem Phys Lett* 29:277.
48. Rayleigh, J. W. S. 1888. On the remarkable phenomenon of crystalline reflexion described by Prof. Stokes. *Phil Mag* 26:256–65.
49. Yablonovitch, E. 1987. Inhibited spontaneous emission in solid state physics and electronics. *Phys Rev Lett* 58:2059.
50. John, S. 1987. Strong localization of photons in certain disordered dielectric superlattices. *Phys Rev Lett* 58:2486.
51. Yablonovitch, E., T. J. Gmitter, and K. M. Leung. 1991. Photonic band structure: the face-centered cubic case employing nonspherical atoms. *Phys Rev Lett* 67: 2295–98.
52. Veselago, V. G. 1968. The electrodynamics of substances with simultaneously negative values of e and m. *Sov Phys Usp* 10:509–14.
53. Pendry, J. B. 1999. Photonic gap materials. *Curr Sci* 76:1311.
54. Pendry, J. B. 2000. Negative refraction makes a perfect lens. *Phys Rev Lett* 85:3966.
55. Pendry, J. B. 2004. Manipulating the near field with metamaterials. *Optics Photonics News* 15:33–37.
56. Pendry, J. B. 2007. Metamaterials and the control of electromagnetic fields. *Proceedings of the Ninth Rochester Conference on Coherence and Quantum Optics*. Washington, DC: Optical Society of America.
57. Pendry, J. B., A. J. Holden, D. J. Robbins, and W. J. Stewart. 1998. Low frequency plasmons in thin wire structures. *J Phys Cond Matter* 10:4785.
58. Ramakrishna, S. A., J. B. Pendry, M. C. K. Wiltshire, and W. J. Stewart. 2003. Imaging the near field. *J Mod Optics* 50:1419–30.
59. Wiltshire, M. C. K., J. B. Pendry, I. R. Young, D. J. Larkman, D. J. Gilderdale, and J. V. Hajnal. 2001. Microstructured magnetic material for RF flux guides in magnetic resonance imaging (MRI). *Science* 291:848–51.
60. Shelby, R. A., D. R. Smith, and S. Schultz. 2001. Experimental verification of a negative refractive index. *Science* 292:77–79.
61. Fang, N., H. Lee, and X. Zhang. 2005. Sub-diffraction-limited optical imaging with a silver superlens. *Science* 308:534–37.
62. Dolling, G., M. Wegener, C. M. Soukoulis, and S. Linden. 2007. Negative-index metamaterials at 780 nm wavelength. *Opt Lett* 32:53–55.
63. Drexel, K. E. 1987. *Engines of creation*. New York: Anchor Books.
64. Drexel, K. E. 1992. *Nanosystems, molecular machinery, manufacturing, and computation*. New York: John Wiley & Sons, Inc.
65. Atkinson, W. I. 2003. *Nanocosm: nanotechnology and the big changes coming from the inconceivable small*. New York: AMACOM.
66. Regis, E. 1995. *Nano: remaking the world atom by atom*. Boston: Little, Brown and Company.
67. Murray, C. B., C. R. Kagan, and M. G. Bawendi. 1995. Self-organization of CdSe nanocrystallites into three-dimensional quantum dot superlattices. *Science* 270:1335–38.
68. Reed, M. A., C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour. 1997. Conductance of a molecular junction. *Science* 278:252–54.
69. Geim, A. K., and K. S. Novoselov. 2007. The rise of graphene. *Nature Materials* 6:183–191.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2

Crystallography

The designer Tokujin Yoshioka makes his *Venus – Natural Crystal Chair* by submerging a block of polyester fibers in the shape of a straight-backed dining chair in a vat of water and then adding a mineral to crystallize it. (Courtesy of Mr. Tokujin Yoshioka.)

Outline

Introduction

Bravais Lattice, Unit Cells, and the Basis

Point Groups and Space Groups

Miller Indices

X-Ray Analysis

Reciprocal Space, Fourier Space, **k**-Space, or Momentum Space

Brillouin Zones

Nothing Is Perfect

Acknowledgments

Appendix 2A: Plane Wave Equations

Questions

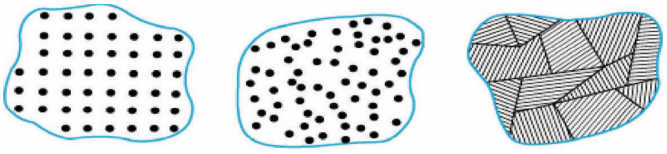
Further Reading

Reference

Introduction

Crystallography is the science of analyzing the crystalline structure of materials. The spatial arrangement of atoms within a material plays a most important role in determining the precise properties of that material. Based on the degree of order, materials are classified as amorphous, with no recognizable long-range order; polycrystalline, with randomly ordered domains (10 Å to a few μm); and single crystalline, where the entire solid is made up of repeating units in an orderly array. This classification is illustrated in Figure 2.1. Amorphous solids (e.g., glasses and plastics) are homogeneous and isotropic because there is no long-range order or periodicity in the internal arrangement.

Many engineering materials are aggregates of small crystals of varying sizes and shapes. The size of the single-crystal grains may be as small as a few nanometers but could also be large enough to be seen by the naked eye. Regions between grains are called grain boundaries. These polycrystalline materials have properties determined by both the chemical nature of the individual crystals and their aggregate properties, such as size and shape distribution, and in the orientation relationships between them. In the case of thin polycrystalline films, material properties might deviate significantly from bulk crystalline behavior, as we discover in



(a) Crystalline (b) Amorphous (c) Polycrystalline

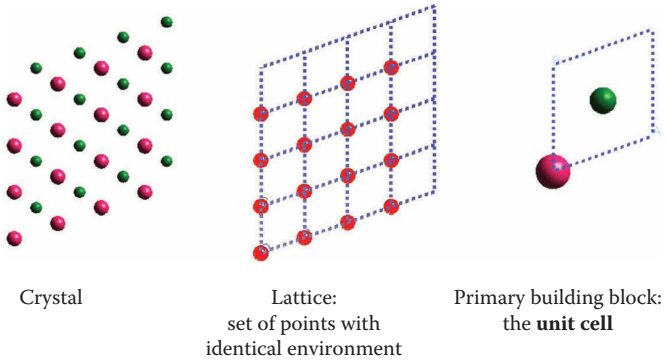
FIGURE 2.1 Classification of materials as crystalline (a), amorphous (b), and polycrystalline (c). (Drawing by Mr. Chengwu Deng.)

Volume II, Chapter 7, where we deal with thin film properties and surface micromachining. In the case of nanoparticles, deviation from expected bulk theory is even more pronounced (see Chapter 3 on quantum mechanics and the band theory of solids in the current volume). The crystal structure of a nanoparticle is not necessarily the same as that of the bulk material. Nanoparticles of ruthenium (2–3 nm in diameter), for example, have body-centered cubic (lattice point at each corner plus one at the center; see below) and face-centered cubic structures (lattice points at each corner as well as in the centers of each face; see below) not found in bulk ruthenium.

Most important, semiconductor devices are based on crystalline materials because of their reproducible and predictable electrical properties. Crystals are anisotropic—their properties vary with crystal orientation. In this chapter we explain the importance of the symmetry of point groups and space groups in determining, respectively, bulk physical properties and microscopic properties of crystalline solids, properties relied on for building miniaturized electronics, sensors, and actuators. We also launch the concept of reciprocal space (also called Fourier space, **k**-space, or momentum space), clarify the conditions for x-ray diffraction in terms of such a reciprocal space, and offer an introduction to Brillouin zones. All these elements are needed for our introduction to the band theory of solids in Chapter 3. We finish Chapter 2 with a description of crystal defects.

Bravais Lattice, Unit Cells, and the Basis

Under special conditions almost every solid can be made into a crystal (helium is the only substance that does not form a solid). Atoms organize themselves into crystals because energy can be minimized



Crystal Lattice: set of points with identical environment Primary building block: the **unit cell**

FIGURE 2.2 Any crystal lattice can be simplified to a three-dimensional array of periodically located points in space. Such a periodic array, specifying how the repeated units of a crystal are arranged, is called a *Bravais lattice*. A real crystal is made up of a basis and a lattice. (Drawing by Mr. Chengwu Deng.)

that way. Any crystal lattice can be simplified to a three-dimensional (3D) array of periodically located points in space as shown in Figure 2.2 in the case of a two-dimensional (2D) crystal. Such a periodic array, specifying how the repeated units of a crystal are arranged, is called a *Bravais lattice*. Bravais, in 1848, demonstrated that there are only 14 ways of arranging points symmetrically in space that do not lead to voids in a crystal (in 2D there are only five such lattices). All crystalline materials, including nanomaterials, assume one of the 14 Bravais lattices. A real crystal can be described in terms of a Bravais lattice, with one specific atom (or ion) or a group of atoms (a molecule), called a *basis*, attached to each lattice point (Figure 2.2). The basis superposed on the Bravais lattice renders the complete crystal structure. The 3D Bravais lattice can be mathematically defined by three noncoplanar basis vectors, $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$, which are the three independent shortest vectors connecting lattice points. These vectors form a parallelepiped called a *primitive cell*, i.e., a cell that can reproduce the entire crystal lattice by translation alone. Such a primitive cell is a minimum volume cell with a density of only one lattice point per cell—there are lattice points at each of the eight corners of the parallelepiped, but each corner point is shared among the eight cells that come together there (1/8 of a point at each corner). The lattice translational vector **r** is given by:

$$\mathbf{r} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \qquad (2.1)$$

where n_1, n_2, and n_3 are integers. A displacement of any lattice point by **r** will result in a new position in the lattice that has the same positional appearance as the original position. A lattice translation vector, **r**, as described in Equation 2.1, connects two points in the lattice that exhibit identical point symmetry.

Nonprimitive unit cells or simple unit cells are also called *conventional* unit cells or *crystallographic* unit cells. They are not necessarily unique and need not be the smallest cell possible. Primitive cells are chosen with the shortest possible vectors, whereas unit cells are chosen for the highest symmetry and may contain more than one lattice point per cell. The unit cell in a lattice, like a primitive cell, is representative of the entire lattice. The simplest unit cell belongs to a cubic lattice, which is further divided into simple cubic (SC), face-centered cubic (FCC), and body-centered cubic (BCC) as illustrated in Figure 2.3.

An FCC lattice has the closest atomic packing, then BCC, and then SC. For a simple cubic crystal (SC) unit cell, as shown in Figure 2.3, $\mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a}_3$, and the axes angles are $\alpha = \beta = \gamma = 90°$. The dimension a ($= \mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a}_3$) is known as the *lattice constant*. For SC the conventional unit cell coincides with the primitive cell. This is not true for FCC and BCC as we shall see in Figure 2.5 below.

The 14 possible Bravais lattices can be subdivided into 7 different "crystal classes" based on the choice of conventional unit cells. These 7 crystal classes are cubic, tetragonal, trigonal, hexagonal, monoclinic, orthorhombic, and triclinic. Each of these systems is characterized by a set of symmetry elements, and the more symmetry elements a crystal exhibits, obviously, the higher its symmetry. A cubic crystal has the highest possible symmetry and a triclinic crystal the lowest. The 14 Bravais lattices categorized according to the 7 crystal systems are shown in Figure 2.4.

A Wigner-Seitz cell is a primitive cell with the full symmetry of the Bravais lattice. It is an important construct for the understanding of Brillouin zones, the boundaries of which satisfy the Laue conditions for diffraction (see below). To appreciate how Wigner-Seitz cells are constructed, we illustrate some simple examples for the case of two different types of 2D lattices in Figure 2.5. Lines are drawn passing through the middle points of dotted lines connecting nearest neighbors. In 3D, the Wigner-Seitz cells

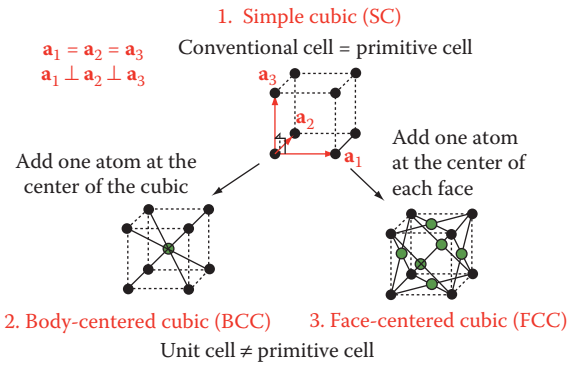


FIGURE 2.3 The simplest unit cell belongs to a cubic lattice, which is further divided into: simple cubic (SC), face-centered cubic (FCC), and body-centered cubic (BCC).

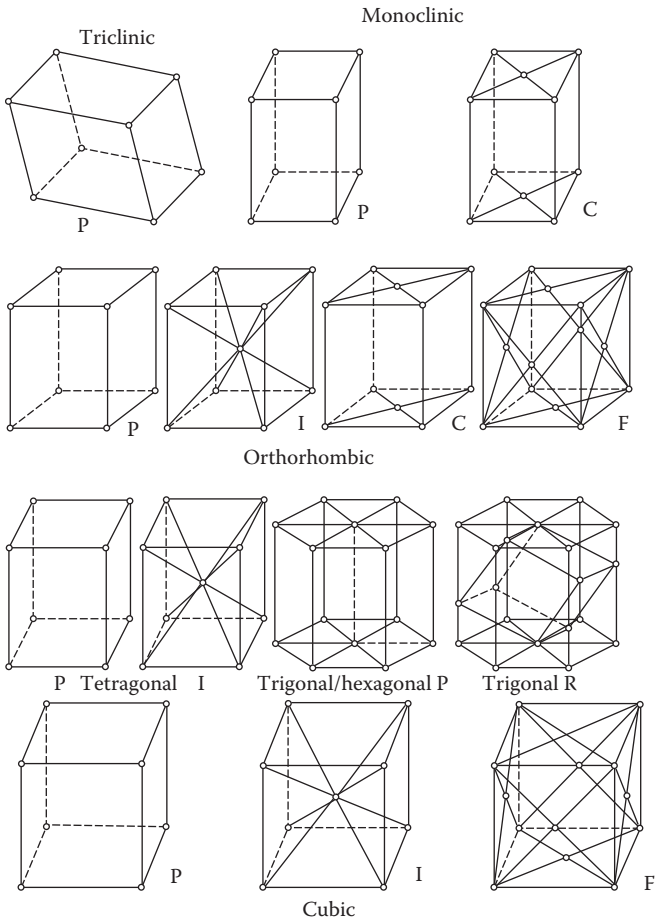


FIGURE 2.4 Conventional unit cells for the 14 Bravais lattices arranged according to the 7 crystal systems. *P* means lattice points on corners only, *C* means lattice points on corners as well as centered on faces, *F* means lattice points on corners as well as in the centers of all faces, and lattice points on corners as well as in the center of the unit cell body are indicated by *I*.

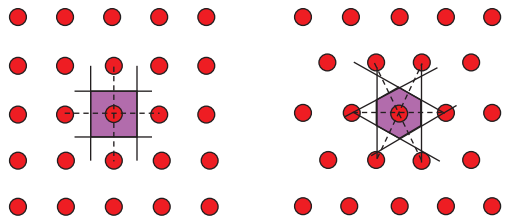


FIGURE 2.5 Wigner-Seitz primitive cells for two types of simple 2D lattices.

are polyhedra constructed about each atom by drawing planes that are the perpendicular bisectors of the lines between nearest neighbors. The Wigner-Seitz cell about a lattice point is the region of space that is closer to that point than any other lattice point. Wigner-Seitz cells for FCC and BCC Bravais lattices are shown in Figure 2.6. In the same figure we also show conventional unit and primitive unit cells for these lattices.

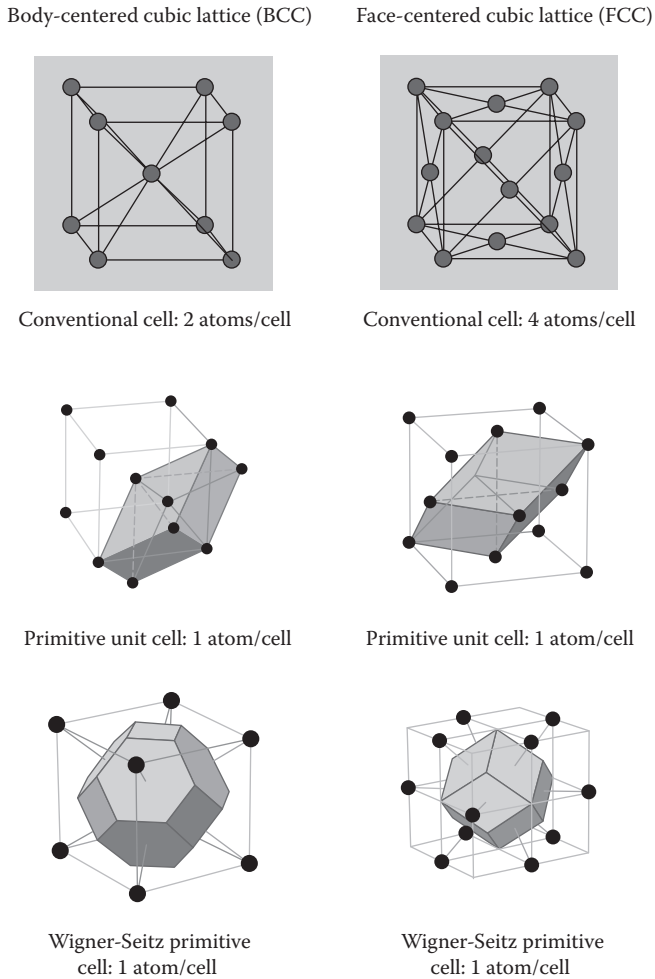


FIGURE 2.6 Conventional unit cells, primitive unit cells, and Wigner-Seitz primitive cells for BCC and FCC lattices. The BCC Wigner-Seitz unit cell is a truncated octahedron. The FCC Wigner-Seitz primitive unit cell is a rhombic dodecahedron.

Point Groups and Space Groups

A lattice translation as described by Equation 2.1 is a type of symmetry operation where a displacement of a crystal parallel to itself carries the crystal structure into itself (Figure 2.7).

Rotation and reflection or a combination of rotation and reflection—a so-called compound symmetry operation—about various points are other symmetry operations that may "carry the crystal into itself" (see Figure 2.8). The point around which the symmetry operation is carried out may be a lattice point or a special point within the elementary parallelepiped. There are five types of rotation axes possible, i.e., one- (360°), two- (180°), three- (120°), four- (90°), and sixfold (60°) rotation. One sees from Figure 2.8 why fivefold rotational symmetry does not occur in nature; it just cannot be stacked without leaving holes. This explains, for example, why we do not see ice crystals with a pentagon shape (Figure 2.9). Mirror reflection takes place about a plane through a lattice or special point. An inversion operation is an example of a compound symmetry operation and is achieved by rotation of π, followed by a reflection in a plane normal to the rotation axis; the effect is also illustrated in Figure 2.8. The collection of point symmetry elements possessed by a crystal is called a *point group* and is defined as the collection of symmetry operations which, when applied about a point, leave the lattice invariant. There are 32 crystallographic point groups in all.

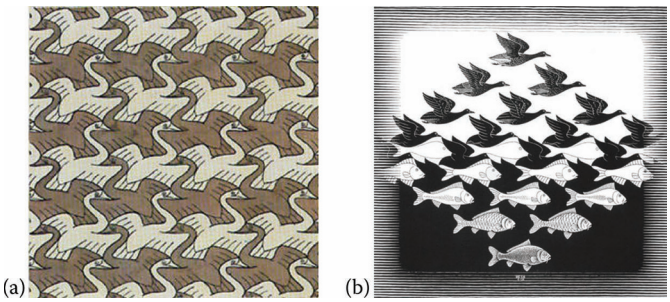


FIGURE 2.7 The drawing on the left (a) is crystal-like and can be carried into itself by a translation that is not possible in the figure on the right (b). The latter is missing a translation vector and is not crystal-like.

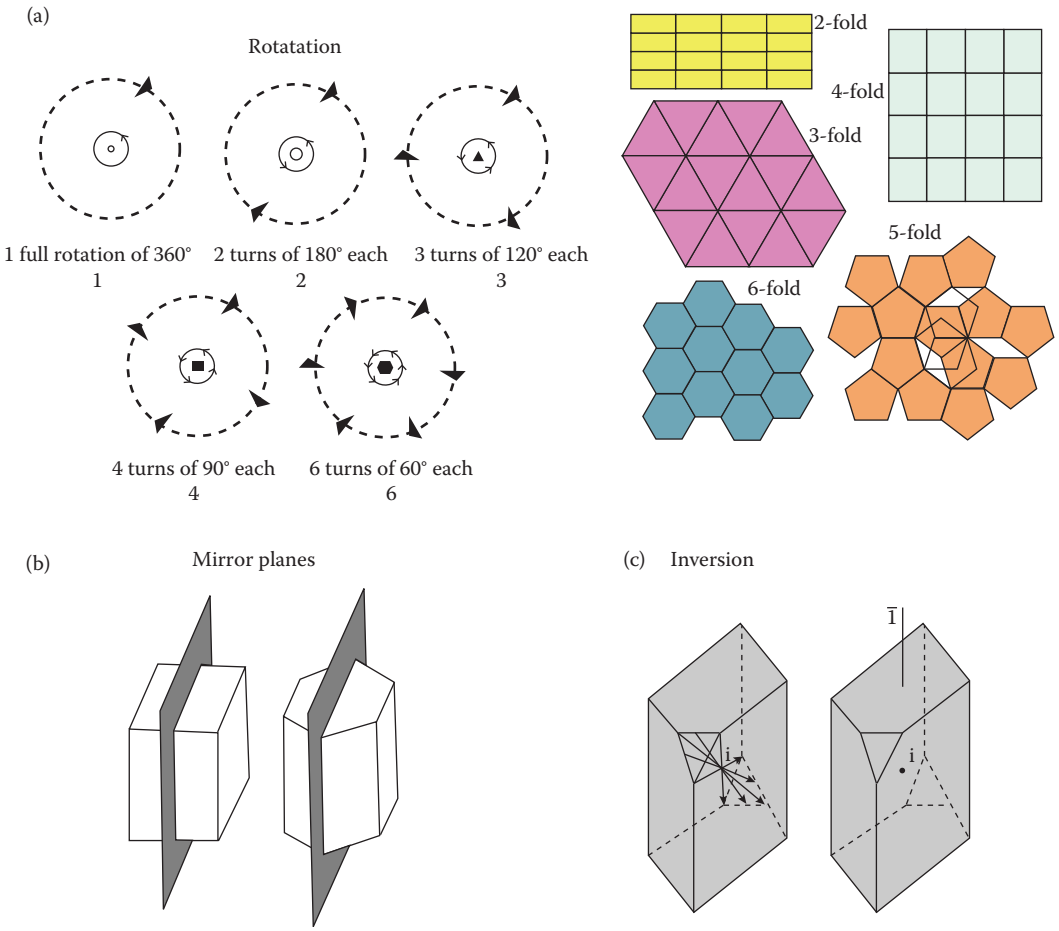


FIGURE 2.8 Point symmetry operations: (a) rotation, (b) reflection, and (c) a compound symmetry operation: inversion. The latter is made up of a rotation of π followed by reflection in a plane normal to the rotation axis. This is also called inversion through a point (i). The symbol for the inversion axis is $\bar{1}$.

The importance of the 32 point group symmetries and corresponding crystal classes is revealed by the important physical properties of crystalline solids they control, including electrical conductivity, thermal expansion, birefringence, piezoresistance, susceptibility, elastic stiffness coefficient, etc. Some properties that depend on the direction along which they are measured relative to the crystal axes are listed in Table 2.1. This orientation dependence of physical and chemical properties is called *anisotropy*. Anisotropy also explains why crystals do not grow into spheres but as polyhedra and why certain crystal directions etch faster than others. The reason for this anisotropy is the regular stacking of atoms in a crystal; as one passes along a given direction, one encounters atoms or groups of atoms at different intervals and from different angles than if one travels through the crystal from another direction. Single molecules in a liquid or a gas can also be anisotropic, but because they are free to move, liquids and gases are isotropic. In a crystal the anisotropy of atoms and groups of atoms is locked into the crystal structure. As a first example we consider a physical quantity such as current density **J** (column 4), and its cause the electrical field **E** (column 5). These quantities are linked, to a first approximation, in a linear relationship described by a tensor equation:

$$\mathbf{J} = \sigma\mathbf{E} \tag{2.2}$$

One remembers that a tensor field represents a single physical quantity that is associated with certain places in three-dimensional space and instants of time. The crystalline property, in column 1, is a tensor field of the rank listed in column 2. Also recall that a scalar (e.g., mass, temperature, charge) is a tensor of zeroth rank, and a vector (e.g., position, velocity, flow of heat) is a tensor of first rank. Other quantities such as stress inside a solid or fluid may be

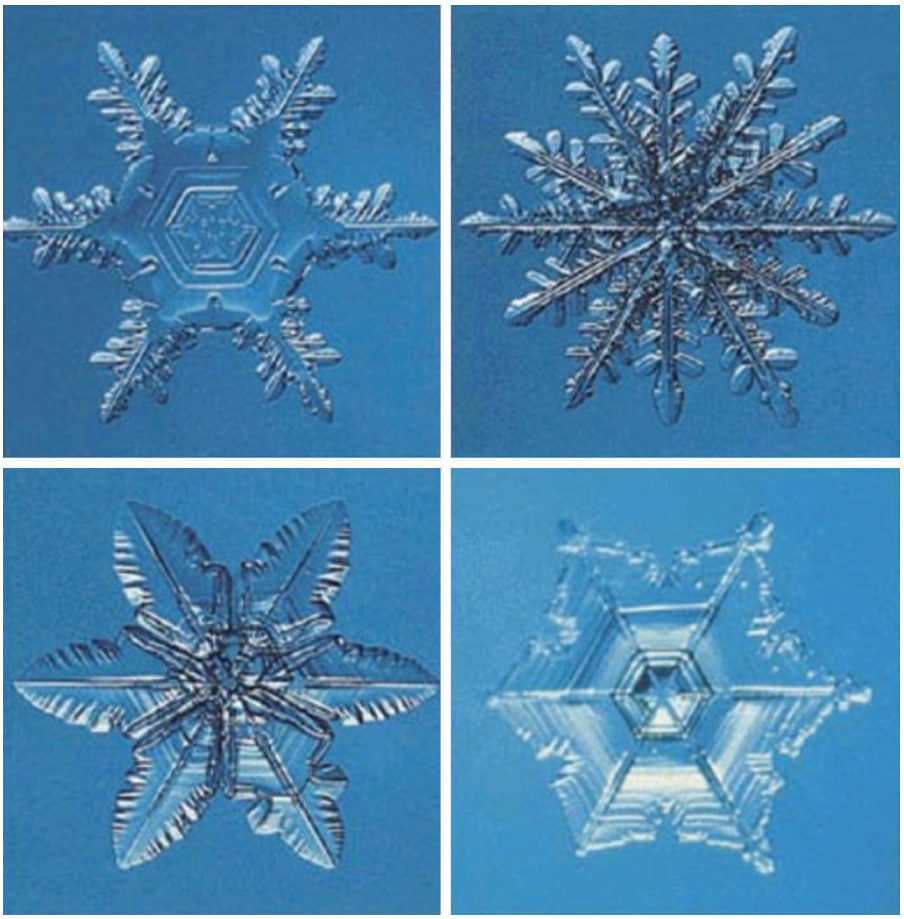


FIGURE 2.9 Ice crystals. No pentagons are found in ice crystal stacking.

characterized by tensors of order two or higher. The physical properties listed in Table 2.1 are exploited to build electronics, sensors, and actuators in MEMS and NEMS.

We elaborate a bit further here on the anisotropy of the electrical conductivity σ in a single crystal, where a causal or forcing term, the electrical field **E**, causes a current density **J**. Because both are vectors, this case is referred to as a *vector-vector* effect. In general, the current vector may not have the same direction as the electric vector. Assuming a linear relationship between electrical field (cause) and

TABLE 2.1 Linear Physical Properties of Solids

Property and Symbol	Rank of Tensor	Number of Independent Components	Dependent Physical Quantity	Causal or Forcing Term
Pyroelectric coefficient p	1	3	Electric polarization dP	Temperature change dT
Conductivity σ	2	6	Current density **J**	Field **E**
Resistivity ρ	2	6	Field **E**	Current density **J**
Susceptibility χ	2	6	Electric polarization P	Electric field **E**
Thermal expansion α	2	6	Strain ε	Temperature change dT
Piezoelectric coefficient d	3	18	Electric polarization P_s	Stress S
Elastic stiffness constants χ	4	21	Stress S	Strain ε
Elastic compliance σ	4	21	Stress ε	Stress S
Piezoresistance π	4	21	Resistivity change ρ	Stress S
Unfortunately, symbols customarily used for these properties do sometimes overlap.				

current (effect), we can describe the components of the current relative to an arbitrarily chosen Cartesian coordinate system as:

$$\begin{aligned} J_x &= \sigma_{xx}E_x + \sigma_{xy}E_y + \sigma_{xz}E_z \\ J_y &= \sigma_{yx}E_x + \sigma_{yy}E_y + \sigma_{yz}E_z \\ J_z &= \sigma_{zx}E_x + \sigma_{zy}E_y + \sigma_{zz}E_z \end{aligned} \tag{2.3}$$

The quantities σ_{ik} are components of a 3×3 "conductivity tensor." The resistivity tensor $\rho(= 1/\sigma)$ tensor, like the conductivity tensor, is a second rank tensor described by:

$$\mathbf{E} = \rho\mathbf{J} \tag{2.4}$$

Based on the basic symmetry of the equations of motion, Onsager demonstrated that the tensor is symmetric, i.e., $\sigma_{ik} = \sigma_{ki}$, so that the nine coefficients are always found to reduce to six. Taking advantage of this symmetry argument and multiplying the expressions in Equation 2.3 by E_x, E_y, and E_z, respectively, one obtains on adding:

$$\begin{aligned} J_xE_x + J_yE_y + J_zE_z &= \sigma_{xx}E^2{}_x + \sigma_{yy}E^2{}_y + \sigma_{zz}E^2{}_z \\ &+ 2\sigma_{xy}E_xE_y + 2\sigma_{yz}E_yE_z + 2\sigma_{zx}E_zE_x \end{aligned} \tag{2.5}$$

To make the mixed terms on the right side of Equation 2.5 disappear, one chooses a new coordinate system with the coordinates along the principal axes of the quadratic surface represented by this right-hand side (rhs) term, and in this new coordinate system one obtains:

$$J_x = \sigma_1E_x;\ J_y = \sigma_2E_y;\ J_z = \sigma_3E_z \tag{2.6}$$

where σ_1, σ_2, and σ_3 are the principal conductivities. The current and field vectors only have the same direction when the applied field falls along any one of the principal axes of the crystal. From Equation 2.6, no matter how low the symmetry of a crystal, it can always be characterized by three conductivities (σ_1, σ_2, and σ_3) or three specific resistivities (ρ_1, ρ_2, and ρ_3). In cubic crystals the three quantities are equal, and the specific resistivity does not vary with direction. In hexagonal, trigonal, and tetragonal crystals, two of the three principal conductivities (or resistivities) are the same. In such a case, the resistivity only depends on the angle θ between the direction in which ρ is measured and the hexagonal, trigonal, or tetragonal axis. One then finds:

$$\rho(\phi) = \rho_{per}\sin^2\phi + \rho_{par}\cos^2\phi \tag{2.7}$$

where the subscripts stand for perpendicular and parallel to the axis.

Another vector-vector effect example in Table 2.1 is the one involving thermal conductivity, where a thermal current vector is caused by a thermal gradient. Scalar-tensor effects lead to similar relations as vector-vector effects. For example, the deformation tensor of a solid resulting from a temperature change (scalar) involves three principal expansion coefficients, α_1, α_2, and α_3. The latter will again all be equal in the case of a cubic crystal, and the angular dependence of α for hexagonal, trigonal, and tetragonal crystals is given by an expression analogous to Equation 2.7.

A simple example of a scalar-vector effect from Table 2.1, illustrating the importance of crystal symmetry or lack thereof, involves pyroelectricity (see first row in Table 2.1). Pyroelectricity is the ability of a material to spontaneously polarize and produce a voltage as a result of changes in temperature. It must be a change in temperature: incident light may heat a pyroelectric crystal, thus changing its dipole moment and causing current to flow, but because pyroelectrics respond to the rate of change of temperature only, the light or heat source must be pulsed or modulated! A pyroelectric is a ferroic material, i.e., a class of smart multifunctional materials having both sensing and actuating functions. *Ferroic materials* is a simplified term to represent ferroelastic, ferromagnetic, and ferroelectric materials. In pyroelectricity, the opposite faces of certain crystals [e.g., tourmaline $(Na, Ca)(Li, Mg, Al)(Al, Fe, Mn)_6(BO_3)_3(Si_6O_{18})(OH)_4$ the "Ceylon magnet," ZnO, $BaTiO_3$, and $PbTiO_3$] become electrically charged as a result of a change in temperature. This is illustrated for tetragonal $BaTiO_3$ in Figure 2.10. In Table 2.1 we consider the linear relationship between the electric polarization **P** (a vector) and a temperature change (a scalar). Electric polarization of materials is covered from a theoretical point of view in Chapter 5. For practical applications of pyroelectricity in actuator construction refer to

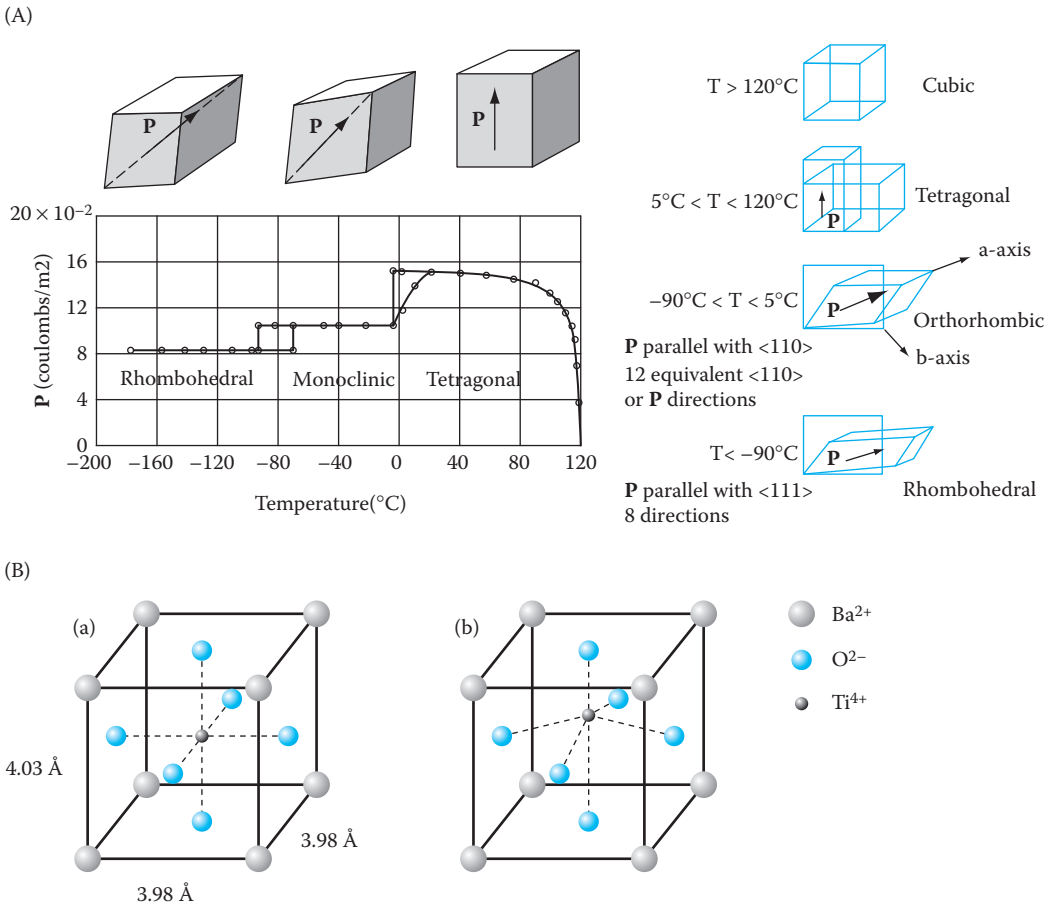


FIGURE 2.10 (A) In the pyroelectric crystal $BaTiO_3$, **P** changes with temperature only when the material is in its tetragonal state. Pyroelectricity only occurs in a crystal lacking an inversion center. This is clear from (B) (a). In cubic $BaTiO_3$ the oxygen ions are at face centers; Ba^{2+} ions are at cube corners; and Ti^{4+} is at cube center. (B) (b) In tetragonal $BaTiO_3$, the Ti^{4+} is off-center, and the unit cell has a net polarization. (Drawing by Mr. Chengwu Deng.)

Volume III, Chapter 8. The crystalline property, the pyroelectric coefficient **p**, is a tensor of rank 1 (a vector in other words). One can predict readily that pyroelectricity only occurs in crystals that lack an inversion center or a center of symmetry, i.e., in noncentrosymmetric crystals with one or more polar axes. Indeed, one could not have a crystal with one face positively charged and one negatively charged—i.e., with a polar axis—as a result of a uniform change in temperature if these crystal faces were equivalent.

More complicated cases involve vector-tensor effects and tensor-tensor effects. Piezoelectricity is an example of a vector-tensor effect; an electric field (vector) causes a mechanical deformation (tensor). Elastic deformation under the influence of a stress tensor is an example of a tensor-tensor effect. These effects require many more constants than the simple examples presented above. In discussing actuators in Volume III, Chapter 8, we will find out that an increase in symmetry introduces major simplifications in the coefficient matrices and that to describe the tensors correctly the point-group symmetry of the crystal must be known.

The combined effects of rotation or reflections from the point groups with translation from the Bravais lattice results in two additional symmetries: screw axes and glide planes. A screw axis combines rotation and translation, and a glide plane combines reflection with translation (Figure 2.11). Considering the various combinations involving the 32 point groups, screw axes, and glide planes, as well as the different Bravais lattices, a total of 230 different possible "space groups" results. In protein crystals there are only 65 space groups because all natural products are chiral so that inversion and mirror symmetry

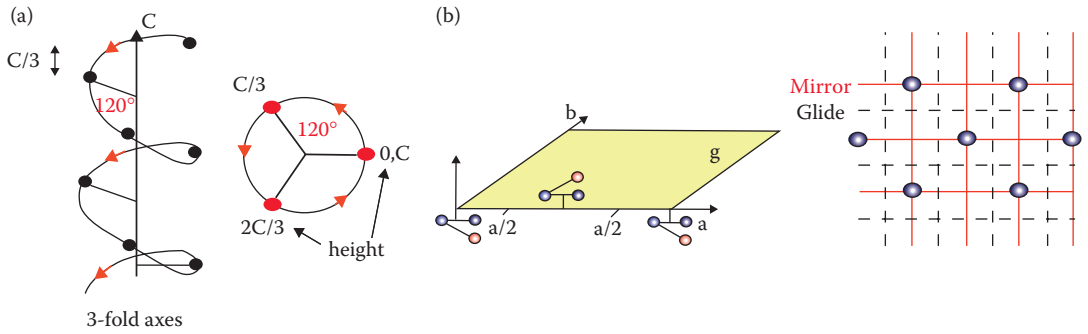


FIGURE 2.11 Example of a screw axis and a glide plane. (a) *N*-fold screw axes *C*: a combination of a rotation of 360°/*n* around *C* and a translation by an integer of *C*/*n*. (b) Glide plane: a translation parallel to the glide plane *g* by *a*/2.

operations are not allowed. A space group is a group that includes both the point symmetry elements and the translations of a crystal. These space groups are most important when studying the microscopic properties of solids. The space groups for most inorganic substances are known and can be found in tables in the Inorganic Crystal Structure Database (ICSD). These tables make it possible to calculate the exact distances and angles between different atoms in a crystal. The external shape of a crystal is referred to as the *habit*. Not all crystals have well-defined external faces. Natural faces always have low indices, i.e., their orientation can be described by Miller indices that are small integers as introduced next. The faces that we see are the lowest energy faces as the surface energy is minimized during growth.

Miller Indices

To identify a plane or a direction in a crystal, a set of integers *h*, *k*, and *l*, called the *Miller indices*, are widely used. To determine the Miller indices of a plane, one takes the intercept of that plane with the axes and expresses these intercepts as multiples of the base vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$. The reciprocal of these three integers is taken, and, to obtain whole numbers, the three reciprocals are multiplied by the smallest common denominator. The resulting set of numbers is written down as (*hkl*). By taking the reciprocal of the intercepts, infinities (∞) are avoided in plane identification. Parentheses or braces are used to specify planes.

The rules for determining the Miller indices of a direction or an orientation in a crystal are as follows: translate the orientation to the origin of the unit cell, and take the normalized coordinates of its other vertex. For example, the body diagonal in a cubic lattice as shown in the right-most panel in Figure 2.12 is 1**a**, 1**a**, and 1**a** or the [111] direction. The Miller indices for a direction are thus established using the same procedure for finding the components of a vector. Brackets or carets specify directions.

Directions [100], [010], and [001] are all crystallographically equivalent and are jointly referred to as the family, form, or group of <100> directions. A form, group, or family of faces that bear like relationships to the crystallographic axes—e.g., the planes (001), (100), (010), (00$\bar{1}$), ($\bar{1}$00), and (0$\bar{1}$0)—are all equivalent, and they are marked as {100} planes (see Figure 2.13). A summary of the typical representation for Miller indices is shown in Table 2.2. The orientation of a plane is defined by the direction

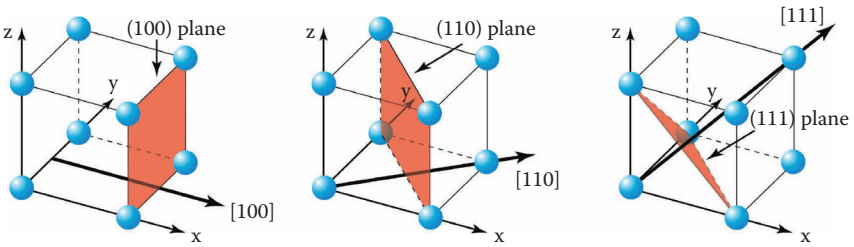


FIGURE 2.12 Miller indices for planes and directions in an SC cubic crystal. Shaded planes are from left to right (100), (110), and (111). (Drawing by Mr. Chengwu Deng.)

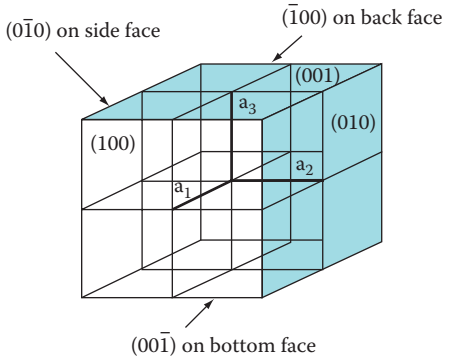


FIGURE 2.13 Miller indices for the planes of the {100} family of planes.

of a normal to the plane or the vector product ($\mathbf{A} \times \mathbf{B} = \mathbf{C}$). For a cubic crystal (such as silicon or gallium arsenide), the plane (*hkl*) is perpendicular to the direction [*hkl*]. In other words, the indices of a plane are the same numbers used to specify the normal to the plane. Using a simple cubic lattice as an example, you can check that crystal planes with the smallest Miller indices, such as {100}, {110}, {111}, have the largest density of atoms. Usually crystals are cleaved along these planes and are grown in directions perpendicular to them.

When one comes across more complicated planes than the ones considered above, the mathematical vector algebra approach to calculate the Miller indices becomes useful. For examples, consider the plane in Figure 2.14, defined by three points P1, P2, and P3, where P1: (400), P2: (020), and P3: (003).

Step 1. Define the following vectors:

$$\mathbf{r} = x\mathbf{a}_1 + y\mathbf{a}_2 + z\mathbf{a}_3$$

$$\mathbf{r}_1 = 4\mathbf{a}_1 + 0\mathbf{a}_2 + 0\mathbf{a}_3$$

$$\mathbf{r}_2 = 0\mathbf{a}_1 + 2\mathbf{a}_2 + 0\mathbf{a}_3 \qquad (2.8)$$

$$\mathbf{r}_3 = 0\mathbf{a}_1 + 0\mathbf{a}_2 + 3\mathbf{a}_3$$

TABLE 2.2 Miller Indices Symbols

Notation	Interpretation
(*hkl*)	Crystal plane
{*hkl*}	Equivalent planes
[*hkl*]	Crystal direction
<*hkl*>	Equivalent directions

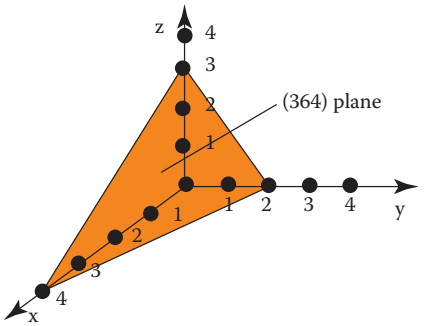


FIGURE 2.14 The (364) plane in a SC cubic lattice.

and find the differences:

$$\mathbf{r} - \mathbf{r}_1 = (x - 4)\,\mathbf{a}_1 + (y - 0)\,\mathbf{a}_2 + (z - 0)\,\mathbf{a}_3$$

$$\mathbf{r}_2 - \mathbf{r}_1 = (0 - 4)\,\mathbf{a}_1 + (2 - 0)\,\mathbf{a}_2 + (0 - 0)\,\mathbf{a}_3 \qquad (2.9)$$

$$\mathbf{r}_3 - \mathbf{r}_1 = (0 - 4)\,\mathbf{a}_1 + (0 - 0)\,\mathbf{a}_2 + (3 - 0)\,\mathbf{a}_3$$

Step 2. Calculate the scalar triple product of these three vectors [$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$], which in this case is a plane and its volume is zero because the vectors are coplanar [$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = 0$]:

$$(\mathbf{r} - \mathbf{r}_1) \cdot [(\mathbf{r}_2 - \mathbf{r}_1) \times (\mathbf{r}_3 - \mathbf{r}_1)] = 0 \qquad (2.10)$$

For vectors **A**, **B**, **C** with coordinates (A_1, A_2, A_3), (B_1, B_2, B_3), and (C_1, C_2, C_3), the requirement in Equation 2.10 is equivalent to:

$$\mathbf{A} \cdot \mathbf{B} \times \mathbf{C} = \begin{vmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \\ C_1 & C_2 & C_3 \end{vmatrix} = A_1(B_2C_3 - B_3C_2)$$

$$+ A_2(B_3C_1 - B_1C_3) + A_3(B_1C_2 - B_2C_1) \qquad (2.11)$$

For our example this leads to:

$$\begin{vmatrix} x - 4 & y & z \\ -4 & 2 & 0 \\ -4 & 0 & 3 \end{vmatrix} = (x - 4)6 + 12y + 8z = 0$$

or $3x + 6y + 4z = 12$

Once we have the equation for the plane, we easily find the Miller indices.

Step 3. Determine the Miller indices.

1. The intercepts with the axes: $x = 4$ (for $y = z = 0$), $y = 2$ (for $x = z = 0$), and $z = 3$ (for $x = y = 0$)
2. The reciprocals 1/4, 1/2, and 1/3, or
3. The Miller indices for the plane are (364)

Adjacent planes (*hkl*) in a simple cubic crystal are spaced a distance d_{hkl} from each other, with d_{hkl} given by:

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \tag{2.12}$$

where *a* is the lattice constant. Equation 2.12 provides the magnitude of d_{hkl} and follows from simple analytic geometry. To generalize this expression, notice that for a plane (*hkl*) with $hx + ky + lz = a$, the distance from any point (x_1, y_1, z_1) to this plane is:

$$d_{hkl} = \left| \frac{hx_1 + ky_1 + lz_1 - a}{\left(h^2 + k^2 + l^2\right)^{\frac{1}{2}}} \right| \tag{2.13}$$

Hence when that point is at origin (0, 0, 0) we find Equation 2.12 back.

- Example 2.1: With a = 5 Å, we find d = a = 5 Å for (100) planes and d = a/$\sqrt{2}$ = 3.535 Å for (110) planes.

Because a, b = |a||b| cos θ the angle between plane (h_1, k_1, l_1) and plane (h_2, k_2, l_2) can be calculated as:

$$\cos\theta = \frac{(h_1h_2 + k_1k_2 + l_1l_2)}{\sqrt{h_1^2 + k_1^2 + l_1^2}\sqrt{h_2^2 + k_2^2 + l_2^2}} \tag{2.14}$$

- Example 2.2: The angle between *a* (100) and *a* (111) plane is cos θ = 1/$\sqrt{1}\sqrt{3}$ = 0.58 or θ = 54.74°.

X-Ray Analysis

Introduction

X-ray analysis reveals the symmetries of crystals (lattice type), distances between atomic planes (lattice parameter), the positions of atoms in crystals, the types of atoms from the intensities of diffracted x-rays, and the degree of crystallinity (ordering). To perform x-ray crystallography, it is necessary to grow crystals with edges of around 0.1–0.3 mm. This is usually not a problem for inorganic materials, but in the case of organic materials, such as proteins (for example, see the crystal structure of the GFP protein in Figure 7.108) and nucleic acids (see x-ray diffraction image in Figure 2.17), it often is a challenge: imagine trying to crystallize a molecule with 10,000 atoms! A crystallographer must combine ingenuity and patience to trick these molecules into crystallizing. In this section we learn about diffraction and the all-important Bragg and Laue x-ray diffraction laws and how the latter are used to deduce the 3D structure of molecules.

Fourier Transforms

Diffraction forms the basis for x-ray crystallography. The first step toward interpreting diffraction patterns was a mathematical trick discovered by the French mathematician Joseph Fourier, who in 1807 introduced Fourier transforms for solving heat conduction problems. The result of a Fourier transform is that periodic functions in the time domain, e.g., light waves, can be completely characterized by information in the frequency domain, i.e., by frequencies and amplitudes of sine, cosine functions. Fourier analysis provides us with the tools to express most functions as a superposition of sine and cosine waves of varying frequency. For periodic signals a discrete sum of sines/cosines of different frequencies is multiplied by a different weighting coefficient in a so-called Fourier series (FS). For nonperiodic functions, one needs a continuous set of frequencies so the integral of sines/cosines is multiplied by a weighting function in a so-called Fourier transform (FT) (Figure 2.15). One important property of Fourier transforms is that they can be inverted. If you apply a Fourier transform to some function, you can take the result and run it through an inverse Fourier transform to get the original function back. The inverse Fourier transform is essentially just another Fourier transform. Fourier and inverse Fourier transforms, which take the signal back and forth between time and frequency domains, are:

$$\text{From time to frequency: } X(f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t)\, e^{-2\pi i f t} dt \tag{2.15}$$

$$\text{From frequency to time: } x(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} X(f)\, e^{2\pi i f t} df \tag{2.16}$$

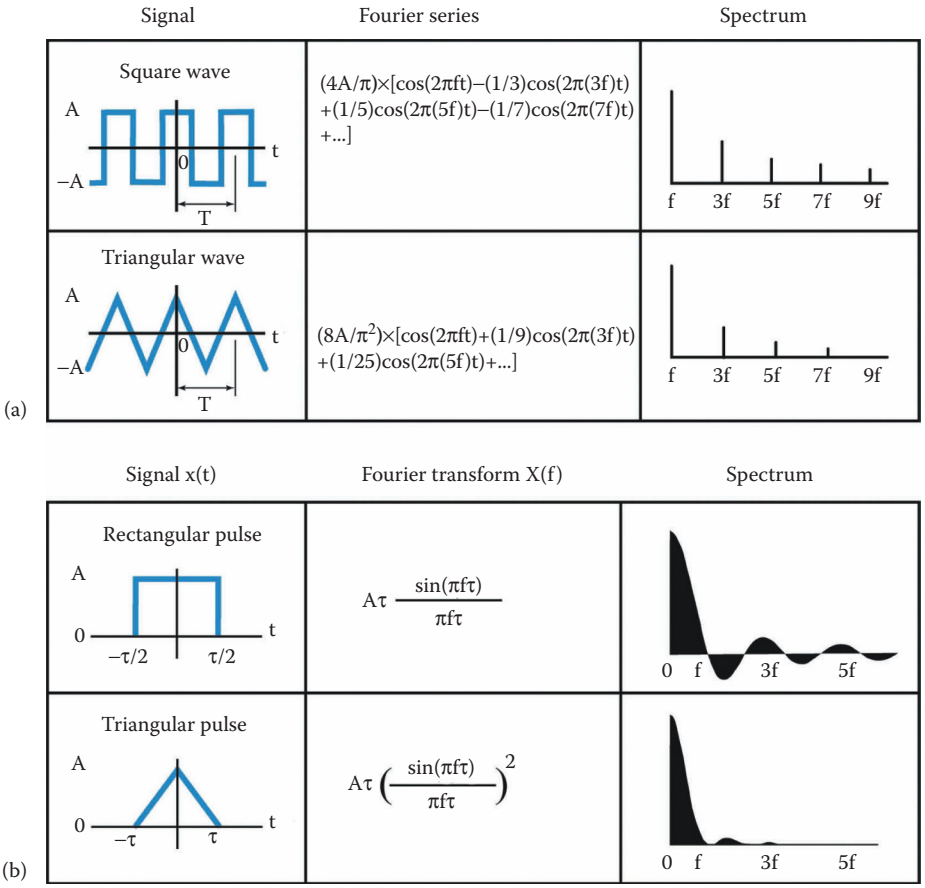


FIGURE 2.15 (a) For periodic signals a discrete sum of sines/cosines of different frequencies is multiplied by a different weighing coefficient in a so-called Fourier series (FS). (b) For nonperiodic functions, one needs a continuous set of frequencies so the integral of sines/cosines is multiplied by a weighting function in a so-called Fourier transform (FT).

In Chapter 5 we will see that the formation of an image, according to Abbe's theory, is a two-stage, double-diffraction process: an image is the diffraction pattern of the diffraction pattern of an object. In x-ray diffraction there is no lens to focus the x-rays, so we have to use a computer to reassemble the image: the x-ray diffraction patterns from a crystal are related to the object diffracting the waves through a Fourier transform.

Fourier transforms are actually even more general than revealed here: a FT allows for a description given in one particular "space" to be transformed to a description in the reciprocal of that space, and time-frequency transformation is just one example. It is interesting to note that strong criticism by peers blocked publication of Fourier's work until 1822 (*Theorie Analytique De La Chaleur*). Today, Fourier analysis is used in GSM (global system for mobile communications)/cellular phones, most DSP (digital signal processing)-based applications, music, audio, accelerator control, image processing, x-ray spectrometry, chemical analysis (FT spectrometry), radar design, PET scanners, CAT scans and MRI, speech analysis (e.g., voice-activated "devices," biometry), and even stock market analysis.

X-Ray Diffraction

Introduction

Wilhelm Conrad Roentgen discovered x-rays in 1895 and received the Nobel Prize in Physics in 1901 for his discovery. X-rays are scattered by the electrons in atoms because electromagnetic radiation (including x-rays) interacts with matter through its fluctuating electric field, which accelerates charged particles. You can think of electrons oscillating in position and, through their accelerations, re-emitting electromagnetic radiation. The scattered radiation

interferes both constructively and destructively, producing a diffraction pattern that can be recorded on a photographic plate. For x-rays, electrons, and neutrons incident on a single crystal, diffraction occurs because of interference between waves scattered elastically from the atoms in the crystal. Intensity of scattered radiation is proportional to the square of the charge/mass ratio, and the proton is about 2000 times as massive as the electron. Because electrons have a much higher charge-to-mass ratio than atomic nuclei or even protons, they are much more efficient in this process. With x-rays the interaction is with the electron mantle of the atoms. In the case of electron beams, say in an electron microscope, scattering is from both the electron mantle and the atom nuclei, and neutrons interact with the nucleus only.

The final result of a crystallographic experiment is not a picture of the atoms, but a map of the distribution of electrons in the molecule, i.e., an electron density map (Figure 2.16). Because the electrons are mostly tightly localized around the nuclei, the electron density map gives us a pretty good picture of the molecule. As we do not have a lens we do not get the electron density map directly; the x-ray diffraction patterns from a crystal are related to the object diffracting the waves through a Fourier transform.

If one thinks of electron density as a mathematical function:

$$\rho(x, y, z) \tag{2.17}$$

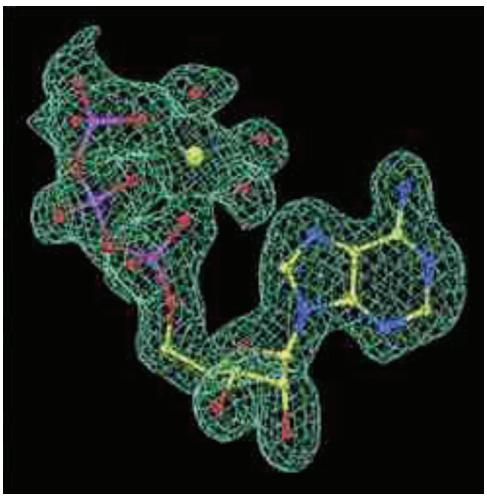


FIGURE 2.16 Electron density map of adenosine triphosphate (ATP).

with x, y, z indices for real space, then the diffraction pattern is the Fourier transform of that electron density function and given as:

$$F(hkl) = T[\rho(x, y, z)] \tag{2.18}$$

where $F(hkl)$ is the structure factor (a scattered wave, therefore a complex number with amplitude and phase) with hkl indices in reciprocal space, and T is the forward Fourier transform of $\rho(x, y, z)$. As we saw in Equation 2.16, the reverse relationship holds also, namely:

$$\rho(x, y, z) = T^{-1}\{T[\rho(x, y, z)]\} \tag{2.19}$$

where T^{-1} is the inverse Fourier transform. This expression tells us that the inverse Fourier transform of the Fourier transform of an object is the original object. The latter is a rewording of Abbe's treatment of image generation as presented in Chapter 5. The intensity in an x-ray diffraction photograph is the square of the amplitude of the diffracted waves, $|F(X)|^2$, or the recorded diffraction pattern of an object is the square of the Fourier transform of that object. In a diffraction pattern, each point arises from the interference of rays scattered from all irradiated portions of the object. To determine the image, one must measure or calculate the structure factor $F(X)$ at many or all points of the diffraction pattern. Each $F(X)$ is described by an amplitude and a phase, but in recording the intensity of the diffracted x-rays, the phase information is lost. This is referred to as the *phase* problem. X-ray phases could be obtained directly if it were possible to rediffract (focus) the scattered rays with an x-ray lens to form an image; unfortunately, an x-ray lens does not exist. With x-rays we can thus detect diffraction from molecules, but we have to use a computer to reassemble an image as shown in Figure 2.16. The process is summarized in Figure 2.17. We will now learn how the x-ray diffraction pattern, $F(X)$, comes about and how the expected intensities in the diffraction pattern are calculated to solve Equation 2.19, i.e., how to reconstitute the image that led to the measured x-ray diffraction pattern.

Bragg's Law

In 1913 W.H. and W.L. Bragg (a father and son team) proposed that the condition for constructive

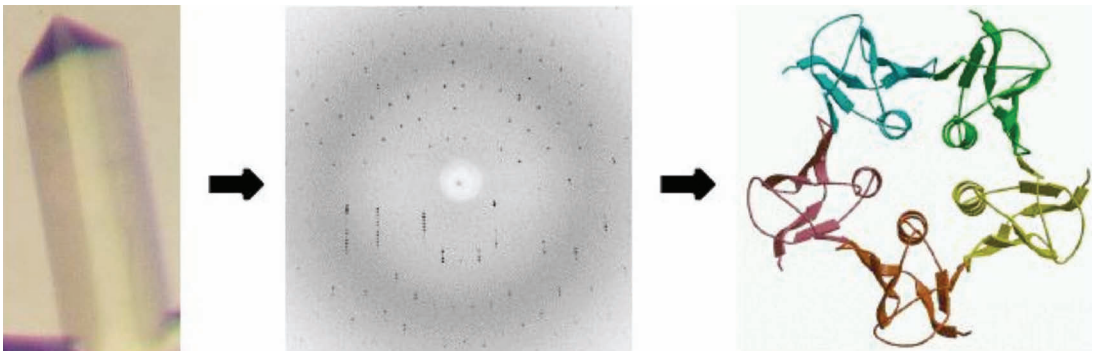


FIGURE 2.17 With x-rays, we can detect diffraction from molecules, but we have to use a computer to reassemble the electron density/molecular structure image.

specular reflection of x-rays from a set of crystal planes separated by a distance d_{hkl} could be represented as:

$$2d_{hkl} \sin \theta = n\lambda \tag{2.20}$$

This expression basically tells us that constructive interference of waves reflected by successive crystal planes occurs whenever the path difference ($2d_{hkl}\sin\theta$) is an integral multiple (n) of the wavelength λ. Also, for each (hkl) family of planes, x-rays will only diffract at one angle θ. The integer n is known as the order of the corresponding reflection. Because Bragg reflection can only occur for $\lambda \leq 2d$, one needs x-rays with wavelengths in the Ångstrom range to resolve crystal planes. The Bragg equation is easily derived from an inspection of Figure 2.18. Bragg's law is a result of the periodicity of the lattice with the atoms in the crystal basis controlling the relative intensity of the various orders (n) of diffraction from a set of parallel (hkl) planes. This basic equation is the starting point for understanding crystal diffraction of x-rays, electrons, neutrons, and any other particles that have a de Broglie wavelength

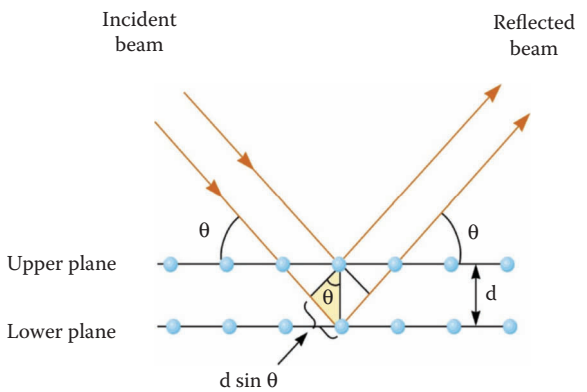


FIGURE 2.18 Schematic used to derive the Bragg equation.

(Chapter 3) less than the interatomic spacing. Although the reflection from each plane is specular, only for certain values of θ will the reflections from all planes add up in phase to give a strong reflected beam. Each plane reflects only 10^{-3} to 10^{-5} of the incident radiation, i.e., it is not a perfect reflector. Hence, 10^3 to 10^5 planes contribute to the formation of the Bragg-reflected beam in a perfect crystal.

The composition of the basis determines the relative intensity of the various orders of diffraction.

Laue Equations

In 1912 von Laue predicted that diffraction patterns of x-rays on crystals would be entirely analogous to the diffraction of light by an optical grating.* In the von Laue approach there is no ad hoc assumption of specular reflection, as in the case of Bragg's law. Instead, this more general approach considers a crystal as composed of sets of ions or atoms at the sites of a Bravais lattice that reradiate the incoming x-rays. For both optical gratings and crystals only the repeat distances of the periodic structure and the wavelength of the radiation determine the diffraction angles.

Let us inquire first into the interference conditions for waves originating from different but identical atoms in a single row—the one-dimensional diffraction case. The scattering atoms in a line form secondary, coherent x-ray sources (scattering from two atoms is shown in Figure 2.19).

* We will encounter the diffraction of light by an optical grating again in Volume II, Chapter 1 on photolithography, where we discuss patterning of a photoresist with UV light, using a mask with a grating structure on it.

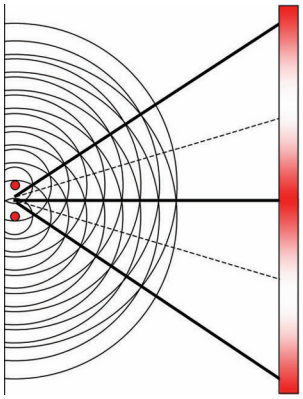


FIGURE 2.19 Two scattering atoms act as coherent secondary sources.

Constructive interference will occur in a direction such that contributions from each lattice point differ in phase by 2π. This is illustrated for the scattering of an incident x-ray beam by a row of identical atoms with lattice spacing $\mathbf{a}_1$ in Figure 2.20. The direction of the incident beam is indicated by wave vector $\mathbf{k}_0$ or the angle α_0, and the scattered beam is specified by the direction of $\mathbf{k}$ or the angle α. Because we assume elastic scattering, the two wave vectors $\mathbf{k}_0$ and $\mathbf{k}$ have the same magnitude, i.e., $2\pi/\lambda$ but with differing direction. A plane wave* $e^{ik.r}$ is constant in a plane perpendicular to $\mathbf{k}$ and is periodic parallel to it, with a wavelength $\lambda = 2\pi/\mathbf{k}$ (see Appendix 2A). The path difference $A_1B - A_2C$ in Figure 2.20 must equal $e\lambda$ with e = 0, 1, 2, 3, …. For a fixed incident x-ray with wavelength λ and direction $\mathbf{k}$, and an integer value of e, there is only one possible scattering angle α defining a cone of rays drawn about a line through the lattice points (see Figure 2.20). Because crystals are periodic in three directions, the Laue equations in 3D are then:

$$\begin{aligned} \mathbf{a}_1(\cos\alpha - \cos\alpha_0) &= e\lambda \\ \mathbf{a}_2(\cos\beta - \cos\beta_0) &= f\lambda \\ \mathbf{a}_3(\cos\gamma - \cos\gamma_0) &= g\lambda \end{aligned} \qquad (2.21)$$

For constructive interference from a three-dimensional lattice to occur, the three equations above must all be satisfied simultaneously, i.e., six angles α, β, γ, a, α_0, β_0, and γ_0; three lattice lengths $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$; and three integers (e, f, and g) are fixed. Multiplying both sides of Equation 2.21 with $2\pi/\lambda$ and rewriting the expression in vector notation we obtain:

$$\begin{aligned} \mathbf{a}_1 \cdot (\mathbf{k} - \mathbf{k}_0) &= 2\pi e \\ \mathbf{a}_2 \cdot (\mathbf{k} - \mathbf{k}_0) &= 2\pi f \\ \mathbf{a}_3 \cdot (\mathbf{k} - \mathbf{k}_0) &= 2\pi g \end{aligned} \qquad (2.22)$$

with $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ being the primitive vectors of the crystal lattice. When two of the conditions in Equation 2.22 are met, one entire plane array scatters in phase. This is depicted in Figure 2.21, where two cones of allowed diffracted rays are depicted. The two conditions are met simultaneously only in two directions along which the cones intersect. To satisfy all three Laue equations simultaneously, the diffracted beam can only have one allowed direction because three cones can mutually intersect along only one line.

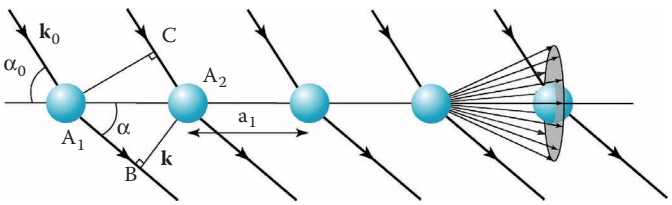


FIGURE 2.20 Scattering of an incident x-ray beam (incident direction is $\mathbf{k}_0$) by a row of identical atoms with lattice spacing $\mathbf{a}_1$. The scattered beam is specified by the direction $\mathbf{k}$. The path difference $A_1B - A_2C$ must equal $e\lambda$, with e = 0,1,2,3,.... (Drawing by Mr. Chengwu Deng.)

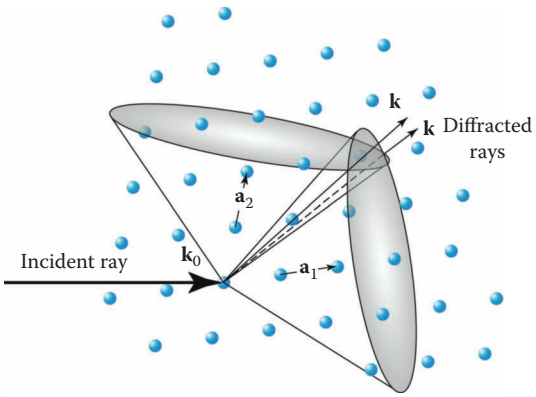


FIGURE 2.21 Each Laue condition produces a cone of allowed rays. In a plane array the entire plane scatters in phase in two directions. These two directions are along the intersection of the two cones. (Drawing by Mr. Chengwu Deng.)

* The complex exponential representation for periodic functions is convenient for adding waves, taking derivatives of wave functions, and so on. It is equivalent to a linear combination of a sine and cosine function, because $e^{i\theta} = \cos\theta + i\sin\theta$ (Euler) (see also Appendix 2A).

FIGURE 2.22 Max von Laue (1897–1960).

If we further define a vector $\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$, Equation 2.22 simplifies to:

$$\begin{aligned} \mathbf{a}_1 \cdot \Delta\mathbf{k} &= 2\pi e \\ \mathbf{a}_2 \cdot \Delta\mathbf{k} &= 2\pi f \\ \mathbf{a}_3 \cdot \Delta\mathbf{k} &= 2\pi g \end{aligned} \qquad (2.23)$$

Dealing with 12 variables for each reflection simultaneously [six angles (α, β, γ, a, α_0, β_0, and γ_0), three lattice lengths ($\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$), and three integers (e, f, and g)] is a handful; this is the main reason why the Laue equations are rarely referred to directly, and a simpler representation is used instead. The reflecting conditions can indeed be described more simply by the Bragg equation. Historically, von Laue (Figure 2.22) developed his equations first; it was one year after his work that the father and son team William Henry and William Lawrence Bragg (Figure 2.23) introduced

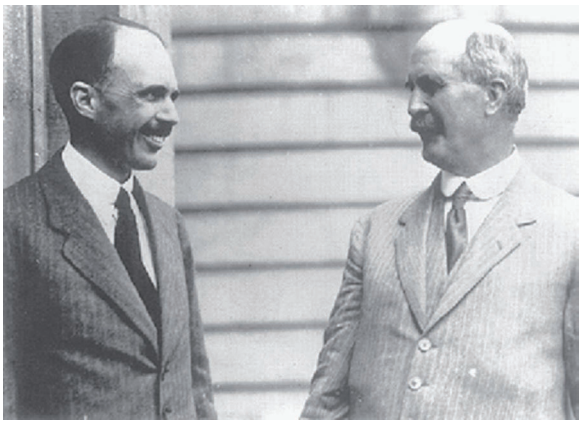


FIGURE 2.23 Father and son Bragg: Sir William Henry and William Lawrence Bragg.

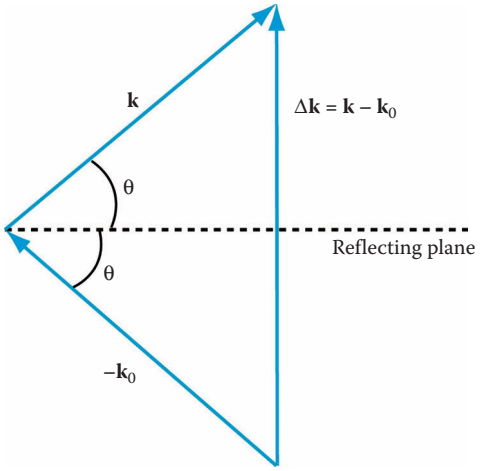


FIGURE 2.24 In case of mirror-like Bragg reflection, the vector **Δk**, the summation of the unit vectors representing incoming ($\mathbf{k}_0$) and reflected rays (**k**), is normal to the plane that intersects the 2θ angle between them. (Drawing by Mr. Chengwu Deng.)

the simpler Bragg's law. Max von Laue and the Braggs received the Nobel Prize in Physics in 1914 and 1915, respectively.

Further below we will learn that constructive interference of diffracted x-rays will occur provided that the change in wave vector, $\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$, is a vector of the reciprocal lattice.

Bragg's law is equivalent to the Laue equations in one dimension as can be appreciated from an inspection of Figures 2.24 and 2.25, where we use a two-dimensional crystal for simplicity. Suppose that vector $\Delta\mathbf{k}$ in Figure 2.24 satisfies the Laue condition; because incident and scattered waves have the same magnitude (elastic scattering), it follows that incoming ($\mathbf{k}_0$) and reflected rays (**k**) make the same angle θ with the plane perpendicular to $\Delta\mathbf{k}$.

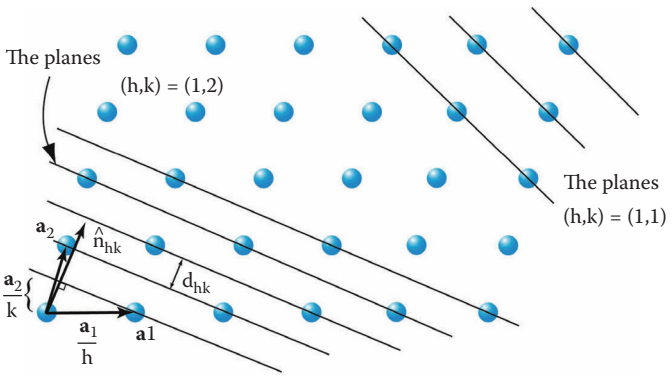


FIGURE 2.25 Connecting Bragg's law with Laue equations and Miller indices. (Drawing by Mr. Chengwu Deng.)

The magnitude of vector $\Delta\mathbf{k}$, from Figure 2.24, is then given as:

$$|\Delta\mathbf{k}| = 2\mathbf{k}\sin\theta \qquad (2.24)$$

We now derive the relation between the reflecting planes to which $\Delta\mathbf{k}$ is normal and the lattice planes with a spacing d_{hkl} (see Figure 2.25 and Bragg's law in Equation 2.20). The normal unit vector $\hat{\mathbf{n}}_{hk}$ and the interplanar spacing d_{hk} in Figure 2.25 characterize the crystal planes (*hk*). From Equation 2.23 we deduce that the direction cosines of $\Delta\mathbf{k}$, with respect to the crystallographic axes, are proportional to e/a_1, f/a_2, and g/a_3 or:

$$e/a_1{:}f/a_2{:}g/a_3 \qquad (2.25)$$

From the definition of the Miller indices, an (*hkl*) plane intersects the crystallographic axes at the points a_1/h, a_2/k, and a_3/l, and the unit vector $\hat{\mathbf{n}}_{hkl}$, normal to the (*hkl*) plane, has direction cosines proportional to:

$$h/a_1,\ k/a_2,\ \text{and}\ l/a_3 \qquad (2.26)$$

Comparing Equations 2.25 and 2.26 we see that $\Delta\mathbf{k}$ and the unit normal vector $\hat{\mathbf{n}}_{hkl}$ have the same directions; all that is required is that e = nh, f = nk, and g = nl, where *n* is a constant. The factor *n* is the largest common factor of the integers *e*, *f*, and *g* and is itself an integer. From the above, Laue's equations can also be interpreted as reflection from the *h,k,l* planes. From Figure 2.25 it can be seen that the spacing between the (*hk*) planes, and by extension between (*hkl*) planes, is given as*:

$$d_{hkl} = \frac{\hat{\mathbf{n}}_{hkl}\cdot\mathbf{a}_1}{h} = \frac{\hat{\mathbf{n}}_{hkl}\cdot\mathbf{a}_2}{k} = \frac{\hat{\mathbf{n}}_{hkl}\cdot\mathbf{a}_3}{l} \qquad (2.27)$$

Because $\Delta\mathbf{k}$ is in the direction of the normal $\hat{\mathbf{n}}_{hkl}$ and comes with a magnitude given by Equation 2.24, we obtain Bragg's law from the Laue's equations as:

$$\mathbf{a}_1\cdot\Delta\mathbf{k} = \mathbf{a}_1\cdot\hat{\mathbf{n}}_{hkl}\,2k\sin\theta = 2\pi e \qquad (2.28)$$

or:

$$hd_{hkl}\frac{4\pi}{\lambda}\sin\theta = 2\pi e \qquad (2.29)$$

* Notice that in the case of a cubic crystal, Equation 2.27 can be simplified to give the distance between planes as in Equation 2.12.

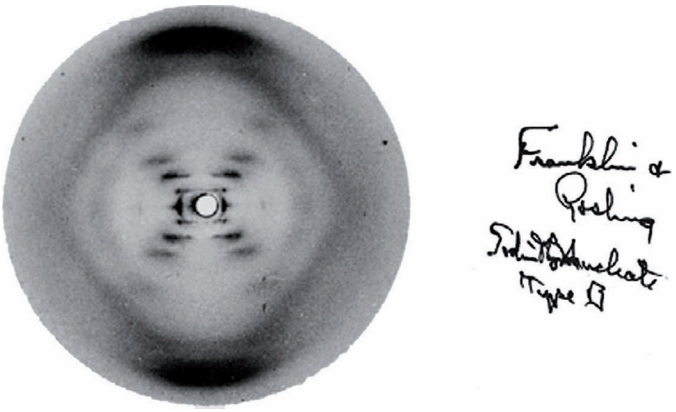


FIGURE 2.26 Sodium deoxyribose nucleate from calf thymus. (Structure B, Photo 51, taken by Rosalind E. Franklin and R.G. Gosling.) Linus Pauling's annotations are to the right of the photo (May 2, 1952).

and with e = nh:

$$2d_{hkl}\sin\theta = n\lambda \qquad (2.30)$$

In the Bragg equation we treat x-ray diffraction from a crystal as a reflection from reciprocal lattice planes rather than scattering from atoms. This construction has fewer variables than the Laue equations because reflections are wholly represented in two dimensions only.

In Figure 2.26 we reproduce what is possibly the most famous x-ray diffraction photograph. It was this photograph—photo 51 of DNA taken by Rosalind Franklin and R.G. Gosling—that convinced Watson and Crick that the DNA molecule was helical. The discovery of the double helix followed soon after, as well as an enduring controversy that Franklin probably did not get the credit she deserved in the elucidation of the DNA structure.

X-Ray Intensity and Structure Factor F(hkl)

So far we have considered only the condition for diffraction from simple lattices for which only corner points of the unit cell are occupied. The intensity of a beam diffracted from an actual crystal will depend on the grouping of atoms in the unit cell and on the scattering power of these atoms. In this section we discuss the intensity and phase of the diffracted rays and the structure factor, *F*(*hkl*). If we treat the incident x-ray waves as plane waves and the atoms as ideal point scatterers, the scattered waves are spherical waves (Figure 2.19) close to the source, i.e., near-field or Fresnel diffraction patterns form at finite

distances from the crystal and Fraunhofer diffraction patterns at infinity (far-field). In a typical x-ray diffraction experiment, a Fraunhofer diffraction pattern is registered at about 50 to 150 mm behind the specimen. This distance is relatively large compared to the size of the diffracting crystal unit cell (~1–50 nm in dimension) and the wavelength of the incident radiation (typically 1.54178 Å for Cu Kα and 0.71073 Mo Kα). Thus, x-ray diffraction methods provide a direct way to display the decomposition of x-rays in component waves (frequencies), and for this reason x-ray diffraction may be called *spatial frequency spectrum* analysis or *harmonic* analysis. Near-field and far-field optics are compared in more detail in Chapter 5.

The number of scattered x-ray photons picked up by the detector results in an intensity proportional to the square of the amplitude (peak height) of the diffracted waves. The scattering intensity depends on the number and distribution of electrons associated with the scattering atoms in the basis, i.e., the structure factor $F(hkl)$. The phase of the diffracted rays is the relative time of arrival of the scattered radiation at a particular point in space, say at the emulsion of a photographic film. The phase information is lost when the x-ray diffraction pattern is recorded on the film—one could say that the film integrates intensity over time—in other words, we cannot measure x-ray phases directly. When using a lens such as in a microscope, light first strikes the imaged object and is diffracted in various directions. The lens then collects the diffracted rays and reassembles them to form an image. The *phase problem* is a major concern in structure analysis as we need both intensity and phase to feed into the Fourier transform. Today several techniques exist to regenerate the *lost phase* information of x-ray diffraction, but the topic falls outside the scope of this book. We need to extract the structure factor $F(hkl)$ from the intensities and phase of the diffraction spots and then do an inverse Fourier transform (T^{-1}) to obtain the crystal structure/electron density function, $\rho(x, y, z)$ (see Figure 2.17).

We now mathematically derive the intensity profile of x-rays scattered from a crystal. The result, as Laue predicted, is the same as for visible light diffracted from an optical grating. When an incident

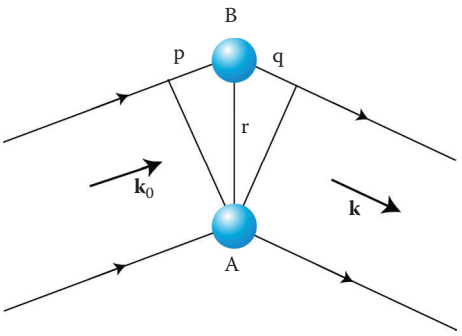


FIGURE 2.27 Scattering of x-rays from two nearby atoms A and B with identical scattering density. (Drawing by Mr. Chengwu Deng.)

x-ray beam travels inside a crystal, we assume that the beam is not much influenced by the presence of the crystal; in other words, the refractive index for x-rays is close to unity, and there is not much loss of energy from the beam through scattering, i.e., elastic scattering dominates!

With reference to Figure 2.27, we assume two parallel plane x-ray waves of wavelength λ and frequency ν (hence velocity $c = \lambda\nu$), scattered elastically from two nearby atoms A and B of identical scattering density. The wave vector for the incoming wave is $\mathbf{k}_0$ and that of the diffracted beam is $\mathbf{k}$. Because we assume elastic scattering:

$$|\mathbf{k}_0| = |\mathbf{k}| = k = 2\pi/\lambda \tag{2.31}$$

Scattering atom A is at the origin, and scattering atom B is at a distance $\mathbf{r}$ away from the origin. The path difference, or phase factor, between the waves can be calculated from Figure 2.27 as:

$$p + q = \mathbf{r}\cdot(\mathbf{k} - \mathbf{k}_0) = \mathbf{r}\cdot\Delta\mathbf{k} \tag{2.32}$$

The equations for the wave amplitudes are (for those readers less familiar with wave equations, consult Appendix 2A for more details):

$$\psi_A(x,t) = \frac{Af}{l_A} e^{i(\mathbf{k}_0\cdot\mathbf{l}_A - \omega\cdot t)} \tag{2.33}$$

and:

$$\psi_B(x,t) = \frac{Af}{l_B} e^{i(\mathbf{k}_0\cdot\mathbf{l}_B - \omega t + \mathbf{r}\cdot\Delta k)} \tag{2.34}$$

where A is the amplitude and f is the atomic scattering factor defined as the ratio of the amplitude of an electromagnetic wave scattered by atom A and that

of a wave scattered by a free electron. The position of the detector with respect to atom A is given by l_A, and the position of the detector relative to atom B is l_B. The product $\mathbf{r}\cdot\Delta\mathbf{k}$ is the phase factor we calculated in Equation 2.32. Generalizing, an x-ray wave scattered from the j_{th} atom in a crystal is:

$$\Psi_j(x,t) = \frac{Af_j}{l_j} e^{i(\mathbf{k}_0\cdot\mathbf{l}_j - \omega t + \mathbf{r}_j\cdot\Delta\mathbf{k})} \tag{2.35}$$

where $\mathbf{r}_j$ is the position of scatterer j relative to scatterer A, and l_j is the position of the detector with respect to scatterer j. The total scattered wave amplitude at the detector is the sum of all the contributing atoms:

$$\begin{aligned}\Psi(x,t) &= \sum_{\text{all atoms}} \frac{Af_j}{l_j} e^{i(\mathbf{k}_0\cdot\mathbf{l}_j - \omega t)} e^{i(\mathbf{r}_j\cdot\Delta\mathbf{k})} \\ &\cong \frac{A}{L} e^{i(\mathbf{k}_0\cdot\mathbf{L} - \omega t)} \sum_{\text{all atoms}} f_j e^{i(\mathbf{r}_j\cdot\Delta\mathbf{k})}\end{aligned} \tag{2.36}$$

For a small sample, the distances l_j are all about the same (the crystal is small compared to the distance between it and the detector), so in Equation 2.36 we can replace l_j with L.

Constructive and destructive interference between the scattered waves that reach the detector is the result of the sum of all scatterers and the scattering vector, $\Delta\mathbf{k}$, and determines where the detector should be put. The term $F(hkl)$:

$$F(hkl) = \sum_{\text{all atoms}} f_j e^{i(\mathbf{r}_j\cdot\Delta\mathbf{k})} \tag{2.37}$$

from the expression 2.36 for the wave amplitude, is the geometrical structure factor. It is defined, in analogy with the atomic scattering factor, as the ratio of the amplitude of the wave scattered by all atoms in a unit cell and that scattered by a free electron for the same incident beam. It incorporates the scattering of all atoms of the unit cell and sums up the extent to which interference of the waves scattered from atoms within the basis diminishes the intensity of the diffraction peaks. So even if the Laue condition is met, the structure factor may be zero and no diffraction will be observed. In the case that there is only one type of atom in the basis, the atomic scattering factor f_j disappears ($f_j = 1$)—identical atoms have identical scattering factors.

Assume now a crystal with base vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ and a total number of atoms along each axis of M, N, and P, respectively, and also accept that there is only a single atom at each lattice point (i.e., $f_j = 1$), then the amplitude of the total wave will be proportional to:

$$\begin{aligned}\Psi(x,t) \propto F(h,k,l) &= \sum_{\text{all atoms}} e^{i(\mathbf{r}_j\cdot\Delta\mathbf{k})} \\ &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{p=0}^{P-1} e^{i[(m\mathbf{a}_1 + n\mathbf{a}_2 + p\mathbf{a}_3)\cdot\Delta\mathbf{k}]}\end{aligned}$$

after rearrangement we obtain:

$$\Psi(x,t) \propto F(h,k,l) = \sum_{m=0}^{M-1} e^{im\mathbf{a}_1\cdot\Delta\mathbf{k}} \sum_{n=0}^{N-1} e^{in\mathbf{a}_2\cdot\Delta\mathbf{k}} \sum_{p=0}^{P-1} e^{ip\mathbf{a}_3\cdot\Delta\mathbf{k}} \tag{2.38}$$

The intensity of the scattered wave is the square of the wave amplitude, and taking the value of one of the sums in Equation 2.38 for a crystal of dimension $M\mathbf{a}_1$ in the direction of $\mathbf{a}_1$, we obtain:

$$\sum_{m=0}^{M-1} e^{im\mathbf{a}_1\cdot\Delta\mathbf{k}} = \frac{1 - e^{iM(\mathbf{a}_1\cdot\Delta\mathbf{k})}}{1 - e^{i(\mathbf{a}_1\cdot\Delta\mathbf{k})}}$$

$$\text{and } I \propto |\Psi(x,t)|^2 \propto \frac{[1 - e^{-iM(\mathbf{a}_1\cdot\Delta\mathbf{k})}][1 - e^{iM(\mathbf{a}_1\cdot\Delta\mathbf{k})}]}{[1 - e^{-i(\mathbf{a}_1\cdot\Delta\mathbf{k})}][1 - e^{i(\mathbf{a}_1\cdot\Delta\mathbf{k})}]}$$

which simplifies to

$$I \propto \frac{2 - e^{-iM(\mathbf{a}_1\cdot\Delta\mathbf{k})} - e^{iM(\mathbf{a}_1\cdot\Delta\mathbf{k})}]}{[2 - e^{-i(\mathbf{a}_1\cdot\Delta\mathbf{k})} - e^{i(\mathbf{a}_1\cdot\Delta\mathbf{k})}]} = \frac{2 - 2\cos(M\mathbf{a}_1\cdot\Delta\mathbf{k})}{2 - 2\cos(\mathbf{a}_1\cdot\Delta\mathbf{k})}$$

$$I \propto \frac{1 - \cos(M\mathbf{a}_1\cdot\Delta\mathbf{k})}{1 - \cos(\mathbf{a}_1\cdot\Delta\mathbf{k})} \quad \text{and with the identity}$$

$$\cos 2x = \cos^2 x - \sin^2 x = 1 - 2\sin^2 x$$

$$I \propto \frac{\sin^2\left(\frac{1}{2}M\mathbf{a}_1\cdot\Delta\mathbf{k}\right)}{\sin^2\left(\frac{1}{2}\mathbf{a}_1\cdot\Delta\mathbf{k}\right)} \tag{2.39}$$

This is the same result as the light intensity profile expected from an M-slit diffraction grating (see Chapter 5 on photonics). If M is large ($\sim 10^8$ for a macroscopic crystal), it has very narrow, intense peaks. Between the peaks the intensity is essentially zero. In Figure 2.28 we have plotted $y = \sin^2 Mx/\sin^2 x$.

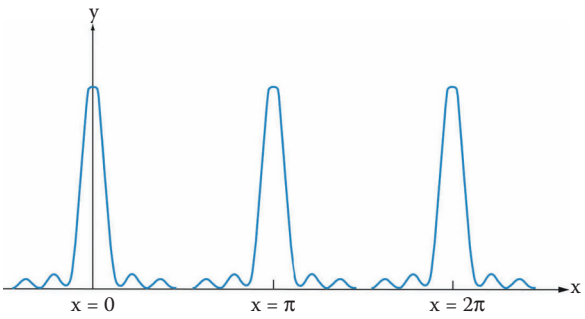


FIGURE 2.28 Graphical presentation of $y = \sin^2 Mx/\sin^2 x$. The width of the peaks and the prominence of the ripples are inversely proportional to *M*.

This function is virtually zero except at the points where $x = n\pi$ (*n* is an integer including zero), where it rises to the maximum value of M^2. The width of the peaks and the prominence of the ripples are inversely proportional to *M*.

Remember that there are three sums in Equation 2.38. For simplicity we only evaluated one sum to calculate the intensity in Equation 2.39. The total intensity equals:

$$I \propto \frac{\sin^2\left(\frac{1}{2}M\mathbf{a}_1 \cdot \Delta\mathbf{k}\right)}{\sin^2\left(\frac{1}{2}\mathbf{a}_1 \cdot \Delta\mathbf{k}\right)} \times \frac{\sin^2\left(\frac{1}{2}N\mathbf{a}_2 \cdot \Delta\mathbf{k}\right)}{\sin^2\left(\frac{1}{2}\mathbf{a}_2 \cdot \Delta\mathbf{k}\right)} \times \frac{\sin^2\left(\frac{1}{2}P\mathbf{a}_3 \cdot \Delta\mathbf{k}\right)}{\sin^2\left(\frac{1}{2}\mathbf{a}_3 \cdot \Delta\mathbf{k}\right)} \tag{2.40}$$

so that the diffracted intensity will equal zero unless all three quotients in Equation 2.40 take on their maximum values at the same time. This means that the three arguments of the sine terms in the denominators must be simultaneously equal to integer multiples of 2π, or the peaks occur only when:

$$\mathbf{a}_1 \cdot \Delta\mathbf{k} = 2\pi e$$

$$\mathbf{a}_2 \cdot \Delta\mathbf{k} = 2\pi f$$

$$\mathbf{a}_3 \cdot \Delta\mathbf{k} = 2\pi g$$

These are, of course, the familiar Laue equations. As we will see below, to solve for the wave vector $\Delta\mathbf{k}$ it is very convenient to introduce the concept of a reciprocal lattice so that the set of all wave vectors $\Delta\mathbf{k}$ yields plane waves with the periodicity of a given Bravais lattice.

From the above, the intensity of the x-ray diffraction is:

$$I \propto |F(X)|^2 = \left|\sum_{\text{all atoms}} f_j e^{i(\mathbf{r}_j \cdot \Delta\mathbf{k})}\right|^2 \tag{2.41}$$

So what we measure from the x-ray film is intensity *I*, and the dependence of that intensity on the atomic position follows from the Fourier expansion of the electron density function in Equation 2.17:

$$\rho(x,y,z) = \sum_h \sum_k \sum_l F(hkl) e^{i(\mathbf{r}_j \cdot \Delta\mathbf{k})} \tag{2.42}$$

where $F(hkl)$ is the coefficient to be determined, and *h*, *k*, and *l* are integers over which the series is summed. Because of the 3D periodicity, a triple summation is required:

$$F(hkl) = \int_0^1 \int_0^1 \int_0^1 \rho(x,y,z) e^{i(\mathbf{r}_j \cdot \Delta\mathbf{k})} dx dy dz \tag{2.43}$$

If we knew all the $F(hkl)$ values, we could calculate the electron density, and vice versa. Unfortunately, as remarked above, Equation 2.42 requires values of $F(hkl)$, but the measured intensities only give us $|F(hkl)|^2$. The apparent impasse is known as the *phase problem* and arises because we need to know both the amplitude and the phase of the diffracted waves to compute the inverse Fourier transform.

Reciprocal Space, Fourier Space, k-Space, or Momentum Space

Introduction

The crystal lattice discussed above is a lattice in real space. Practice has shown the usefulness of defining a lattice in reciprocal space for the interpretation of diffraction patterns of single crystals, the simple reason being that a diffraction pattern reveals the reciprocal lattice. We did already consider some reciprocal features when we introduced the Miller indices, which are derived as the reciprocal of the unit cell intercepts. Because the lattice distances

between planes are proportional to the reciprocal of the distances in the real crystal (i.e., they are proportional to $1/d_{hkl}$), the array is called a reciprocal lattice. We will learn in this section that the condition for nonzero intensity in scattered x-rays is that the scattering vector $\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$ is a translation vector, $\mathbf{G}_{hkl}$, of the reciprocal lattice, i.e., $\Delta\mathbf{k} = \mathbf{G}_{hkl}$. One can describe a reciprocal lattice the same way we describe a real one, but one must keep in mind that in one case the points are the position of real objects (atoms or the base), whereas in the other case they mark the positions of abstract points—magnitude and direction of momentum. Reciprocal space, also called Fourier space, **k**-space, or momentum space, plays a fundamental role in most analytical studies of periodic structures. After the introduction of the reciprocal lattice we will discuss the conditions for x-ray diffraction in terms of such a lattice and end with an introduction to Brillouin zones. The concepts introduced here are not only needed to understand x-ray diffraction better, they also are required preparation for Chapter 3, where we deal with band theory of solids.

The Reciprocal Lattice

Graphical Presentation of the Reciprocal Lattice

Before introducing a reciprocal lattice mathematically, let us learn how to draw one as we demonstrate in Figure 2.29. In this example we start with the two-dimensional unit cell of a monoclinic crystal with $\mathbf{a}_1$ and $\mathbf{a}_2$ in the plane of the paper. The figure also shows the edges of four (*hkl*) planes: (100), (110), (120), and (010). These planes all are perpendicular to the face of the paper. To construct the reciprocal lattice we draw from a common origin a normal to each plane. Next we place a point on the normal to each plane (*hkl*) at a distance $2\pi/d_{hkl}$ from the origin. Each of the points thus obtained maintains the important features of the stack of parallel planes it represents. The direction of the point from the origin preserves the orientation of the planes, and the distance of the point from the origin preserves the interplanar distance. A doubling of the periodicity in real space will produce twice as many diffraction spots in reciprocal space. It is convenient to let the reciprocal lattice vector, $\mathbf{G}_{hkl}$, be 2π times the reciprocal of the

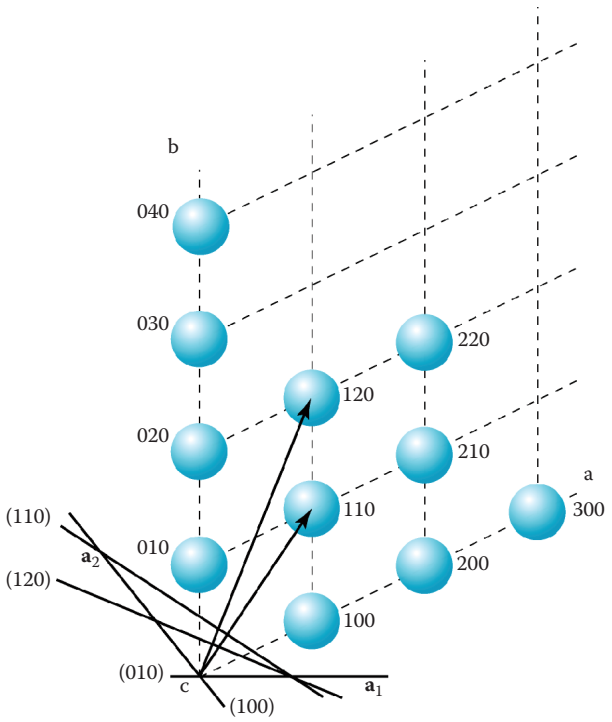


FIGURE 2.29 Graphical construction of the reciprocal lattice. (Drawing by Mr. Chengwu Deng.)

interplanar distance, d_{hkl}. This convention converts the units from periods per unit length to radians per unit length. This simplifies comparison of different periodic phenomena, e.g., a crystal lattice and light that interacts with it. For instance, $\Delta\mathbf{k}$, the wave vector, has an absolute value of $2\pi/\lambda$. If we choose this scaling factor, we are able to compare the two values directly, and all it really does in our drawing is to expand the size of the reciprocal lattice.

Definition of the Reciprocal Lattice

Now that we can draw a reciprocal lattice in 2D, let us define what a reciprocal lattice is. Imagine a set of points **R** constituting a Bravais lattice and a plane wave, $e^{i\mathbf{k}\cdot\mathbf{r}}$, interacting with that lattice. For most values of **k**, a plane wave will not have the periodicity of the Bravais lattice of points **R**. Only a very special set of the wave vectors, $\mathbf{G}_{hkl} = h\cdot\mathbf{a}_1^* + k\cdot\mathbf{a}_2^* + l\cdot\mathbf{a}_3^*$, of the reciprocal lattice satisfies that condition. Mathematically, $\mathbf{G}_{hkl}$ belongs to the reciprocal lattice of points **R** of the Bravais lattice, if the relation:

$$e^{i\mathbf{G}_{hkl}\cdot(\mathbf{r}+\mathbf{R})} = e^{i\mathbf{G}_{hkl}\cdot\mathbf{r}} \qquad (2.44)$$

holds for any **r**, and for all **R** in the Bravais lattice. From Equation 2.44 it follows that we can describe

the reciprocal lattice as the set of wave vectors $\mathbf{G}_{hkl}$ satisfying:

$$e^{i\mathbf{G}_{hkl}\cdot\mathbf{R}} = 1 \tag{2.45}$$

for all $\mathbf{R}$ in the Bravais lattice. Because lattice scatterers are displaced from one another by the lattice vectors $\mathbf{R}$, the condition that all scattered rays interfere constructively is that the Laue equations hold simultaneously for all values of Bravais lattice vectors $\mathbf{R}$, or:

$$\mathbf{R}\cdot\Delta\mathbf{k} = 2\pi m \tag{2.46}$$

with integer m. Because $\exp[i2\pi(\text{integer})] = 1$, this can be written in the equivalent form:

$$ei^{\mathbf{R}}\cdot\Delta^{\mathbf{k}} = 1 \tag{2.47}$$

We see from Equation 2.47 that if $\Delta\mathbf{k}$ is equal to any reciprocal lattice vector $\mathbf{G}_{hkl}$, then the Laue equations for wave diffraction are satisfied. The diffraction condition is thus simply:

$$\Delta\mathbf{k} = \mathbf{G}_{hkl} \tag{2.48}$$

or from Equation 2.23 (with e = h, f = k and g = l):

$$\begin{aligned}\mathbf{G}_{hkl}\cdot\mathbf{a}_1 &= 2\pi h\\ \mathbf{G}_{hkl}\cdot\mathbf{a}_2 &= 2\pi k\\ \mathbf{G}_{hkl}\cdot\mathbf{a}_3 &= 2\pi l\end{aligned} \tag{2.49}$$

The expression $\Delta\mathbf{k} = \mathbf{G}_{hkl}$ can be represented in the form of a vector triangle as illustrated in Figure 2.30. Because $\mathbf{k}$ and $\mathbf{k}_0$ are of equal length, i.e. $2\pi/\lambda$, the triangle O, O', O'' has two equal sides. The angle between $\mathbf{k}$ and $\mathbf{k}_0$ is 2θ, and the *hkl* plane dissects it.

In the two-dimensional reciprocal lattice shown at the bottom of Figure 2.31, the $\mathbf{G}_{hkl}$ vectors give the outline of the unit cell. Vector $\mathbf{a}_1^* = \mathbf{G}_{100}$ and $\mathbf{a}_3^* = \mathbf{G}_{001}$ and:

$$|\mathbf{a}_1^*| = \frac{2\pi}{d_{100}} \text{ and } |\mathbf{a}_3^*| = \frac{2\pi}{d_{001}} \tag{2.50}$$

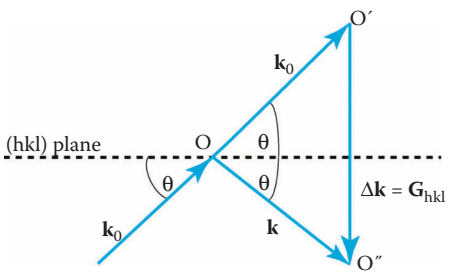


FIGURE 2.30 Vector triangle representation of $\Delta\mathbf{k} = \mathbf{G}_{hkl}$. (Drawing by Mr. Chengwu Deng.)

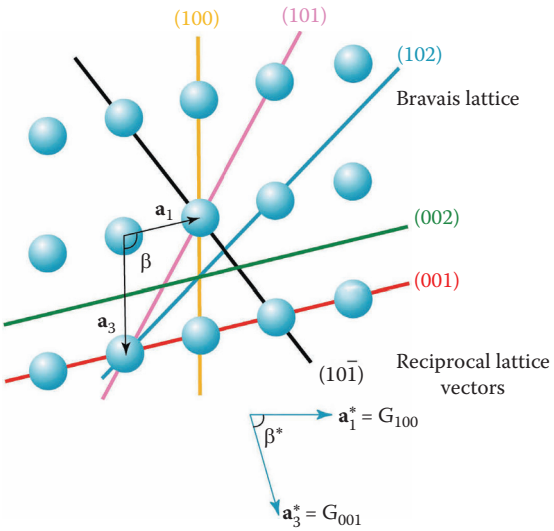


FIGURE 2.31 Reciprocal lattice vectors $a_1^* = G_{100}$ and $a_3^* = G_{001}$ in a monoclinic unit cell and their relation to the Bravais lattice. (Drawing by Mr. Chengwu Deng.)

The angle β^* between the reciprocal lattice vectors in Figure 2.31 is the complement of β in the Bravais lattice.

The components of any vector referred to the reciprocal lattice represent the Miller indices of a plane whose normal is along that vector, whereas the spacing of the plane is given by the inverse of the magnitude of that vector multiplied by 2π. For example, the reciprocal lattice vector $\mathbf{u}^* = [123]$ is normal to the planes with Miller indices (123) and has an interplanar spacing $2\pi/|\mathbf{u}^*|$.

Explicit Algorithm for Constructing the Reciprocal Lattice

We can write out the change in wave vector, $\mathbf{G}_{hkl} = \Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$, in the following expression:

$$\mathbf{G}_{hkl} = \Delta\mathbf{k} = e\mathbf{a}_1^* + f\mathbf{a}_2^* + g\mathbf{a}_3^* \tag{2.51}$$

where *e*, *f*, and *g* are the integers from the Laue equations (Equation 2.23) and $\mathbf{a}_1^*$, $\mathbf{a}_2^*$ and $\mathbf{a}_3^*$ are basis vectors of the reciprocal lattice to be determined. Equation 2.51 is a solution of the Laue Equation 2.23 if all of the following relations are satisfied:

$$\begin{aligned}&\mathbf{a}_1^*\cdot\mathbf{a}_1 = \mathbf{a}_2^*\cdot\mathbf{a}_2 = \mathbf{a}_3^*\cdot\mathbf{a}_3 = 2\pi\\ &\mathbf{a}_1^*\cdot\mathbf{a}_2 = \mathbf{a}_2^*\cdot\mathbf{a}_1 = \mathbf{a}_3^*\cdot\mathbf{a}_2\\ &= \mathbf{a}_1^*\cdot\mathbf{a}_3 = \mathbf{a}_2^*\cdot\mathbf{a}_3 = \mathbf{a}_3^*\cdot\mathbf{a}_1 = 0\end{aligned} \tag{2.52}$$

From Equation 2.52, $\mathbf{a}_1^*$ is perpendicular to primitive lattice vectors $\mathbf{a}_2$ and $\mathbf{a}_3$ of the direct lattice. A vector

TABLE 2.3 Bravais Lattice after Fourier Transform

Real Space	Fourier Space
Normals to the planes (vectors)	Points
Spacing between planes	2π/distance planes
λ is distance, wavelength	$2\pi/\lambda$ is momentum or wave number
First Brillouin zone (see below)	Wigner-Seitz cell

that is perpendicular to $\mathbf{a}_2$ and $\mathbf{a}_3$ is given by the vector product $\mathbf{a}_2 \times \mathbf{a}_3$. To construct $\mathbf{a}_1^*$ of the reciprocal lattice completely, we must normalize $\mathbf{a}_2 \times \mathbf{a}_3$ to satisfy the expression $\mathbf{a}_1^* \cdot \mathbf{a}_1 = 2\pi$. All the equations in 2.52 are thus satisfied when choosing:

$$\mathbf{a}_1^* \equiv \frac{2\pi \mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3}$$
$$\mathbf{a}_2^* \equiv \frac{2\pi \mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3}$$
$$\mathbf{a}_3^* \equiv \frac{2\pi \mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3} \qquad (2.53)$$

The denominators have been written the same way because of the property of the triple scalar product: $\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3 = \mathbf{a}_2 \cdot \mathbf{a}_3 \times \mathbf{a}_1 = \mathbf{a}_3 \cdot \mathbf{a}_1 \times \mathbf{a}_2$. The magnitude of this triple product is the volume of the primitive cell. In Table 2.3 we summarize the properties of a Bravais lattice before and after a Fourier transform.

The Ewald Construction

The Ewald construction is a way to visualize the Laue condition that the change in wave vector, $\Delta\mathbf{k}$, must be a vector of the reciprocal lattice. We reconsider for a moment the vector representation of the von Laue condition $\Delta\mathbf{k} = \mathbf{G}_{hkl}$ represented in the form of a vector triangle in Figure 2.30. Because $\mathbf{k}$ and $\mathbf{k}_0$ are of equal length ($= 2\pi/\lambda$) the triangle O, O', O'' has two equal sides, and we can draw a sphere with $|\mathbf{k}| = |\mathbf{k}_0| = 2\pi/\lambda$ as illustrated in Figure 2.32. The angle between $\mathbf{k}$ and $\mathbf{k}_0$ is 2θ, and the *hkl* plane dissects it. Diffraction of the incoming beam represented by the vector $\mathbf{k}_0$ giving the vector $\mathbf{k}$ may be thought of as reflection from the dotted line in this diagram (as in Bragg's Law). We will now see that with $\Delta\mathbf{k} = \mathbf{G}_{hkl}$ the diffraction vector corresponds to the distance between planes in reciprocal space. We thus superimpose an imaginary sphere of x-ray radiation on the reciprocal lattice as illustrated

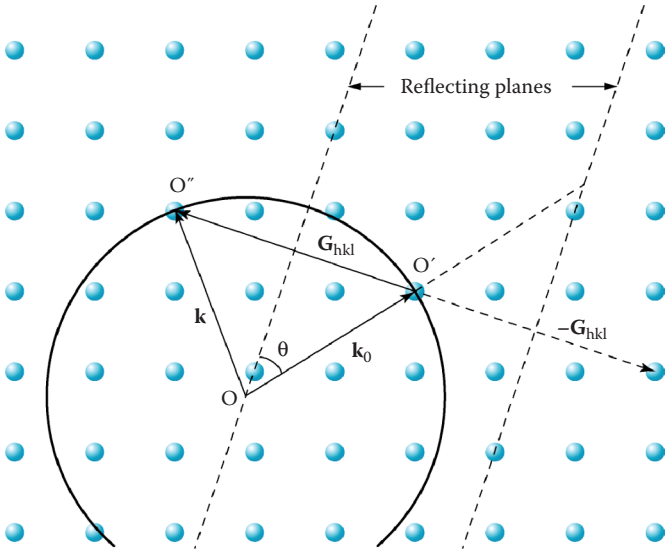


FIGURE 2.32 The reciprocal lattice and the geometry of diffraction clarified by the Ewald sphere. The sphere with center *O* intersects the reciprocal lattice center *O′*. (Drawing by Mr. Chengwu Deng.)

in Figure 2.32. Somewhat confusingly, one might consider two origins: *O*, which is the center of the sphere and may be considered as the position of the crystal, is the source of the secondary beam $\mathbf{k}$, and *O′* is the origin of reciprocal space, the origin of the diffraction vector $\Delta\mathbf{k} = \mathbf{G}_{hkl}$, and the center of the reciprocal lattice. As the crystal rotates, the reciprocal lattice rotates in exactly the same way. Any points in reciprocal space that intersect the surface of the sphere reveals where diffraction peaks will be observed if the structure factor is nonzero. In other words, reflection is only observed if the sphere intersects a point where $\Delta\mathbf{k} = \mathbf{G}_{hkl}$. The diffraction angle, θ, is then half the angle between the incident and diffracted wave vectors.

It is useful to think of the crystal at the center of the Ewald sphere (*O*) being linked to the center (origin) of the reciprocal lattice (*O′*) by something like a bicycle chain—the two "objects" rotate exactly in step with each other.

The Ewald construction also makes for a good link with the Brillouin zones discussed in the next section. Notice in Figure 2.32 that both $\mathbf{G}_{hkl}$ and $-\mathbf{G}_{hkl}$ are vectors in reciprocal space. It can be seen that a reflecting plane bisects the vector $\mathbf{G}_{hkl}$ at $\mathbf{G}_{hkl}/2$ and another reflecting plane cuts $-\mathbf{G}_{hkl}$ at $-\mathbf{G}_{hkl}/2$. The incident wave vector $\mathbf{k}_0$ starting from the Ewald circle's center *O* must terminate at the $-\mathbf{G}_{hkl}/2$ reflecting plane for

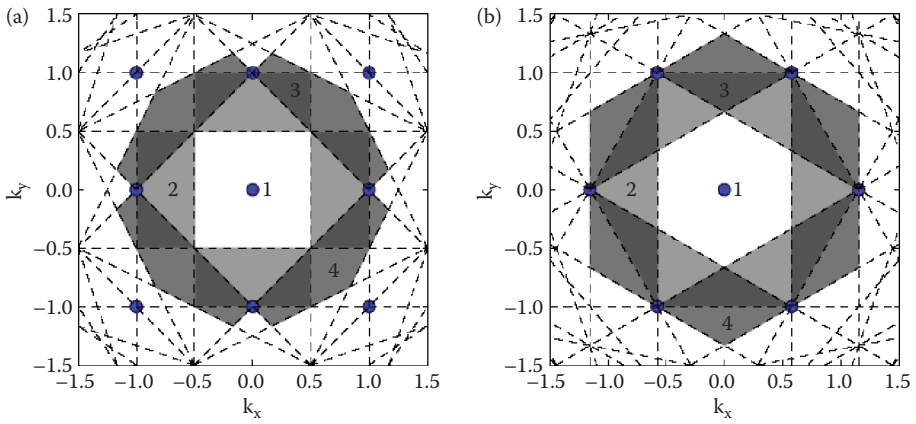


FIGURE 2.33 The Brillouin zones for (a) a square 2D lattice and (b) a triangular 2D lattice. The solid circles are the lattice points, and the dashed lines are the Bragg lines. The first four Brillouin zones are marked with different gray scales.

diffraction to occur. This observation gives rise to the idea of using the reciprocal lattice to construct Brillouin zones, the boundaries of which satisfy the Laue conditions for diffraction. Bragg planes bisect the lines joining the origin to points of the reciprocal lattice, and one can define the first Brillouin zone as the set of points that can be reached from the origin without crossing any other Bragg planes. Recall that this was also the way we learned how to construct Wigner-Seitz cells. So the first Brillouin zone is the Wigner-Seitz cell of the reciprocal lattice.

Brillouin Zones

Leon Brillouin (1889–1969) was a French-American physicist who was a professor at the Sorbonne, Collège de France, Wisconsin, Columbia, and Harvard. He also worked briefly at IBM. In the previous section we used the Ewald sphere (Figure 2.32) as a way to introduce the important concept of Brillouin zones. A Brillouin zone is defined as a Wigner-Seitz cell in the reciprocal lattice and gives a geometric interpretation of the diffraction condition. The Wigner-Seitz cell of the reciprocal lattice is the set of points laying closer to $\Delta\mathbf{k} = 0$ than to any other reciprocal lattice point. The Brillouin construction exhibits all wave vectors $\mathbf{k}$ that can be Bragg reflected by the crystal. The constructions divide the Fourier space into zones, out of which the first Brillouin zone is of greatest importance.

We have just seen that x-ray waves traveling in a crystal lattice undergo Bragg reflection at certain wave vectors. The first Brillouin zone is the set of points in $\mathbf{k}$-space that can be reached from the origin without crossing any Bragg reflection plane. The second Brillouin zone is the set of points that can be reached from the first zone by crossing only one Bragg plane. The nth Brillouin zone can be defined as the set of points that can be reached from the origin by crossing $n - 1$ Bragg planes, but no fewer. Each Brillouin zone is a primitive cell of the reciprocal lattice. In Figure 2.33 we illustrate the first four Brillouin zones in square and triangular two-dimensional Bravais lattices. For examples of 3D Wigner-Seitz cells, corresponding to Brillouin zones, see Figures 2.6 and 2.34.

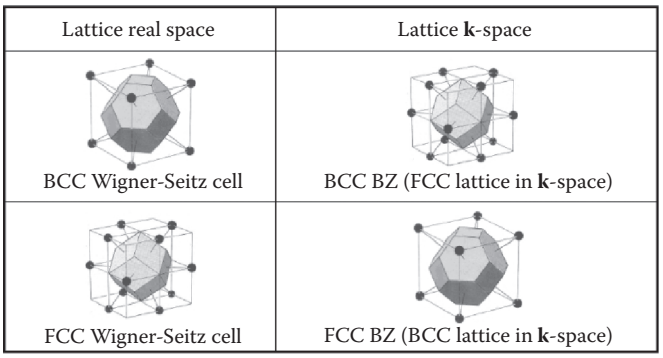


FIGURE 2.34 The Wigner-Seitz cell of BCC lattice in real space transforms to a Brillouin zone in an FCC lattice in reciprocal space, whereas the Wigner-Seitz cell of an FCC lattice transforms to a Brillouin zone of a BCC lattice in reciprocal space.

Nothing Is Perfect

Even a crystal having a mass of less than 0.1 g is likely to have more than 10^{23} ions, and it is unrealistic to

expect that the growth of such crystal from solution can lead to zero impurities or imperfections. Successive purification steps may remove impurities, but even in the purest crystal, thermodynamics predicts the existence of some structural intrinsic imperfections. There are at least four types of defects one distinguishes: 1) point defects where the irregularity in structure extends over only a few atoms in size (0D), 2) linear defects where the irregularity extends across a single row of atoms (1D), 3) planar defects with an irregularity extending across a plane of atoms (2D), and 4) volumetric defects with the irregularity taking place over 3D clusters of atoms. In addition, defects can be categorized as either intrinsic, where defects are induced because of physical laws, and extrinsic, where defects are present because of the environment and/or processing conditions.

A physical law imposing the presence of intrinsic defects in a crystal is the minimization of the Gibbs free energy (G). The Gibbs free energy is higher in a crystal without vacancies than one with vacancies (G = H – TS = minimum). This is because defects are energetically unfavorable but are entropically favorable. Formation of vacancies (n) does indeed increase the enthalpy H of the crystal because of the energy required to break bonds ($\Delta H = n\Delta H_f$), but vacancies also increase S of the crystal as a result of an increase in configurational entropy. The configurational entropy is given as:

$$S = k\ln W \tag{2.54}$$

where W is the number of microstates. If the number of atoms in the crystal is N and the number of vacancies is n, then the total number of sites is $n + N$, and the number of all possible microstates W may be calculated from:

$$W = \frac{(N+n)!}{n!N!} \tag{2.55}$$

and the increase in entropy ΔS is then given by:

$$\Delta S = k\ln W = k\ln\frac{(N+n)!}{n!N!} \tag{2.56}$$

and the total free energy change as:

$$\Delta G = n\Delta H_f - T\Delta S \tag{2.57}$$

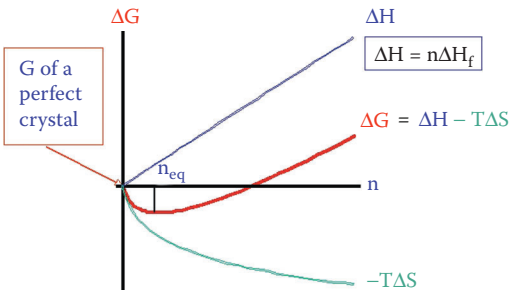


FIGURE 2.35 Change in Gibbs free energy G of a crystal as a result of the number of vacancies n.

This expression is plotted in Figure 2.35. It can be seen that for a crystal in equilibrium, vacancies are required to be present at any temperature above 0 K. The equilibrium concentration n_{eq} is calculated from:

$$\left|\frac{\partial \Delta G}{\partial n}\right|_{n=n_{eq}} = 0 \tag{2.58}$$

or we obtain:

$$\frac{n_{eq}}{N} = \exp\left(-\frac{\Delta H_f}{kT}\right) \tag{2.59}$$

so the n_{eq} should increase with increasing temperature. For Al, $\Delta H_f = 0.70$ eV/vacancy and for Ni, $\Delta H_f = 1.74$ eV/vacancy, leading to the values for n_{eq}/N at three different temperatures as shown in Table 2.4.

Defects, even in very small concentrations, can have a dramatic impact on the properties of a material. Actually, without defects solid-state electronic devices could not exist, metals would be much stronger, ceramics would be much tougher, and some crystals would have no color. Vacancies even make a small contribution to the thermal expansion of a crystal. Some commonly observed defects are summarized in Table 2.5; they are categorized here as point defects (0D), line defects (1D), and plane defects (2D).

The simplest sorts of defects are those based on points (0D). In Figure 2.36 we review point defects, including vacancies, interstitial atoms, small and

TABLE 2.4 n_{eq}/N Values for Al and Ni at Three Different Temperatures

n_{eq}/N	0 K	300 K	900 K
Al	0	$1.45\ 10^{-12}$	$1.12\ 10^{-4}$
Ni	0	$5.59\ 10^{-30}$	$1.78\ 10^{-10}$

Table 2.5 Common Defects in Crystals

Type of Imperfection	Description
Point defects:	
Interstitial	Extra atom in an interstitial site
Schottky defect	Atom missing from correct site
Frenkel defect	Atom displaced to interstitial site creating nearby vacancy
Line defects:	
Edge dislocations	Row of atoms marking edge of a crystallographic plane extending only part way in crystal
Screw dislocations	Row of atoms about which a normal crystallographic plane appears to be spiral
Plane defects:	
Lineage boundary	Boundary between two adjacent perfect regions in the same crystal that are slightly tilted with respect to each other
Grain boundary	Boundary between two crystals in a polycrystalline solid
Stacking fault	Boundary between two parts of a closest packing having alternate stacking sequences

large substitutional atoms, and the intrinsic Frenkel and Schottky defects.

A missing atom is a vacancy; a dissimilar atom in a nonlattice spot is an interstitial; and a different atom in a lattice position is substitutional. Vacancies are required in ionic solids, just like they are in metals, but the vacancies must be formed in such a way that the solid remains charge neutral. The two main ways

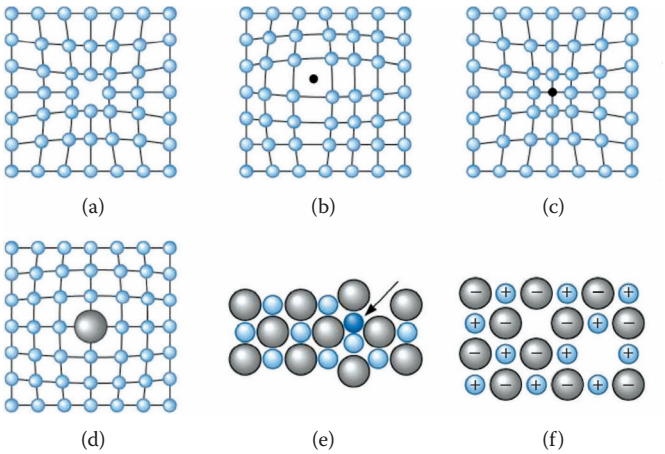


FIGURE 2.36 Point defects: vacancy (a), interstitial atom (b), small (c) and large (d) substitutional atom, Frenkel (e), and Schottky defect (f).

to create point defects in ionic solids without causing a charge imbalance are through correlated vacancies (Schottky defects, for the German who described them in 1930) and through correlated vacancy/interstitial pairs (Frenkel defects, for the Russian who first described them in 1924). Impurities, such as dopants in single-crystal Si (see Chapter 4), are atom(s) of a type that do not belong in the perfect crystal structure. An impurity atom in a pure crystal will generally raise the enthalpy (H) somewhat and increase the entropy (S = klnW) a lot. In other words, there will always be some impurities present for the same reason as vacancies form in a pure crystal. When doping a crystal we introduce impurities on purpose (see Chapter 4), perhaps turning the material into an extrinsic semiconductor. Because of defects, metal oxides may also act as semiconductors. Some nonstoichiometric solids are engineered to be n-type or p-type semiconductors. Nickel oxide (NiO) gains oxygen on heating in air, producing Ni^{3+} sites acting as electron traps, resulting in p-type semiconductor behavior. On the other hand, ZnO loses oxygen on heating, and the excess Zn metal atoms in the sample are ready to donate electrons, leading to n-type semiconductor behavior.

Color centers are imperfections in crystals that cause color. The simplest color center is found in sodium chloride, normally a colorless crystal. When sodium chloride is bombarded with high-energy radiation, a Cl^- can be ejected, creating a vacancy. Momentarily the crystal is no longer electrically neutral, and to regain stability, it grabs an available electron and sticks it in the vacancy previously occupied by the ejected Cl^-. With the electron replacing the ejected Cl^-, there are now equal numbers of positive and negative charges in the crystal, and the electron is held firmly in its site by the surrounding positively charged Na^+ ions. This process turns the colorless salt crystal into an orange/brown. These electrons are color centers, often referred to as F-centers, from the German word *Farben*, meaning color. The color center can also exist in an excited state, and the energy needed to reach that excited state is equal to the energy of a visible photon. The color center absorbs a "violet" photon, causing a jump to the excited state, and the crystal appears with the color orange/brown (the complement of violet). Analogous color

FIGURE 2.37 Color centers in some well-known minerals: (a) the Dresden green diamond, (b) smoky quartz, and (c) amethyst or violet quartz.

centers occur in several minerals. For example, a diamond with C vacancies (missing carbon atoms) absorbs light, and these centers lead to a green color (Figure 2.37). In some cases, impurities are involved in forming the color centers. Replacement of Al^{3+} for Si^{4+} in quartz gives rise to the color of smoky quartz (Figure 2.37). An iron impurity is responsible for the violet color in amethyst through the creation of a color center. The color center, not the iron impurity, is responsible for absorbing the "yellow" photon that makes amethyst violet (Figure 2.37).

Line defects are 1D imperfections in a crystal structure in which a row of atoms has a local structure that differs from the surrounding crystal. These defects are extrinsic because their presence is not required by thermodynamics. They are created by material processing conditions and by mechanical forces acting on the material. In a typical material, about 5 of every 100 million atoms (0.000005%) belong to a line defect. Line defects or dislocations have a dramatic impact on the mechanical properties of metals and some ceramics. Two Hungarians, Egon Orowan and Michael Polanyi, and an Englishman, G.I. Taylor, discovered dislocations in 1934. As we will see, dislocations really govern the mechanical properties of solids, and their discovery was a very important milestone in the material science field. There are two pure types of dislocations, edge dislocations and screw dislocations, but sometimes a mixed type is displayed. An edge dislocation is the simplest type of dislocation and can be viewed as an extra half-plane of atoms inserted into the crystal. This plane terminates somewhere inside the crystal. The boundary of the half-plane is the dislocation, shown in Figure 2.38A. A screw dislocation, shown in Figure 2.38B, is a dislocation produced by skewing a crystal so that one atomic plane produces a spiral ramp about the dislocation. A mixed dislocation is a dislocation that contains some edge components and some screws components as illustrated in Figure 2.38C. Line dislocations cannot terminate inside an otherwise perfect crystal but do end at a crystal surface, an internal surface, or interface (e.g., a grain boundary) or they form dislocation loops. Line defects are mostly caused by the misalignment of ions or the presence of vacancies along a line. When lines of ions are missing in an otherwise perfect array of ions, an edge dislocation appears.

Dislocations are visible in transmission electron microscopy (TEM) where diffraction images of dislocations appear as dark lines. When a crystal surface is etched the rate of material removal at the location of a dislocation is faster, and this results in an etch pit. Symmetrical strain fields of edge dislocations produce a conical pit, whereas for a screw dislocation a spiral etch pit results.

A Burgers vector **b** is a measure of the magnitude and direction of a dislocation. Imagine going around a dislocation line, and exactly going back as many atoms in each direction as you have gone forward, you will not come back to the same atom where you have started. The Burgers vector points from the start atom to the end atom of your journey. This "journey" is called Burgers circuit in dislocation theory. A Burgers circuit is thus a clockwise trace around the core of a dislocation, going from lattice point to lattice point, and it must go an equal number of steps left and right and an equal number

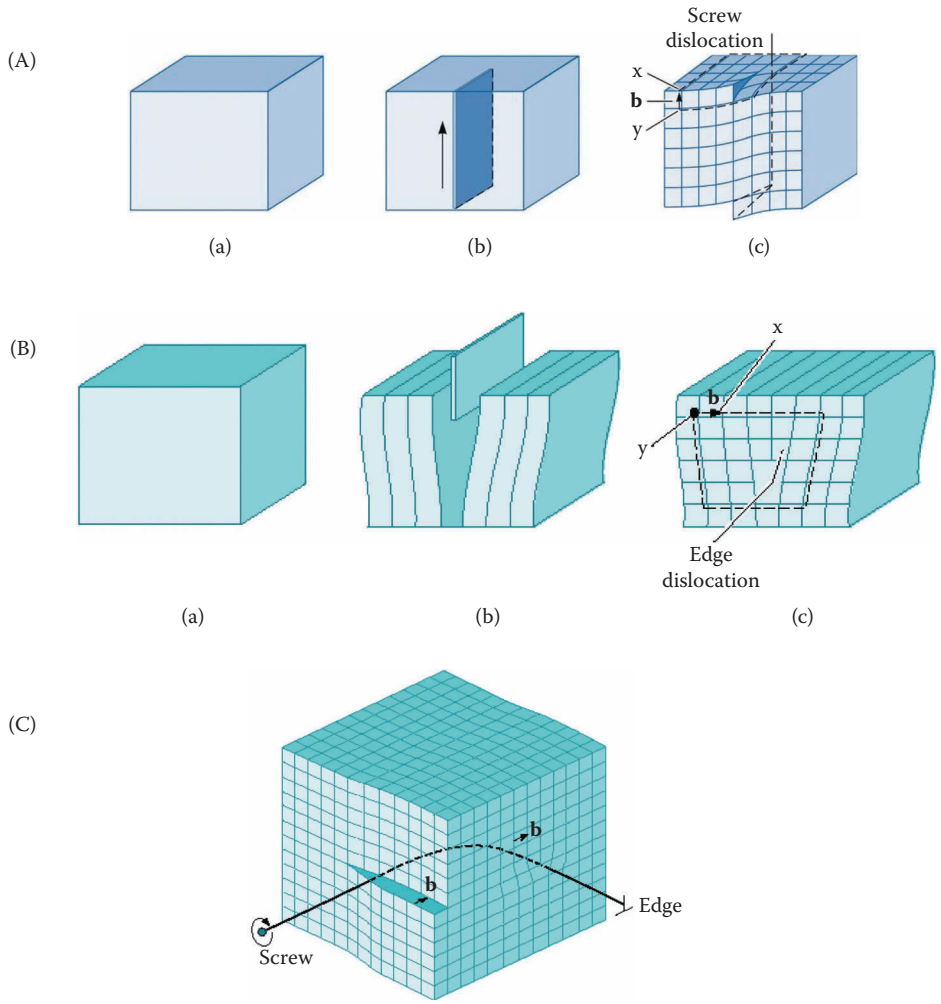


FIGURE 2.38 (A) Edge dislocation: The perfect crystal in (a) is cut, and an extra plane of atoms is inserted (b). The bottom edge of the extra plane is an edge dislocation (c). A Burgers vector **b** is required to close a loop of equal atom spacings around the edge dislocation. (B) Screw dislocation: The perfect crystal (a) is cut and sheared over one atom spacing (b and c). The line along which shearing occurs is a screw dislocation. (C) A mixed dislocation: The screw dislocation at the front face of the crystal gradually changes to an edge dislocation at the side of the crystal.

of steps up and down. Referring to Figure 2.38A, the extra half-plane of lattice points in the edge dislocation causes the Burgers circuit to be open. The vector that points from the end of the Burgers circuit to its beginning is shown here as the Burgers vector, **b**. The vector always points from one lattice point to another; it always has the same length and direction for a given dislocation, regardless where the circuit starts. For an edge dislocation, **b** is always perpendicular to the dislocation line. If **b** is parallel to the dislocation line, the dislocation is a screw dislocation. If **b** is neither perpendicular nor parallel to the line, a mixed dislocation is involved.

Plastic deformation of a material refers to irreversible deformation or change in shape that remains even when the force or stress causing it is removed, whereas elastic deformation is deformation that is fully recovered when the stress causing it is removed. Before the discovery of dislocations, materials scientists faced a big theoretical problem; they calculated that plastic deformation of a perfect crystal should require stresses 100 to 1000 times higher than those observed in tensile tests! So the problem was to explain why metals yielded so easily to plastic deformation. How they got these very large theoretical numbers can be understood from Figure 2.39. Planes of atoms in a crystal slip with respect to each other, in contrast to flow in a fluid, where the solid remains crystalline. The theoretical maximal shear stress or yield strength (failure), τ_{max}, needed

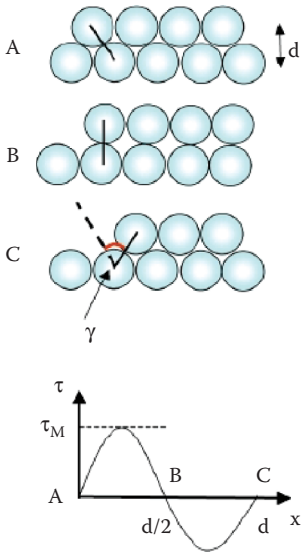


FIGURE 2.39 Calculation of the theoretical shear stress in a crystal.

to produce slip in an ideal crystal is calculated from the type and strength of the bonds involved, the spacing of the crystal, *d*, and the crystal symmetry. In yield, atoms slide tangentially from one equilibrium position to another. Thus, the shear stress, τ, is a periodic function:

$$\tau = \tau_{max} \sin \frac{2\pi x}{d} \qquad (2.60)$$

where *x* is the direction of the shear. With *x* small this may be rewritten as:

$$\tau = \tau_{max} \frac{2\pi x}{d}$$
$$x = \gamma d = \frac{\tau}{\mu} d \therefore \tau = \tau_{max} \frac{2\pi}{d} \frac{\tau}{\mu} d \qquad (2.61)$$
$$\tau_{max} = \frac{\mu}{2\pi}$$

where we have used $\gamma = \tau/\mu$ (Hooke's law) as the shear strain, with μ as the shear modulus (see Figure 2.40).

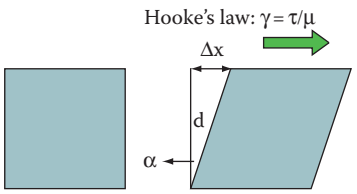


FIGURE 2.40 Shear stress, τ; shear strain, γ; and shear modulus, μ. The shear stress τ produces a displacement Δx of the upper plane as indicated; the shear strain, γ, with $\Delta x/d = \tan \alpha$ is defined $\gamma = \tau/\mu$.

TABLE 2.6 Theoretical Yield Strengths (Shear and Tensile) of Some Important Materials

Material	Shear Strength (GPa)	Tensile Strength (GPa)
Metallic Cu	1.2	3.9
NaCl	2.84	0.43
Quartz	4.4	16
Diamond	121	205

Calculated values for μ are in the range of 1–150 GPa (see also Table 2.6).

But when the experiment is carried out, permanent deformation takes place by a stress as low as 0.5 MPa! The reason is that real materials have lots of dislocations, from $10^2/cm^2$ in the best Ge and Si crystals to $10^{12}/cm^2$ in heavily deformed metals. Therefore, the strength of the material depends on the force required to make dislocations move, not the bonding energy between all atoms in two planes as calculated above. Dislocations can move if the atoms from surrounding planes break their bonds and rebond with the atoms at the terminating edge as shown in Figure 2.41. Instead of all the atoms in

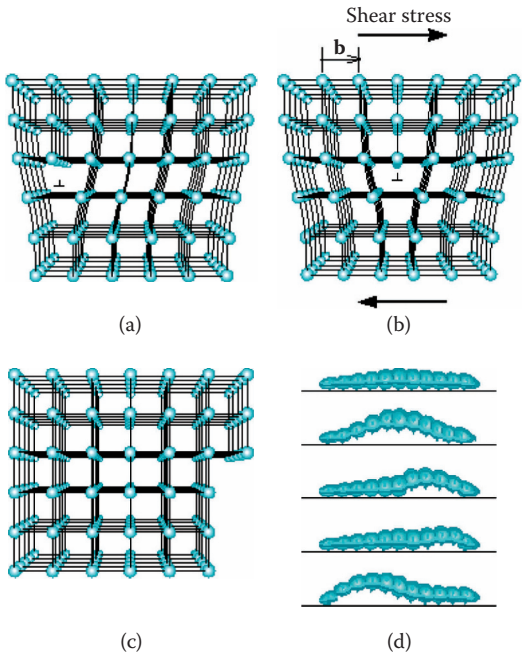


FIGURE 2.41 Dislocation movement: When a shear stress is applied to the dislocation in (a), the atoms are displaced, causing the dislocation to move one Burgers vector **b** in the slip direction (b). Continued movement of the dislocation eventually creates a step (c), and the crystal is deformed. (d) The caterpillar does not move its complete body at a single time, but it moves one segment at a time as it pulls itself forward.

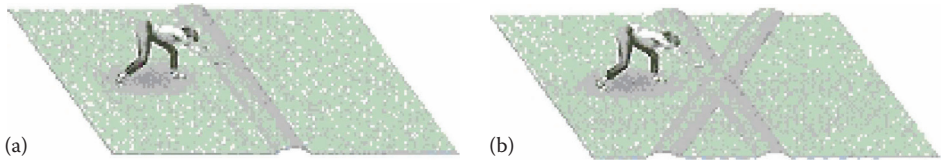


FIGURE 2.42 Moving a carpet over the floor to illustrate the effect of a line dislocation in a crystal: (a) dislocation; (b) work hardening.

a plane breaking at the same time, atoms are gliding gently in the direction of applied stress normal to dislocation lines. Thus, the direction of movement, i.e., the deformation, is the same as the Burgers vector, **b**. As a consequence **b** is also called the slip vector.

The interatomic forces in a crystal offer little resistance to the gliding motion of dislocations. An analogy is that of moving a carpet across the floor. This is difficult because of the friction developed from the contact of the whole surface of the carpet with the floor. But with a wrinkle in the carpet, as shown in Figure 2.42, the carpet can now be moved by pushing the wrinkle across the floor. The work involved is much less because only the friction between a small section of carpet and the floor has to be overcome. A similar phenomenon occurs when one plane of atoms moves past another by means of a dislocation defect.

Figure 2.43 shows how an applied shear stress, τ, exerts a force on a dislocation and is resisted by a frictional force, **F**, per unit length. The work done by the shear stress (W_τ) equals the work done by the frictional force (W_F), or:

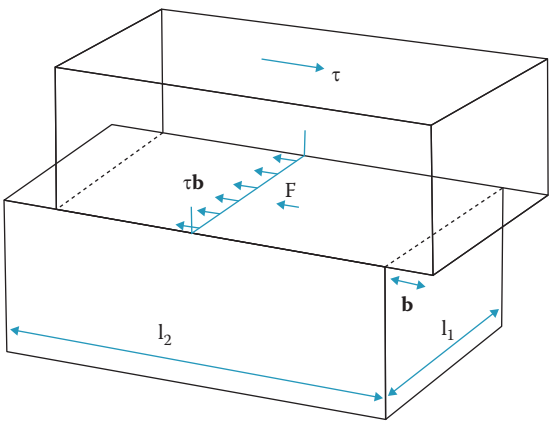


FIGURE 2.43 An applied shear stress, τ, exerts a force on a dislocation and is resisted by a frictional force, **F**, per unit length. The slip vector or Burgers vector is **b**. (Drawing by Mr. Chengwu Deng.)

$$W_\tau = (\tau\, l_1 l_2) \times \mathbf{b}$$

and (2.62)

$$W_F = (F\, l_1) \times l_2$$

With $W_\tau = W_F$, the lattice friction stress for dislocation motion is:

$$\tau = \frac{\mathbf{F}}{\mathbf{b}} \tag{2.63}$$

where **F** is the force per unit length of the dislocation and **b** is its Burgers vector or slip vector. Thus, the applied stress produces a force per unit length everywhere along the dislocation line equal to $\tau\mathbf{b}$ and perpendicular to the line element. It can be shown that:

$$\tau = \frac{\mathbf{F}}{\mathbf{b}} = \mu e^{\frac{2\pi \mathbf{a}}{\mathbf{b}}} \tag{2.64}$$

This lattice friction stress is much less than the theoretical shear strength. Dislocation motion, from Equation 2.64, is most likely on closed packed planes (large **a**, interplanar spacing) in closed packed directions (small **b**, in-plane atomic spacing). It is thus easiest to create dislocations in the closest packed crystals, and they are typically very soft and ductile.

When dislocations move it is said that slip occurred, and the lattice planes that slipped are called—imaginatively—slip planes as elucidated in Figure 2.44 for an edge dislocation and a screw dislocation. As mentioned above, the preferred slip planes are those with the greatest interplanar distance, e.g., (111) in FCC crystals, and the slip directions are those with the lowest resistance, i.e., the closest packed direction. Slip lines are the intersection of a slip plane with a free surface.

The motion of dislocations can be blocked by another dislocation, a grain boundary, or a point

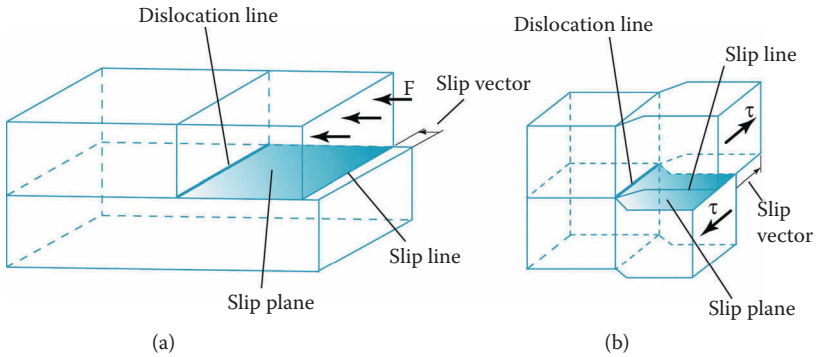


FIGURE 2.44 Schematic of slip line, slip plane, and slip (Burgers) vector for (a) an edge dislocation and (b) for a screw dislocation. (Drawing by Mr. Chengwu Deng.)

defect. Two dislocations may repel or attract each other, depending on their directions. Atoms near the core of a dislocation have a higher energy because of distortion. To minimize this energy, dislocations tend to shorten as if the line had a line tension, *T*, i.e., strain energy per unit length. The line tension is given as:

$$T \approx \frac{1}{2}\mu \mathbf{b}^2 \tag{2.65}$$

Dislocations might get pinned by interstitials and bow with a radius *R* when subjected to a shear stress like the string shown in Figure 2.45:

$$\tau \mathbf{b} L = 2T \sin\frac{\theta}{2} \tag{2.66}$$

With Equation 2.65 this leads to:

$$R = \frac{\mu \mathbf{b}}{2\tau} \tag{2.67}$$

Plastic deformation (or yielding) occurs by sliding (or slip) of parallel lattice planes past each other. Pure metals have low resistance to dislocation motion, thus they exhibit low yield strength. Adding impurities in *solution strengthening* may increase the yield strength. A well-known example is alloying Zn and Cu to form brass with a strength increase of up to 10 times over pure Cu. The bigger Zn atoms make the slip plane "rougher," thus increasing the resistance to motion. In general, impurities diffuse to dislocations and form "clouds" that pin dislocations, increasing the elastic limit. Small particles, precipitates, can also promote strengthening by impeding dislocation motion. Precipitates cause bowing of dislocations as illustrated in Figure 2.45. A critical condition is reached when the dislocation takes on a semicircular configuration between two particles separated by a distance *L*. Beyond this point a dislocation may loop and escape between the finely dispersed particles in the metal. In the semicircular situation $\tau \mathbf{b} L = 2T$ or $\tau = 2T/\mathbf{b}L$, and with Equation 2.67 $\tau = \mu \mathbf{b}/L$; thus, making *L* smaller will disadvantage dislocation looping. Finally, metals can be hardened by work hardening. Dislocations move when metals are subjected to "cold work," and their density can increase up to 10^{12} dislocations/cm^2 because of the formation of new dislocations and dislocation multiplication (see below). The consequent increasing overlap between the strain fields of adjacent dislocations gradually increases the resistance to further dislocation motion. This causes a hardening of the metal as the deformation progresses (Figure 2.42b). This effect is known as strain hardening. The effect of strain hardening can be removed by appropriate heat treatment (annealing), which promotes the recovery and subsequent recrystallization of the material. Annealing decreases the dislocation density to around 10^6 dislocations/cm^2.

Some dislocations form during the process of crystallization, but more are created during plastic deformation. Frank and Read elucidated one possible mechanism by which dislocations multiply;

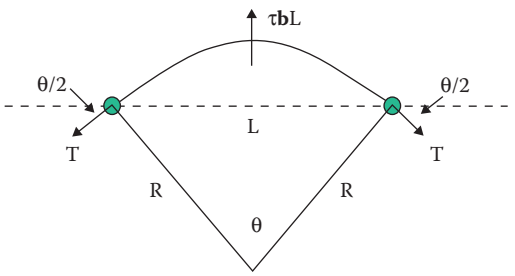


FIGURE 2.45 A pinned dislocation bows under a shear stress.

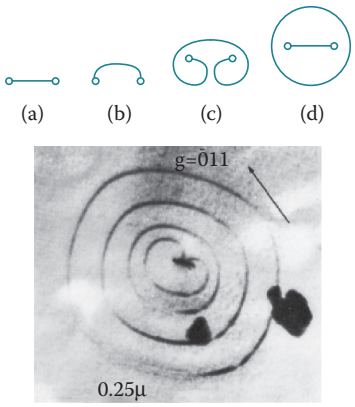


FIGURE 2.46 A Frank-Read source can generate dislocations. (a) A dislocation is pinned at its ends by lattice defects. (b) As the dislocation continues to move, the dislocation bows, eventually bending back on itself. (c) Finally the dislocation loop forms, and (d) a new dislocation is created. (e) Electron micrograph of a Frank-Read source (×330,000).

they found that a pinned dislocation (as shown in Figure 2.45), under an applied stress, produces additional dislocations. This mechanism is at least partly responsible for strain hardening. The Frank-Read dislocation multiplication mechanism is illustrated in Figure 2.46. First, a dislocation is pinned at its ends by lattice defects, and as the dislocation continues to move, it bows and eventually bends back on itself. Then the dislocation loop forms, and a new dislocation is created. Figure 2.46 also shows an electron micrograph of a Frank-Read source.

Schmid's Law, illustrated in Figure 2.47, gives the relationship between shear stress, the applied stress, and the orientation of the slip system. Slip on a given

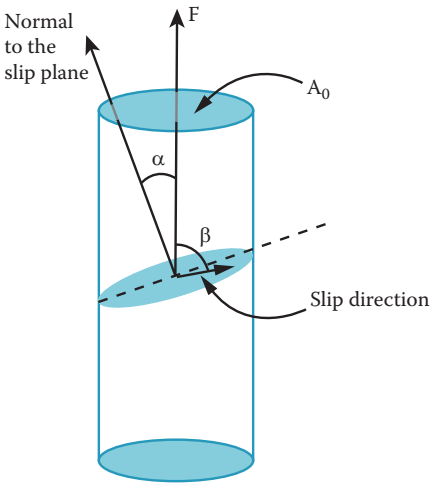


FIGURE 2.47 Geometry of slip plane, slip direction, and tensile force F. (Drawing by Mr. Chengwu Deng.)

slip system begins when the shear stress resolved on that system reaches a critical value. Consider a cylindrical crystal of cross-section A_0 under the influence of a tensile force F. Let the normal to the active slip plane make an angle α with F, and let the angle between the slip direction and F be β. The resolved shearing force, i.e., the force acting per unit area of the slip plane in the slip direction, is then given by:

$$\tau = \frac{F}{A}\cos\alpha\cos\beta \tag{2.68}$$

Because the area of the slip plane is $A/\cos\alpha$, the tensile stress per unit area normal to the slip plane is:

$$\sigma = \frac{F}{A}\cos^2\alpha \tag{2.69}$$

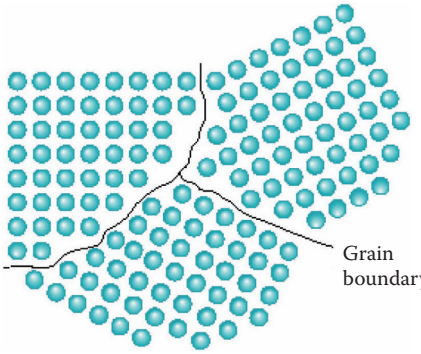
When increasing the force F, the rate of plastic flow increases very rapidly when the resolved shear stress τ reaches a critical value τ_c. In general τ_c decreases with increasing temperature and increases as a result of alloying or cold working (see above).

Besides point and line defects there are also surface defects to reckon with in single crystals. Surface defects include grain boundaries, phase boundaries, and free surfaces. The crystal structure is disturbed at grain boundaries, and different crystal orientations are present in different grains (Figure 2.48).

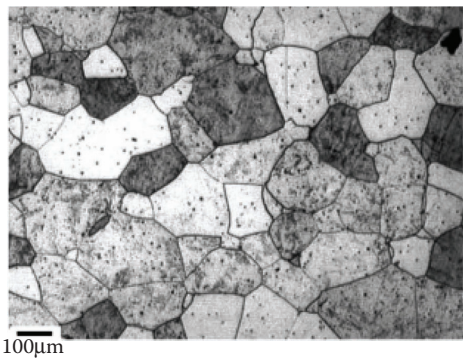
Dislocations are blocked by grain boundaries, so slip is blocked. The smaller the grain size, the larger the surface of the grain boundaries and the larger the elastic limit. The latter is expressed in the empirical Hall-Petch equation for the maximum elastic yield strength (stress at which the material permanently deforms) of a polycrystalline material:

$$\sigma_Y = \sigma_0 + \frac{K}{\sqrt{d}} \tag{2.70}$$

where d is the average diameter of the grains in micrometers, with σ_0 the frictionless stress (N/m^2) that opposes dislocation, and K a constant. The strength of a material thus depends on grain size. In a small grain, a dislocation gets to the boundary and stops, i.e., slip stops. In a large grain, the dislocation can travel farther. So small grain size equates to more strength. For example, the elastic limit of copper doubles when the grain size falls from 100 μm



(a)



(b)

FIGURE 2.48 Crystal structure is disturbed at grain boundaries. Schematic representation of grain boundaries (a) and microscope picture (b). (From Askeland D. R., and P. P. Phule, *The science and engineering of materials*, Brooks/Cole, Pacific Grove, CA, 2003.)

to 25 μm. Instead of yield strength one might also plot the hardness (H) of the material as a function of grain size, and a similar relationship is obtained (see Volume II, Figure 7.54): smaller grain size corresponds to a harder material. Indentation (hardness) testing is very common for bulk materials where the direct relationship between bulk hardness and yield strength is well known. As we will learn in Volume II, Chapter 7, in the section on thin film properties, the Hall-Petch relationship has been well established for grain sizes in the millimeter through submicrometer regimes but is less well-known in the nano regime. Based on Equation 2.70 one expects that nano-sized grains would produce materials with yet greater mechanical integrity, but in reality this is not the case! There is a reverse Hall-Petch effect, i.e., the strength of materials, from a small grain size on, starts to decrease with decreasing grain size. Plastic deformation occurs at lower and lower stresses as the grains shrink. In other words, an optimal grain size (d_c) exists as suggested by the plot in Volume II, Figure 7.54. The classic Hall-Petch relationship is based on the idea that grain boundaries act as obstacles to dislocation movement, and because dislocations are carriers of plastic deformation, this manifests itself macroscopically as an increase in material strength. The Hall-Petch behavior breaks down when the smallest dislocation loop no longer fits inside a grain. Lattice dislocation cannot be the way very small-grained materials deform. The deformation mechanism for materials with very small grains (<20 nm) is indeed different, and it has been suggested that plastic deformation in this case is no longer dominated by dislocation motion; it is believed that individual atoms migrate, diffuse, and slide along grain boundaries and through triple junctions (Y-shaped grain boundary intersections).[1]

In Volume III, Chapter 7, on scaling, we learn that the surface-to-volume ratio (S/V) of particles scales as $1/r$, where r is the characteristic dimension of the particles. The smaller a particle, the more of its atoms find themselves at its surface. A bulk solid material will typically have less than 1% of its atoms on its surface but 10-nm particles have about 15% of surface atoms (Figure 2.49). The high S/V ratio of nanoparticles makes them

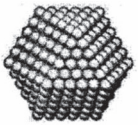
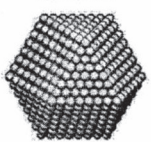
Full-Shell Clusters		Total Number of Atoms	Surface Atoms (%)
1 Shell		13	92
2 Shells		55	76
3 Shells		147	63
4 Shells		309	52
5 Shells		561	45
7 Shells		1415	35

FIGURE 2.49 Nanoparticles, clusters of atoms in shells.

more reactive as catalysts in chemical reactions and lowers their melting temperature, T_m. There is an inverse linear relationship between melting temperature and the surface-to-volume ratio. This makes sense because atoms on a surface are more easily accessed and rearranged than atoms in the bulk. As a consequence, the melting temperature of particles is always lower than the bulk. Grains in a solid are analogous to particles, and the melting temperature of solids decreases with grain size. Although the temperature dependence of σ_0 and K in Equation 2.70 can be neglected for conventional grain sizes, for grains smaller than 20 nm this assumption breaks down. In nanomaterials the melting temperature decreases because of the smaller grain size.

If we could grow a crystal without dislocations, it should approach the theoretical shear stress τ calculated in Equation 2.61. Sometimes fine metallic whiskers can be grown virtually free of dislocations, and these are very strong indeed.

In Figure 2.50 we show a picture of a zinc whisker. Metal whiskers are a crystalline metallurgical phenomenon whereby metal grows tiny, filiform hairs. The effect is primarily seen with elemental metals but is also observed with alloys. The mechanism behind metal whisker growth is not well understood but seems to be encouraged by compressive mechanical stresses, including residual stresses caused by electroplating, mechanically induced stresses, stresses induced by diffusion of different metals, and thermally induced stresses. Metal whiskers differ from metallic dendrites in several respects; dendrites are fern-shaped and grow across the surface of the metal, whereas metal whiskers are hair-like and project at a right angle to the surface. Although the precise mechanism for whisker formation remains unknown, it is known that whisker formation does not require either dissolution of the metal or the presence of an electromagnetic field. Whiskers, like carbon nanotubes, approach ideal lattices.

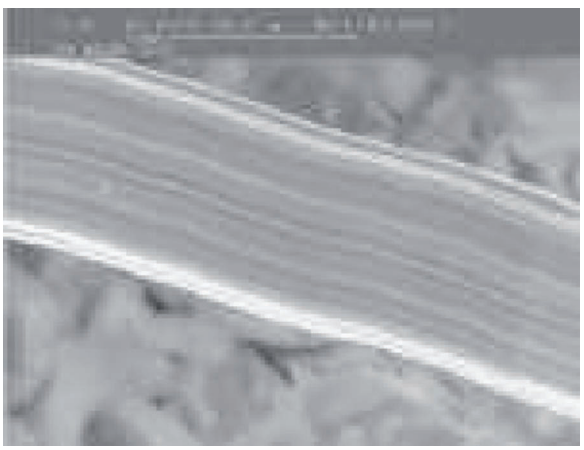


FIGURE 2.50 SEM of a zinc whisker; diameter is 10 μm.

As we are moving to smaller and smaller functional devices in MEMS and NEMS, surfaces and crystal imperfections contribute more and more importantly to overall performance and behavior. Many of the observations described here will have to be revisited when exploring nanomaterials and nanodevices. These effects will be treated in Volume II, Chapter 7 on thin films and in Volume III, Chapter 3 on nanotechnology.

Acknowledgments

Special thanks to Xavier Casadevall I Solvas, Robin Gorkin, Chengwu Deng, Kartikeya Malladi, Leyla Esfandiari, Fatima Alim, and Drs. Sean Parkin, Han Xu, and Guangyao Jia.

Appendix 2A: Plane Wave Equations

We define a plane wave,* $e^{i\mathbf{k}.\mathbf{r}}$, as a wave that is constant in a plane perpendicular to $\mathbf{k}$ (in rad/m) and is periodically parallel to it, with a wavelength $\lambda = 2\pi/\mathbf{k}$, where $\mathbf{k}$ is the wave number (because it measures the number of wavelengths in a complete cycle) with a value of $2\pi/\lambda$ and where $\omega = 2\pi\nu = c\mathbf{k}$ (free space) is the period of the wave (see also Table 2A.1). A bit simpler: a plane wave is a continuous wave (CW) whose amplitude and phase are constant in the directions transverse to the propagation direction.

* The complex exponential representation for periodic functions is convenient for adding waves, taking derivatives of wave functions, and so on. It is equivalent to a linear combination of a sine and cosine function because $e^{i\theta} = \cos\theta + i\sin\theta$.

Table 2A.1 Plane Wave Parameters

Parameter	Symbol/Value
Amplitude	$\mathbf{E}_m, \mathbf{B}_m$
Phase	$\phi = kx - \omega t$
Velocity	$v = \frac{\omega}{\mathbf{k}}$
Wavelength	λ
Period	T
Wave vector	$\mathbf{k} = \frac{2\pi}{\lambda}$
Angular frequency	$\omega = \frac{2\pi}{T}$
Wave number	$\bar{\mathbf{k}} = \frac{1}{\lambda}$
Cyclic frequency	$v = \frac{1}{T}$

The amplitude of a wave propagating in the x-direction is mathematically introduced as

$$\psi = \psi_0 \sin(kx + \omega t + \delta)$$
$$\omega = 2\pi f, \text{ the cyclic frequency}$$
$$k = \frac{2\pi}{\lambda}, \text{ the wave vector}$$
$$\delta, \text{ the phase at } t = 0 \text{ and } x = 0$$

Calculations are greatly simplified by using complex numbers:

$$\psi = \psi_0 \text{Im } e^{i(kx+\omega t+\delta)}$$
$$i = \sqrt{-1}$$
$$|e^{ix}| = 1$$
$$\text{Im } (e^{ix}) = \sin x$$
$$\text{Re } (e^{ix}) = \cos x$$
$$\sin x = \frac{e^{ix} + e^{-ix}}{2i}$$
$$\cos x = \frac{e^{ix} + e^{-ix}}{2}$$

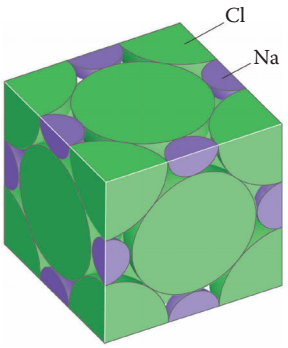
Questions

2.1: Scientists are considering using nanoparticles of magnetic materials such as iron-platinum (Fe-Pt) as a medium for ultrahigh density data storage. Arrays of such particles potentially can lead to storage of trillions of bits of data per square inch—a capacity that is 10 to 100 times higher than any current storage devices such as computer hard disks. If these scientists consider iron (Fe) particles that are 3 nm in diameter, what is the number of atoms in one such particle?

2.2: Assuming that silica (SiO_2) has 100% covalent bonding, describe how oxygen and silicon atoms in silica (SiO_2) are joined.

2.3: What is the difference between lattice and basis and between unit cell and primitive cell?

2.4: What are the net numbers of Na^+ and Cl^- ions in the NaCl unit cell represented below? The crystal is an example of which type of cubic lattice? Identify the atom positions of the Na and Cl atoms in the NaCl structure.



2.5: Consider the plane defined by the three points, P1(2,4,–3), P2(–1,2,1), and P3(3,0,–2). Calculate the points where this plane intersects with the axes and derive the Miller indices associated with this plane.

2.6: Calculate (a) the angle between [111] and the direction normal to (111) plane in a simple cubic crystal and (b) the angle between the [121] direction and the direction normal to (113) plane in a simple cubic crystal.

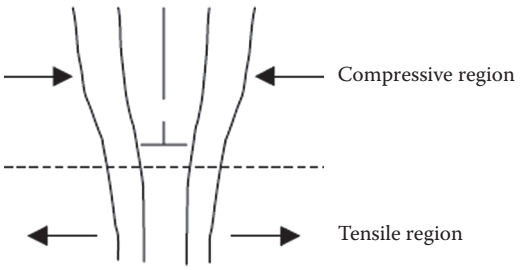
2.7: A signal $x(t)$ has a Fourier transform $X(f)$. Express the Fourier transforms of x_1 in terms of $X(f)$, the function $x_1(t) = x(3 - t)$.

2.8: In an x-ray set-up with a crystal-to-detector distance of 100 mm (D), you find that the highest resolution reflection is at $x = 20$ mm, $y = 15$ mm, relative to the direct beam position. The wavelength $\lambda = 1.54$ Å. What is the Bragg angle of this reflection? What is the d-spacing of the crystal? If the detector is

circular and has a radius of 100 mm, and you would like to collect data so that the highest resolution reflection is at the detector edge, would you move the detector closer to or further away from the crystal?

2.9: Describe the difference between a reciprocal lattice and a real one.

2.10: Edge dislocations may be used to getter impurities in semiconductors. In an edge dislocation, there are always two regions, a compressive region, where the layer is inserted, and a tensile region (see figure below). Which of these two regions will best accommodate (via elastic interactions) *substitutional* atoms whose radius is larger than that of the host atoms?

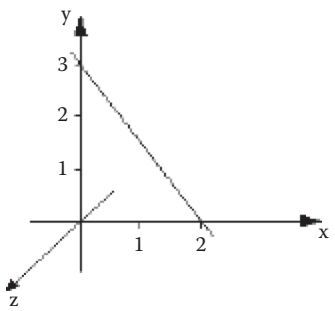


2.11: Show that the reciprocal lattice of a face-centered cubic (FCC) lattice is a body-centered cubic (BCC) lattice or, conversely, that the reciprocal lattice of a BCC lattice is an FCC lattice.

2.12: An electron moves with speed u = 0.7c. Calculate its total energy and its kinetic energy in eV.

2.13: Calculate the number of atoms in a 100 μm long Ag line (1 μm wide and 1 μm high). If using STM we put one atom down per second, how long will it take to finish this Ag line?

2.14: What is the Miller index for the plane shown below?



2.15: What is the number of nearest neighbors for the following crystal lattices:
(a) simple cubic (SC)
(b) face-centered cubic (FCC)
(c) body-centered cubic (BCC)

2.16: Calculate the angle θ of reflection for an x-ray experiment with λ = 1.54 Å, for a cubic crystal with a lattice parameter of **a** = 5 Å.

2.17: X-rays with wavelength 1.54 Å are "reflected" from the (110) planes of a cubic crystal with unit cell **a** = 6 Å. Calculate the Bragg angle, θ, for all orders of reflection, *n*.

2.18: What is the closest packed crystal? Simple cubic (SC), body-centered cubic (BCC), or face-centered cubic (FCC)?

2.19: Is five-fold symmetry ever found in crystal lattices? Why or why not?

2.20: Calculate the number of atoms in 100 g of silver.

2.21: The derivation of Bragg's law results in $n\lambda = 2d\sin\theta$. What does *n* represent and why is it usually omitted? Can you give an example to show why n is not needed?

2.22: Calculate the packing factor in a FCC lattice.

2.23: What is the rule for determining the *slip direction* in a close-packed material?

2.24: What causes work hardening?

2.25: How can we determine the direction cosine between two vectors?

2.26: How can we identify the direction of slip in a crystal (Burgers vector)?

2.27: Why do very thin metal wires/whiskers exhibit very high strengths?

2.28: Why does the strength of glass fibers increase as the diameter goes down?

2.29: Find the angle between the planes 110 and 100 in a simple cubic crystal.

2.30: Could you use x-ray diffraction to determine the coefficient of thermal expansion (i.e., change in length of a material due to change in temperature)?

2.31: Diamond and graphite are examples of which type of crystalline solids: molecular, covalent network, ionic, or metallic?

2.32: What value does the atomic scattering factor f approach as 2θ approaches 0?

2.33: How does light affect the color of a crystal?

Further Reading

Vectors and Tensors

Danielson D. A. 1997. *Vectors and tensors in engineering and physics.* New York: Addison-Wesley Publishing Company, Inc.

Crystallography

Books reviewed or listed in *Crystallography News*. See http://img.cryst.bbk.ac.uk/bca/cnews/books/index.html.

The International Union of Crystallography (IUcr) produces a number of books on crystallography, in association with Oxford University Press and Springer. See http://www.iucr.org/iucr-top/publ/index.html.

Hannay, N. B. 1967. *Solid-state chemistry.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Ladd, M. F. C., and R. A. Palmer. 2003. *Structure determination by X-ray crystallography.* 4th ed. New York: Springer.

Sands, D. S. 1969: *Introduction to crystallography.* New York: W.A. Benjamin, Inc.

Solid-State Physics

Brown, F. C. 1967. *The physics of solids.* New York: W.A. Benjamin, Inc.

Dekker, A. J. 1970. *Solid state physics.* New York: Prentice-Hall, Inc.

Kittel, C. 2005. *Introduction to solid state physics*. 8th ed. New York: John Wiley & Sons, Inc.

"Phase Problem"

Eisenberg, D., and D. Crothers. 1979. *Physical chemistry: with applications to the life sciences*. San Francisco: Benjamin-Cummings/Addison Wesley Longman.

Glusker, J. P., and K. N. Trueblood. 1972. *Crystal structure analysis: a primer.* New York: Oxford University Press.

Holmes K. C., and D. M. Blow. 1965. *Use of X-ray diffraction in the study of protein and nucleic acid structure*. New York: Krieger Pub Co/Interscience.

Wilson, H. R. 1966. *Diffraction of X-rays by proteins, nucleic acids, and viruses.* New York: St. Martin's Press.

Dislocations

Hull, D. 1965. *Introduction to dislocations*. Oxford, UK: Pergamon Press.

Weertman, J., and J. R. Weertman. 1964. *Elementary dislocation theory.* New York: MacMillan.

Reference

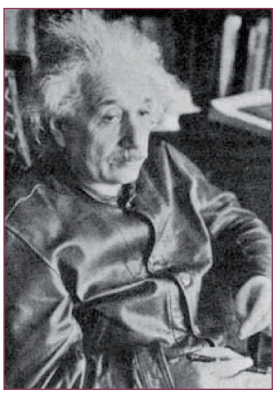
1. Schiotz, J., and K. W. Jacobson. 2003. A maximum in the strength of nanocrystalline copper. *Science* 301:1357–59.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

3

Quantum Mechanics and the Band Theory of Solids

Some of the actors in this chapter: (a) Fermi, (b) Schrödinger, (c) de Broglie, (d) Feynman, (e) Born, (f) Bohr, (g) Rutherford, (h) Dirac; and below is Einstein.



Outline

Introduction
Classical Theory Starts Faltering
Quantum Mechanics to the Rescue
Beyond Schrödinger's Equation

The only reason for time is so that everything doesn't happen at once.

Albert Einstein

Nothing is real.

John Lennon, 1940–1980

Introduction

As we came to understand from Chapter 2, when atoms are brought together in crystals, their outermost electrons are influenced by a periodic potential. Therefore, the possible electron energies form bands of allowed values separated by bands of forbidden values. This band structure is of fundamental importance in explaining the properties of metals, semiconductors, and insulators. The most accepted and most accurate theory in modern physics today, to predict physical, mechanical, chemical, and electrical properties of atoms, molecules, and solids—including the band theory of solids—is quantum mechanics.

At the end of the nineteenth century, physicists thought they had a good grasp of the physical world with mechanics (e.g., Newton, Hamilton, and Lagrange), statistical mechanics (mostly Boltzmann and Gibbs), hydrodynamics (Stokes), and electrodynamics (Maxwell). There had been a growing unease, though, about the incapability of classical theories to explain a wide variety of experiments, such as the magnitude of the electrical and thermal conductivity of metals and the heat capacity of metals and insulators. Confidence was further challenged by a series of new discoveries: radioactivity (1896), the electron (1897), the quantum (1900), the photoelectric effect (1887), and x-rays (1895). The result was the development of a set of new theories, all explaining extreme aspects of nature better than classical theories. Although the classical theories worked well at everyday velocities and scales, at the extremes, new theories were needed. For very fast phenomena, special relativity worked better; for very small particles, quantum mechanics; and for very large phenomena, general relativity was demonstrated to be superior.

Before introducing quantum mechanics, we start out by detailing the areas where classical physics started faltering at the end of the nineteenth century. We specifically describe the shortfalls of the classical theories in explaining electrical and thermal conductivity of metals, heat capacity, the observation of a positive Hall effect, the ultraviolet catastrophe, and the photoelectric effect. All the above phenomena follow logically from quantum theory.

The relevance of quantum mechanics in the context of ICs and NEMS cannot be underestimated, and the profound implications of quantum physics for nanoelectronics and NEMS are a recurring topic throughout this book. In the IC and NEMS world, we are moving fast into the realm of quantum mechanics. Moore's law might remain valid until about 2020, but by then the scale of electronic components will be at the molecular/atomic level, and hence can no longer be described by classical mechanics. Quantum computing and nanotechnology, including nanotubes, nanowires, biological nanostructures, and quantum dots, all require some grounding in quantum mechanics to be understood at all. Quantum mechanics must now become a familiar tool not only to physicists but also to materials scientists, biologists, and electrical, mechanical, and bioengineers.

Classical Theory Starts Faltering

Introduction

In this section, we cover examples where microscopic classical models started failing at explaining a wide variety of experimental results obtained in the late nineteenth century. The examples include early models devised to explain electrical conductivity in metals, heat capacity of metals and insulators, the temperature dependence of electrical conductivity and heat capacity, thermal conductivity, the Hall effect, blackbody radiation, and the photoelectric effect. The failing of each model ultimately originated in the mistaken assumption that lattice vibrations (phonons), electrons, and photons could all take on continuous energy values, an error that quantum mechanics corrects.

Electronic Conductivity

DC Electrical Conductivity

Paul Drude (1863–1906) (Figure 3.1) was intrigued with the huge resistivity range in materials—from $10^{18}\ \Omega\cdot$cm for fused quartz to $10^{-6}\ \Omega\cdot$cm for silver (Figure 3.2)—and developed one of the first models to explain electrical conductivity in metals.

Modern condensed matter physics really started with the discovery of the electron by J.J. Thompson

FIGURE 3.1 Paul Drude (1863–1906).

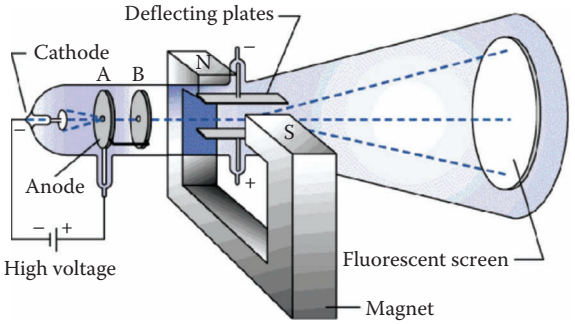


FIGURE 3.3 J.J. Thomson demonstrated that an electron behaves as a particle using an apparatus as shown here (A and B are anodes). The electron stream is deflected by electric and magnetic fields.

in 1897. In a series of experiments performed before the year 1900, Thompson demonstrated that electrons behave as particles of mass m that carry a fixed amount of negative electrical charge, e. These particles move in trajectories governed by the laws of electricity, magnetism, and classical mechanics, and Thompson determined the ratio e/m directly using an instrument as shown in Figure 3.3. Thompson received the Nobel Prize for his work on the electron in 1906. It was in the famous oil-drop experiment of Robert Millikan (1868–1953), carried out in 1909, that the size of the charge on an electron was finally established. Millikan also determined that there was a smallest "unit" charge, or that charge is "quantized."

Drude relied on the newly introduced concept of electrons to postulate a theory of metallic conductivity. His work was before the development of quantum mechanics, so he relied on classical physics models only. He introduced the idea of a free electron gas (FEG), in which he assumed that an electron gas surrounds the positive ion cores of the metal and that electron-ion interactions are negligible. Applying the kinetic theory of gases to the free electron gas, he had electrons only moving in straight lines and colliding only with ion cores, thus neglecting electron-electron interactions. Drude also envisioned collisions in which electrons instantaneously lose all the energy they previously gained from an electric field. He approximated the mean free path of the electrons, λ, with the interionic core spacing in the solid lattice, a. It should be noted that in reality electron densities (10^{22}–10^{23} cm^{-3}) are thousands of times greater

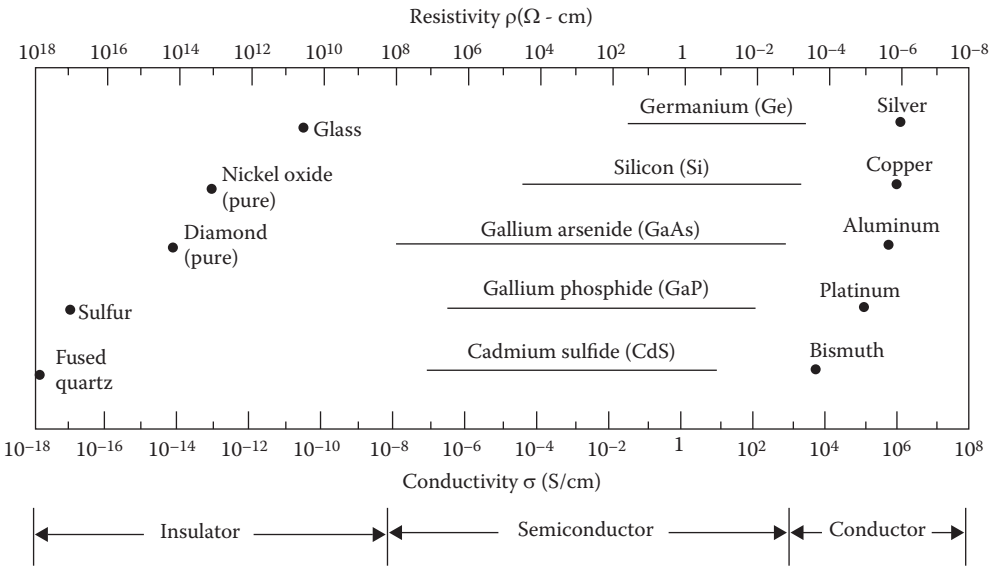


FIGURE 3.2 Conductivity/resistivity range for common materials.

than those of a gas at normal conditions, and the assumption that there are no strong electron-electron and electron-ion interactions is wrong. In spite of this, the model successfully explains the form of Ohm's law. However, it does fail to explain many other important aspects of conductors, such as the exact magnitude of electrical conductivity and heat capacity and their temperature dependence, or the relationship between electrical and thermal conduction.

In what follows we apply macroscopic boundary conditions to a piece of metal. In the nanoworld, as we will learn, things are different: with a macroscale conductor wire, the mean free path of the electrons, λ, is very small compared with the wire length, and the motion of electrons is diffusive. But, with a nanowire, the wire is short compared with λ and the electron motion may turn ballistic.

We start with the experimental observation that the current in a conductor is proportional to the applied voltage ($V \propto \mathbf{I}$), i.e., Ohm's law. If an electrical field **E** (V/m) is applied over a conductor of length *L*, a charge, *Q*, flows and an electrical current, $\mathbf{I} = dQ/dt$, results [SI unit: 1 ampere (A) = 1 coulomb per second (C/s)]. The current density is given as $\mathbf{J} = \mathbf{I}/A$ or $d\mathbf{I}/dA$, with *A* the cross-sectional area of the current pathway or also:

$$\mathbf{I} = \int \mathbf{J}\, dA \tag{3.1}$$

where **J** is given by (see Chapter 2):

$$\mathbf{J} = \sigma\mathbf{E} \tag{2.2}$$

or also (see Chapter 2):

$$\mathbf{E} = \rho\mathbf{J} \tag{2.4}$$

σ is the electrical conductivity, and ρ is the electrical resistivity ($\sigma = 1/\rho$). The SI unit for resistivity is in ohm·meter (Ω·m) and ohm = volt/ampere (V/A); the SI unit for conductivity is in siemens per meter (S/m^{-1}) and siemens = ampere/volt = ohm^{-1}. The conductivity, σ, and resistivity, ρ, are intrinsic characteristics of a material and are independent of sample geometry but are linked to the crystal structure as we may glean from Equations 2.3–2.7 in Chapter 2. From Equation 2.2, σ measures the current density for a given electric field. For now, we assume

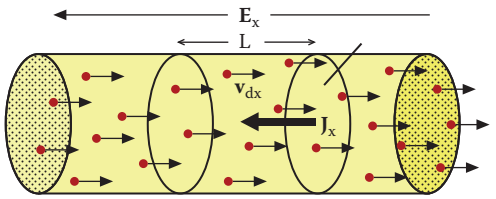


FIGURE 3.4 Electron conduction in a section of metal wire in the presence of an electrical field. The average distance traveled by an electron is the mean free path λ, which Drude assumed to be equal to the lattice constant, *a*.

an isotropic solid, and because $\mathbf{J} = \mathbf{I}/A$ and $\mathbf{E} = V/L$, Equation 2.2 can be rewritten as $\mathbf{J} = \mathbf{I}/A = \sigma\, V/L$ or:

$$V = \mathbf{I}(L/\sigma A) = \mathbf{I}(\rho L/A) = \mathbf{I}R \tag{3.2}$$

i.e., the familiar Ohm's law, with the resistance, *R*, linked to the geometry of the sample (*L* and *A*). The SI unit for resistance is ohm (Ω).

Drude's microscopic interpretation of Ohm's law can be understood from an inspection of Figure 3.4, where we consider the conduction of electrons in a metal wire on application of an electrical field in the *x*-direction, $\mathbf{E}_x$. There are *n* electrons per unit volume (electron density), and they all move in the direction of the current, *I*, with a drift velocity in the *x*-direction, $\mathbf{v}_{dx}$. The drift velocity, $\mathbf{v}_{dx}$, is the net motion of electrons opposite to the electrical field (Figure 3.5). The number of electrons crossing area *A* in a time *dt* is $nA\mathbf{v}_{dx}dt$. In the time segment *dt*, the electrons have traveled a distance *L* along the wire. The current density in the *x*-direction, $\mathbf{J}_x$, is then $\frac{\mathbf{I}}{A}$ or $\frac{\Delta Q}{A\Delta t} = \frac{(ne)(AL)}{AL/\mathbf{v}_{dx}}$, and with *Q* the charge we calculate:

$$\mathbf{J}_x = en\mathbf{v}_{dx} \tag{3.3}$$

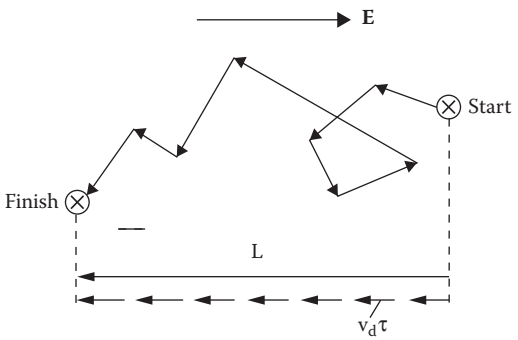


FIGURE 3.5 Drift of an electron opposite the electrical field. During an average time τ, the electron travels a mean free path $\lambda = v_{dx}\tau$. In a time *dt*, the electrons have traveled a distance *L* along the wire.

where e is the charge of the electron. By convention, a positive charge moves with the electrical field, and Equation 3.3, more correctly, should read $\mathbf{J}x = env_{dx}$, with $\mathbf{J}_x$ and $\mathbf{v}_{dx}$ vectors with a direction and magnitude. For simplicity, we often leave the vector notation out, and introduce it only when we want to draw special attention to it. Drude treated the free electron gas as a gas of molecules where distribution of velocities follows the Maxwell-Boltzmann distribution (see below). In such a gas, at T = 0, all the free electrons in a conductor have zero kinetic energy. When heating the conductor, the lattice ions acquire an average kinetic energy of $\frac{3}{2} k_B T$, where k_B is the Boltzmann constant. This average energy is imparted to the electron gas by collisions between electrons and lattice ions. The latter is a consequence of the equipartition theorem. At ordinary temperatures (~300 K) the mean kinetic energy of the electrons based on this model is about 0.04 eV. From this mean kinetic energy, we approximate the thermal velocity, v_{th}, of the electrons in a metal from the root mean square (rms speed) or $v_{rms} = \sqrt{\frac{3k_B T}{m_e}}$ (see Equation 3.20 below), which is slightly larger than the average v_{avg} or mean speed $\bar{v} = \sqrt{\frac{8k_B T}{\pi m_e}}$ (see Equation 3.21 below), as $v_{th} = \bar{v} \approx v_{rms} = 1.57 \times 10^5$ m/s or almost 1000 km/s (this is about 1% of the speed of light!). However, when calculating $\mathbf{v}_{dx}$ in Equation 3.3 for a current of 1 A, one obtains velocities of the order of 10^{-4} m/s or only 0.1 mm/s! The reason for the huge difference between drift and thermal velocity is that electrons travel at fast thermal velocities for a short, average time, τ, and then "scatter" because of collisions with atoms, grain boundaries, impurities, or material surfaces (especially in very thin films or small particles). The drift velocity is also not how fast "electricity travels," for electric fields travel essentially at the speed of light. To reconcile this discrepancy, think of electrons in a wire as a pipe full of water; when a little water enters one end of the pipe, almost immediately some water flows out at the other end.

The very small drift velocity caused by the electric field has essentially no effect on the very large mean speed of the electrons: in other words, v_{th} does not depend on **E**. Drude assumed that all of the electrons' forward velocity is reduced to zero after each collision and must then be accelerated again by the electrical field.

The forces exerted on a colliding electron in an electric field, **E**, are given by Newton's second law, according to which forces give rise to a change in the momentum of particles [$(\mathbf{F} = d(m_e \mathbf{v}_{dx})/dt = d\mathbf{p}_x/dt$] or:

$$\frac{d\mathbf{p}_x}{dt} = -e\mathbf{E}_x + \mathbf{F}_{coll} = -e\mathbf{E}_x - \frac{\mathbf{p}_x}{\tau}$$
$$= -e\mathbf{E}_x - \frac{m_e \mathbf{v}_{dx}}{\tau} \qquad (3.4)$$

where $-e\mathbf{E}_x$ is the force on the electron garnered from the electrical field, m_e the mass of an electron in vacuum, $\mathbf{F}_{coll}$ the collision force, and $\boldsymbol{p}_x$ the electron's momentum in the x-direction ($=m_e \mathbf{v}_{dx}$). Because

$$-e\mathbf{E}_x - m_e \mathbf{v}_{dx}/\tau = 0 \qquad (3.5)$$

the result is a constant average velocity:

$$\mathbf{v}_{dx} = -\frac{e\mathbf{E}_x \tau}{m_e} \qquad (3.6)$$

where we introduced a minus sign because, as remarked above, the drift velocity is opposite the electrical field **E**. Substituting this result in Equation 3.3 and generalizing for three directions, we obtain the famous Drude result (ignoring vector notation):

$$\sigma = \frac{ne^2\tau}{m_e} \quad \text{and} \quad \rho = \frac{m_e}{ne^2\tau} \qquad (3.7)$$

Scatter time τ is also known as the relaxation time, the collision time, or the mean free time a randomly picked electron travels before the next collision. Scatter time τ decreases with increasing temperature T, i.e., more scattering at higher temperatures leads to higher resistivity in a metal. During an average time τ, electrons travel a mean free path λ, i.e., $\lambda = v_{th}\tau$, and in terms of the mean free path and the mean speed the resistivity is given as:

$$\rho = \frac{m_e v_{th}}{ne^2\lambda} \qquad (3.8)$$

According to Ohm's law, the resistivity is independent of the field, and in Equation 3.8 only λ and v_{th} could possibly be dependent on the field. But we

just saw that the very small drift velocity caused by the electric field has essentially no effect on the very large mean speed of the electrons or v_{th} does not depend on **E**. Drude assumed the electron mean free path, λ, to be equal to the lattice constant, a, which is of the order of 1 nm, and of course also independent of the electric field **E**; neither λ nor v_{th} depends on **E** in accordance with Ohm's law. In the absence of an electric field, the electrons perform a random, thermal motion, and there is no conduction, but when an electric field is applied, electrons move into a direction opposite to the field, thus generating a current. The value of the resistance is finite because electrons collide with the lattice ions, and they are stopped frequently in their tracks before being accelerated again. This simple model does explain correctly the form of Ohm's law.

From Equation 3.7 there are two contributions to the conductivity. One is the number of charges (ne), and the second is how easily those charges can be accelerated ($e\tau/m_e$). To really understand the differences between the numbers of "free charges" in different materials requires the quantum description introduced below. However, let us accept for now that in a material such as silver, the number of electrons that are free to move equals $5.8 \times 10^{28}/m^3$ and that in zinc it is even larger, $1.3 \times 10^{29}/m^3$. From these numbers one would expect zinc to be the better conductor! However, as we shall see, this is not the end of the story. The second contribution to the conductivity in Equation 3.7 is the charge carrier's drift mobility, μ_e (in the case of electrons), which, in cm^2/Vs, is given as:

$$\mu_e = \frac{e\tau}{m_e} \tag{3.9}$$

Therefore, the conductivity is the product of the number of free charges and the mobility of those charge carriers:

$$\sigma = \mu_e ne \tag{3.10}$$

The drift mobility, μ_e, may also be defined as the drift velocity per unit applied electric field, or:

$$\mu_e = \frac{v_d}{E} \tag{3.11}$$

The drift mobility, from Equation 3.9, is linked in turn to the mean scattering time between collisions, τ. Coming back to the comparison of the conductivity of silver and zinc, typical mobilities in silver are much higher than those in zinc so that the electrical conductivity is higher, even though zinc has more electrons. This makes silver the better conductor after all. Once we have introduced quantum mechanics, it will become clear that the mean scattering time, τ, has nothing to do with the stagnant lattice ions in the crystal but is determined by lattice vibrations, imperfections, and impurities instead. Consequentially, the mobility of a metal can be reduced by introducing defects (e.g., kinking a wire) or by increasing the number of lattice vibrations, i.e., phonons (by raising the temperature).

The Maxwell-Boltzmann Distribution

The next question is, how does resistivity of a metal depend on temperature? To answer this we calculate, à la Drude, how the drift velocity, v_d, depends on temperature. With electrons behaving like an ideal gas, the distribution of electron speeds is described by a Maxwell-Boltzmann (MB) distribution. In a MB distribution, the probability of finding particles in a particular energy state varies exponentially as the negative of the energy divided by k_BT or:

$$f_{MB}(E) = Ae^{-\frac{E}{k_BT}} \tag{3.12}$$

where A is a normalization constant, and $e^{-\frac{E}{k_BT}}$ is called the Boltzmann factor. Based on Equation 3.12, at a given temperature, particles are distributed among all available levels, and the ground state always contains the bulk of the particles, whereas other levels will contain exponentially less. The Maxwell-Boltzmann distribution is illustrated in Figure 3.6.

The number of particles per unit volume, i.e., density $n(E)$, that have energies between E and $E + dE$ is $n(E)dE$, where $n(E) = G(E)f_{MB}(E)$. The function $G(E)dE$ is the total number of possible (allowed-to-occupy) energy states per unit volume with energies between E and $E + dE$. It is also referred to as the density of state function, often abbreviated as DOS. The other function, $f_{MB}(E)$, is the Maxwell-Boltzmann distribution, representing the probability function

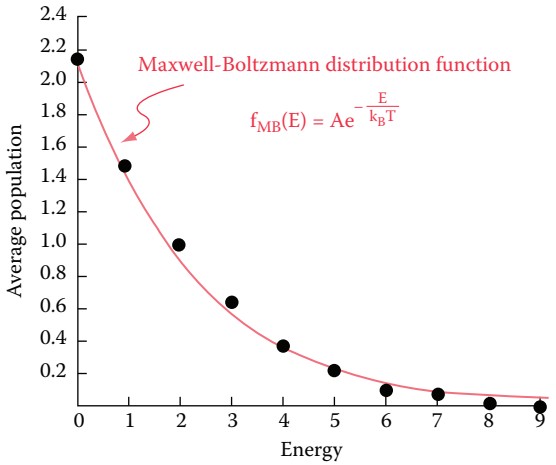


FIGURE 3.6 The Maxwell-Boltzmann distribution function: $f_{MB}(E) = Ae^{-\frac{E}{k_BT}}$.

of occupancy of a state with energy E. Because $E = mv^2/2$, we may, using Equation 3.12, rewrite $n(E)dE$ as:

$$n(E)dE = G(E)Ae^{-\frac{mv^2}{2k_BT}}dE \qquad (3.13)$$

To find the number of gas molecules with a speed in the range dv, irrespective of the direction of the velocity, we need the speed distribution function, $n(v)$. The number of allowed states (velocities) between v and $v + dv$, i.e., the number of states between a sphere of radius v and a sphere of radius $v + dv$, as illustrated in Figure 3.7, is:

$$4\pi v^2 dv = G(E)dE \qquad (3.14)$$

For every value of v there is a value of E in $G(E)$, and substituting Equation 3.14 in Equation 3.13 we obtain:

$$n(E)dE = n(v)dv = A4\pi v^2 e^{-\frac{mv^2}{2k_BT}}dv \qquad (3.15)$$

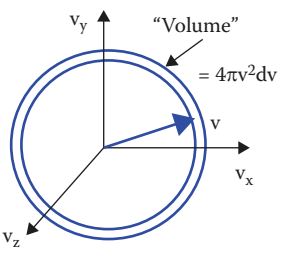


FIGURE 3.7 The velocity is a vector in three-dimensional space, and one needs to take into account that, with the increase of the magnitude of v, the space accessible to a particle increases.

To calculate the value of A, we remember that we can derive the number of molecules per unit volume or volume density of all particles (all velocities/energies) as:

$$\frac{N}{V} = \int_0^\infty n(v)dv$$

or

$$A = \frac{N}{V}\left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}} \qquad (3.16)$$

$$\left(\text{since } \int_0^\infty v^2 e^{-av^2}dv = \frac{1}{4a}\sqrt{\frac{\pi}{a}}\right)$$

From Equations 3.15 and 3.16, the Maxwell-Boltzmann speed distribution is derived as:

$$n(v)dv = \frac{4\pi N}{V}\left(\frac{m}{2\pi k_BT}\right)^{\frac{3}{2}} v^2 e^{-\frac{1}{2}\frac{mv^2}{k_BT}}dv \qquad (3.17)$$

This expression gives us the speed distribution of n particles as a function of their mass and the temperature: it represents the probability that a particle has a speed in the range v to $v + dv$ as illustrated in Figure 3.8. As the temperature increases, the curve

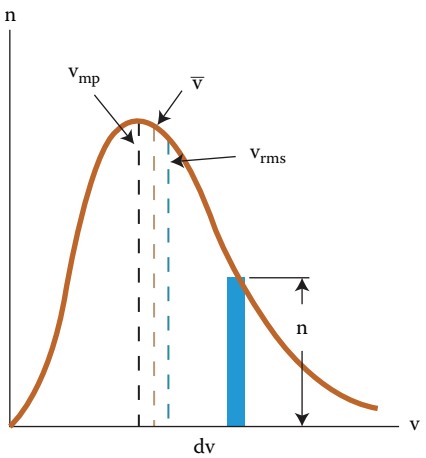


FIGURE 3.8 The Maxwell-Boltzmann distribution of speeds of particles as a function of the temperature and mass of the particles. The most probable speed v_{mp} is the speed at which the distribution curve reaches a peak, $\bar{v}$ is the average speed, and the root mean square speed is v_{rms}.

broadens and extends to higher speeds. Using $dE = mvdv$ ($E = mv^2/2$), the Maxwell-Boltzmann distribution of kinetic energies is:

$$n(E)dE = \frac{2N}{V\sqrt{\pi}}\left(\frac{1}{k_B T}\right)^{\frac{3}{2}}\sqrt{E}\, e^{-\frac{E}{k_B T}}dE \qquad (3.18)$$

Comparing Equation 3.18 with Equation 3.13 yields the 3D density of states function for gas molecules:

$$G(E) = \frac{2N}{V\sqrt{\pi}}\left(\frac{1}{kT_B}\right)^{\frac{3}{2}}\sqrt{E} \qquad (3.19)$$

This function increases smoothly with the square root of the energy.

The most probable speed $v_{mp} = \sqrt{\frac{2k_B T}{m}}$ is the speed at which the speed distribution (Equation 3.17) reaches a peak; this can be calculated from $\frac{dn(v)}{dv}$, setting it to zero, and solving for v. Although Equation 3.17 gives the distribution of speeds or, in other words, the fraction of molecules having a particular speed, we are often more interested in quantities such as the root mean square speed or the average speed of the particles rather than the actual distribution. The root mean square speed is given as:

$$v_{rms} = \sqrt{\frac{\int_0^\infty v^2 n(v)dv}{N/V}} = \sqrt{\frac{3k_B T}{m}} \qquad (3.20)$$

and the average or mean speed $\bar{v}$ is given as:

$$\bar{v} = \sqrt{\frac{\int_0^\infty vn(v)dv}{N/V}} = \sqrt{\frac{8k_B T}{\pi m}} \qquad (3.21)$$

In the case in which the Maxwell-Boltzmann distribution is applied to electrons rather than gas molecules, we must replace m with m_e, the mass of an electron. The Maxwell-Boltzmann distribution, shown in Figure 3.8, is a classical distribution of particles; in quantum statistics, other particle distribution functions are introduced. In the Maxwell-Boltzmann distribution function, it is assumed that all the particles are distinguishable; particles are physically identical but distinguishable in position and trajectory. Particles are considered distinguishable if the distance separating them is large compared with their de Broglie wavelength (see below). Another way of saying this is that the average distance between particles must be large compared with the quantum uncertainty. For an ideal gas, this criterion is certainly fulfilled, but it is not true for electrons and, as we remarked before, this is ultimately the reason why the Drude model fails to explain conductivity correctly: the Maxwell-Boltzmann distribution is not valid for a collection of electrons.

Drude Fails

Above we saw that during an average time τ, electrons travel a mean free path λ, i.e., $\lambda = v_{th}\tau$, and because Drude assumed the electron mean free path, λ, to be equal to the lattice constant, a, which is of the order of 1 nm, this yields a typical value for τ of about 10^{-14} s (because we calculated $v_{th} = \bar{v} \approx v_{rms} = 1.57 \times 10^5$ m/s). Based on Equations 3.8 and 3.21, where we replaced m with m_e and used $\bar{v}$ for calculating v_{th} (Equation 3.21), one derives for the resistivity:

$$\rho = \frac{1}{\sigma} = \frac{m_e v_{th}}{ne^2\lambda} = \frac{m_e}{ne^2 a}\sqrt{\frac{8k_B T}{\pi m_e}} \qquad (3.22)$$

Using the lattice constant, a, for the mean free path and the Maxwell-Boltzmann equation at T = 300 K to calculate v_{th}, values for the resistivity of a metal are obtained that are six times too large. In addition, from Equation 3.22 the temperature dependence of resistivity is determined by v_{th}, which in this model is proportional to $\sqrt{T}$. In practice, the temperature dependence of the resistivity is represented by the empirical relationship:

$$\rho = \rho_0 + \alpha T \qquad (3.23)$$

where ρ_0 is the resistivity at a reference temperature, usually room temperature, and α is the temperature coefficient. An expanded version of Equation 3.23 is known as the Matthiessen rule. This rule is just an approximation (a rule of thumb that works pretty well), not a true physical law. According to Matthiessen rule, the resistivity can be expressed as a sum of

TABLE 3.1 Resistivity Values of Common Metals

Material	ρ_0 ($\mu\Omega$·cm)	α ($\mu\Omega$·cm/K)*	Resistivity at 100°C ($\mu\Omega$·cm)
Aluminum	2.284	0.00390	3.0640
Copper, annealed	1.7241	0.00393	2.5101
Copper, hard-drawn	1.771	0.00382	2.5350
Brass	7.000	0.00200	7.4000
Gold	2.440	0.00340	3.1200
Iron	9.710	0.00651	11.0120
Lead	20.6480	0.00336	21.3200
Nickel	11.000	0.00600	12.2000
Silver	1.590	0.00380	2.3500
Steel	10.400	0.00500	11.4000
Nichrome	100.000	0.00040	100.0800
Platinum	10.000	0.00300	10.6000
Tungsten	5.600	0.00450	6.5000

*Determined at 25°C.

terms resulting from (nearly) independent contributions, for example, $\rho = \rho_0 + \rho_T T + \rho_I C + \rho_{other}$. Here, ρ_0 is the extrapolation of the resistivity to 0 K, $\rho_T T$ is the roughly linear and independent contribution caused by temperature, $\rho_I C$ is the roughly linear contribution caused by solid solution impurities, and ρ_{other} represents the contributions from other scattering centers, such as dislocations and precipitates. Typical values of ρ_0 and α are listed in Table 3.1, along with the calculated resistivity at 100°C.

Experimentally determined resistivity versus temperature plots for a metal, an insulator, and a superconductor are shown in Figure 3.9. Superconductivity is the flow of electric current without resistance. It has been observed in certain metals, alloys, and ceramics at temperatures near absolute zero, and in some cases at temperatures hundreds of degrees above absolute zero. At low temperatures, all materials, other than superconductors, are insulators or metals. For pure metals, the resistivity increases rapidly with increasing temperature, whereas for insulators the resistivity decreases rapidly with increasing temperature [this was first observed by Michael Faraday (1791–1867) in 1833]. Semiconductors have resistivities intermediate between metals and insulators at room temperature. Instead of a square root dependence of temperature, the plot in Figure 3.9 reveals that, for a metal, there is proportionality with T at higher temperatures, and, at low temperatures, the resistivity is proportional to T^5! The latter is known as the Bloch-Gruneisen T^5 law.

From the above observations it is clear that some of the Drude assumptions are very wrong. Moreover, Drude's model cannot explain why one material acts as an insulator and another as a metal. This was all very confusing around Drude's time. Because solids contain a number of atoms and electrons with a similar density, why the large conductivity

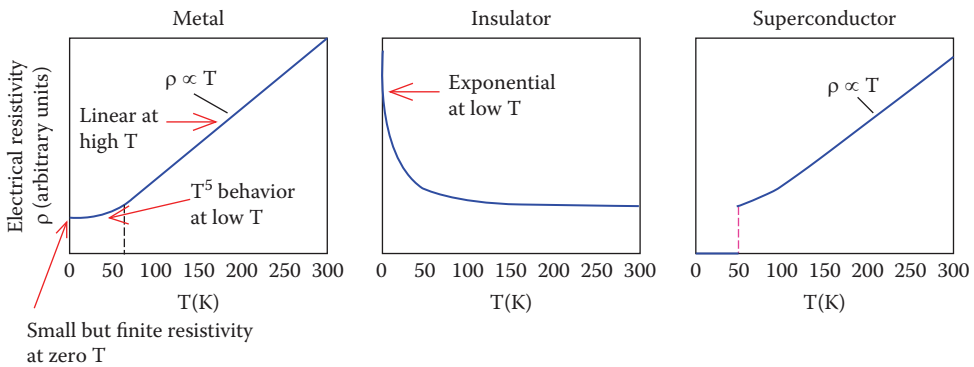


FIGURE 3.9 The three states of solid-state matter as defined by their electrical resistivity in the low temperature limit. The resistivity at low temperatures is finite for a metal, very large for an insulator, and zero for a superconductor.

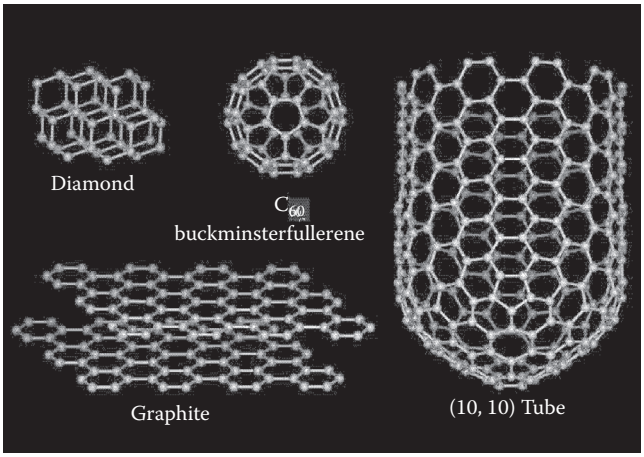


FIGURE 3.10 Carbon may act as an insulator (diamond), as a semimetal (graphite), and as a superconductor (buckminsterfullerene and carbon nanotube).

differences shown in Figure 3.2? More intriguing yet, in the case of carbon, the same material may act as a good insulator (diamond), as a semimetal (graphite), and even as a superconductor (buckminsterfullerenes and carbon nanotubes)* (see Figure 3.10). The answer, we will learn, is that electrons are *fermions*, and only electrons with energy on the order of a few k_BT contribute to the conduction process, at room temperature. A *fermion* is a particle, such as an electron, proton, or neutron, having half-integral spin and obeying statistical rules requiring that not more than one in a set of identical particles may occupy a particular quantum state. For fermions, the Boltzmann distribution must be replaced by a Fermi-Dirac distribution. Because of the wave nature of electrons and the exclusion principle (to be discussed below), the energy distribution of electrons in a metal does not even resemble a Maxwell-Boltzmann distribution. Moreover, the collision between ions and electrons cannot be pictured as that of two hard objects. Instead it involves the scattering of electron waves by lattice ions.

Drude AC Electrical Conductivity and Dielectric Functions

Drude AC Electrical Conductivity In the previous sections we considered the dc conductivity of metals; we now consider their ac electrical conductivity. We assume that the wavelength of the electromagnetic (EM) field is large compared with the electronic mean free path λ, so that electrons "see" a homogeneous field when moving between collisions. In Equation 3.4, we used Newton's second law to launch the change of momentum of a free electron in the presence of an applied dc electrical field. Here we use the same expression to derive the change of momentum of electrons in a time-harmonic electric field, given by:

$$\mathbf{E}(\omega,t) = \mathbf{E}(\omega)e^{-iwt} \tag{3.24}$$

The equation for the momentum per electron is:

$$\frac{d\mathbf{p}(\omega,t)}{dt} = -e\mathbf{E}(\omega,t) - \frac{\mathbf{p}(\omega,t)}{\tau} \tag{3.25}$$

where τ is the relaxation time or the mean free time a randomly picked electron travels before the next collision in a metal. Because $\frac{d\mathbf{p}(\omega,t)}{dt} = -i\omega\mathbf{p}(\omega,t)$, this expression may be rewritten as:

$$\begin{aligned}
\mathbf{p}(\omega,t) &= -\left[\frac{e\mathbf{E}(\omega,t)}{\frac{1}{\tau} - i\omega}\right] \text{ or also} \\
&= -\left[\frac{e\mathbf{E}(\omega,t)}{\frac{1}{\tau^2} + \omega^2}\right]\left(\frac{1}{\tau} + i\omega\right) \\
&= -\left[\frac{e\mathbf{E}(\omega)}{\frac{1}{\tau^2} + \omega^2}\right]\left(\frac{1}{\tau} + i\omega\right)\left(\cos\omega t - i\sin\omega t\right) \\
&= -\left[\frac{e\mathbf{E}(\omega)}{\frac{1}{\tau^2} - \omega^2}\right]\left(\frac{1}{\tau}\cos\omega t + \omega\sin\omega t\right) \\
&\quad + i\left(\omega\cos\omega t - \frac{1}{\tau}\sin\omega t\right)
\end{aligned} \tag{3.26}$$

We analyze this expression now for two different and important situations: 1) with $\omega \ll 1/\tau$, where the electrons can follow the changing electrical field, and 2) with $\omega \gg 1/\tau$, i.e., at very high frequencies

* The latter were not yet discovered at the time; if they had been, this would have led to even more consternation. How can the same material have all these different resistivities?

(optical frequencies we will see), where the electrons cannot follow the fast changing electrical field anymore.

1. With $\omega \ll 1/\tau$:

$$\mathbf{p}(\omega,t) = -\left[\frac{e\mathbf{E}(\omega)}{\frac{1}{\tau^2}}\right]\left(\frac{1}{\tau}\cos\omega t - i\frac{1}{\tau}\sin\omega t\right)$$

$$= -\left[\frac{e\mathbf{E}(\omega)}{\frac{1}{\tau}}\right]e^{-iwt} = -e\tau\mathbf{E}(\omega,t) \qquad (3.27)$$

   At these frequencies, $\mathbf{p}(\omega,t)$ is in phase with $\mathbf{E}(\omega,t)$.

2. With $\omega \gg 1/\tau$:

$$\mathbf{p}(\omega,t) = -\left[\frac{e\mathbf{E}(\omega)}{\omega^2}\right](\omega\sin\omega t + i\omega\cos\omega t)$$

$$= -\left[\frac{e\mathbf{E}(\omega)i\omega}{\omega^2}\right](\cos\omega t - i\sin\omega t)$$

$$= -\left[\frac{e\mathbf{E}(\omega)i\omega}{\omega^2}\right]e^{-iwt} = -\left[\frac{e\mathbf{E}(\omega,t)i}{\omega}\right]$$

$$= \frac{e\mathbf{E}(\omega,t)}{i\omega} \qquad (3.28)$$

At very high frequencies, where the electrons are too slow to follow the very fast changing electrical field vector, $\mathbf{p}(\omega,t)$ is out of phase with $\mathbf{E}(\omega,t)$; moreover, $\mathbf{p}(\omega,t)$ tends to zero. The transition between these two behaviors occurs when the ac frequency exceeds the collision frequency τ. With a typical value for τ of metal of about 10^{-14} s, that transition frequency is at roughly 10^{14} Hz ($\tau \sim 10^{-14}$ s, $v\lambda = c$ or $v = \frac{3.10^{10}\text{cm/s}}{5000\cdot10^{-8}\text{cm}} \approx 10^{14}$ Hz), corresponding to optical frequencies (light waves). The equations derived here apply for the ac behavior of metals as well their interaction with light (metal optics).

Relying on Equation 3.3, rendered in vector format as $\mathbf{J}(\omega) = -env_{dx}$, and given that $\mathbf{p}(t) = m_e\mathbf{v}(t)$, we can also write:

$$\mathbf{J}(\omega) = -\frac{en\mathbf{p}(\omega)}{m_e} = \frac{\left[\frac{e^2n}{m_e}\right]\mathbf{E}(\omega)}{\frac{1}{\tau} - i\omega} = \sigma(\omega)\mathbf{E}(\omega) \qquad (3.29)$$

so that the Drude complex AC conductivity is calculated as:

$$\sigma(\omega) = \frac{\sigma_0}{1 - i\omega\tau} \qquad (3.30)$$

which simplifies to the DC conductivity, i.e., σ_0 ($= \frac{ne^2\tau}{m_e}$, Equation 3.7), in the case of very low ac frequencies ($\omega \ll 1/\tau$).

The real and imaginary parts of the complex conductivity are:

$$\sigma' = \frac{\sigma_0}{1 + \omega^2\tau^2}(\text{Re}) \qquad (3.31)$$

and

$$\sigma'' = \frac{\sigma_0\omega\tau}{1 + \omega^2\tau^2}(\text{Im}) \qquad (3.32)$$

The real (σ') and imaginary (σ'') parts of the complex conductivity $\sigma(\omega)$ are plotted versus $\omega\tau$ in Figure 3.11. The maximum in σ'' is called the Drude peak and is characteristic for each metal.

We consider now what happens with the current and the conductivity as a function of frequency. At very low ac frequencies ($\omega \ll 1/\tau$), electrons have many collisions before the direction of the wave changes; this situation corresponds to the Ohm's law

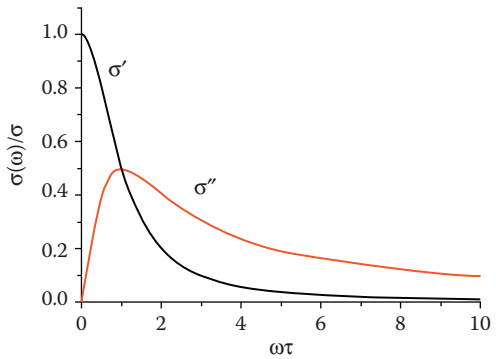


FIGURE 3.11 Frequency dependency of the real (σ') and imaginary (σ'') in parts of the conductivity. The maximum in σ'' at $\sigma''/\sigma_0 = 0.5$ is the Drude peak at $\omega\tau = 1$ and is a characteristic of a metal. The DC conductivity σ_0 is reached at $\omega\tau = 0$ ($\sigma''/\sigma_0 = 1$).

regime—**J** follows **E**—and σ is real. At very high frequencies ($\omega \gg 1/\tau$), electrons might have only one collision or less when the direction of the ac field is changed. In this case **J** is imaginary and out of phase with **E**, and σ is also imaginary. As we noted above, the transition from one regime to the other occurs at optical frequencies, and qualitatively we can say that with $\omega\tau \ll 1$, electrons are in phase and reirradiate, i.e., they are reflected and the metal appears shiny. With $\omega\tau \gg 1$, electrons are out of phase as they are too slow, there is less interaction, and we have transmission. More specifically, free electrons do not influence E-fields in metals with frequencies greater than that of visible light, or the electron gas is transparent in the UV range! No energy is absorbed from the field in this range, and no joule heating occurs.

Next, we will investigate how the permittivity or dielectric constant of a medium changes as a function of frequency of the applied electrical field (again from ac fields to electromagnetic radiation).

Dielectric Functions and an Introduction to Metal Optics The permittivity or dielectric constant of a medium describes how an electric field both affects and is affected by that medium and can be looked at as the quality of a material that allows it to store electrical charge. The dielectric constant ε or permittivity (*As/Vm*) is a materials-dependent "constant," and its frequency dependence defines the so-called dielectric function $\varepsilon(\omega)$. In the most general terms, dielectric functions are dependent on both frequency (ω) and wave-vector **k**: $\varepsilon(\omega,\mathbf{k})$. To understand the dielectric behavior and optical properties of metals better, we need to explain how plasma oscillations come about and analyze how the dielectric "constant" ε of a metal changes with frequency. A plasma, in general, is a medium with equal amounts of positive and negative charges, of which at least one charge type is mobile. In case of a plasma in a metal, the mobile charges are the free electrons, and these charges are balanced by the positive, immobile, metal ion cores. Drude's free electron gas (FEG), with an ac field applied, can exhibit collective longitudinal oscillations of the plasma because the displacement of all the electrons against the ion bodies of the material induces a dipole moment

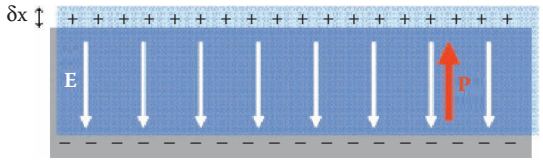


FIGURE 3.12 Drude bulk plasma oscillation in a metal film. **E** is the electrical field, and **P** is the polarization. The entire free electron gas is displaced over a small distance, δ*x*.

and an electric field opposing that displacement, as shown in Figure 3.12. In longitudinal oscillations, the displacement is in the same direction as the wave motion, as in sound waves. To appreciate better how oscillations of free electrons are induced by a time harmonic electrical field $\mathbf{E}(\omega,t) = \mathbf{E}(\omega)e^{-i\omega t}$, we rewrite Equation 3.25 for one dimension with $\mathbf{p}_x(\omega,t) = m_e\mathbf{v}_x(\omega,t) = m_e d_x/dt$ as:

$$m_e \frac{d^2x}{dt^2} + m_e\gamma\frac{dx}{dt} = -e\mathbf{E}(t) = -e\mathbf{E}(\omega)e^{-i\omega t} \qquad (3.33)$$

In the Drude model, electrons are free, and a damping factor γ comes about only because of electron scattering. The damping factor is related to the scattering constant as $1/\gamma = \tau$ (as we saw, typically ~10^{-14} s in a metal). The solution of Equation 3.33—the displacement *x* of the entire free electron gas—is then given as:

$$x(t) = \frac{e\mathbf{E}(\omega)}{m_e(\omega^2 + i\gamma\omega)} \qquad (3.34)$$

Polarization comes about as a result of small displacement of charges in an electrical field. The macroscopic polarization density **P** (C/m^2), illustrated in Figure 3.12, is a vector field that represents the density of permanent or induced electric dipole moments and is given by the product of the displacement $x(t)$ and the electron density ne(**P** = – nex) or:

$$\mathbf{P} = -\frac{ne^2\mathbf{E}(t)}{m_e(\omega^2 + i\gamma\omega)} \qquad (3.35)$$

The polarization **P** is also related to the electrical field as:

$$\mathbf{P}(\omega,t) = \chi_e\varepsilon_0\mathbf{E}(\omega,t) \qquad (3.36)$$

where χ_e is the electric susceptibility tensor of the material, a proportionality constant relating the electrical field **E** to the induced dielectric polarization

density **P**, and ε_0 is the permittivity of free space with a value of 8.854 pF/m. In the case of a linear homogeneous and isotropic material, χ_e becomes a scalar constant. The polarization **P** is further linked to the electric displacement field **D**, expressed in coulombs per square meter (C/m^2), and to the dielectric constant $\varepsilon(\omega,\mathbf{k})$ or permittivity (As/Vm), a materials-dependent constant, as:

$$\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P} = (1 + \chi_e)\varepsilon_0\mathbf{E} = \varepsilon(\omega,\mathbf{k})\mathbf{E} \qquad (3.37)$$

In the expression $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$ we have separated the electric displacement **D** in its materials (**P**) and vacuum parts ($\varepsilon_0\mathbf{E}$). The permittivity of a material is usually given as a relative permittivity $\varepsilon_r(\omega)$ (also called the dielectric constant). The permittivity $\varepsilon(\omega,\mathbf{k})$ of a medium is an intensive parameter and is calculated by multiplying the relative permittivity by ε_0, or $\varepsilon(\omega) = \varepsilon_0\varepsilon_r(\omega)$. The permittivity of air is $\varepsilon_{air} = 8.876$ pF/m, so the relative permittivity of air is $\varepsilon_{r,air} = 1.0005$, and for vacuum it is 1 by definition. From Equation 3.37, it then also follows that $\varepsilon_r(\omega) = 1 + \chi_e$. The electric displacement field, **D**, is in turn related to the electric field **E** (in units of V/m) by the constitutive relation:

$$\mathbf{D} = \varepsilon(\omega,\mathbf{k})\,\mathbf{E} \qquad (3.38)$$

Using Equation 3.36 for the polarization **P**, we can rewrite Equation 3.37 as:

$$\mathbf{D} = \varepsilon(\omega,\mathbf{k})\mathbf{E} = \varepsilon_0\mathbf{E} + \mathbf{P} = \varepsilon_0\mathbf{E} - \frac{ne^2\mathbf{E}}{m_e(\omega^2 + i\gamma\omega)} \qquad (3.39)$$

From Equation 3.39 we arrive at the following expression for the complex dielectric function:

$$\varepsilon(\omega,\mathbf{k}) = \varepsilon_0\left[1 - \frac{ne^2}{m_e\varepsilon_0(\omega^2 + i\gamma\omega)}\right] = \varepsilon_0\left[1 - \frac{\omega_p^2}{(\omega^2 + i\gamma\omega)}\right] \qquad (3.40)$$

The plasma frequency ω_p in Equation 3.40 is defined as:

$$\omega_p = \sqrt{\frac{ne^2}{m_e\varepsilon_0}} \qquad (3.41)$$

where n is the density of electrons in the metal. To obtain $\varepsilon_r(\omega)$, the relative dielectric response of a material to electromagnetic waves at various frequencies, one divides by ε_0 because $\varepsilon(\omega,\mathbf{k})/\varepsilon_0 = \varepsilon_r(\omega,\mathbf{k})$. The corresponding real and imaginary components of the complex dielectric function are:

$$\varepsilon_r'(\omega,\mathbf{k}) = 1 - \frac{\omega_p^2\tau^2}{1 + \omega^2\tau^2} \qquad \text{Re} \qquad (3.42)$$

and

$$\varepsilon_r''(\omega,\mathbf{k}) = \frac{\omega_p^2\tau}{\omega(1 + \omega^2\tau^2)} \qquad \text{Im} \qquad (3.43)$$

In a metal, the damping term γ is just the electron collision rate, which is the inverse of the mean electron collision time, τ, i.e., $\gamma = \tau^{-1}$. As stated before, the collision rate can be quite rapid—tens of femtoseconds. But for optical frequencies (e.g., for $\lambda = 500$ nm, $\omega = 2\pi c/\lambda = 3.8 \times 10^{15}$ rad/s) $(\omega\tau)^2 >> 1$. Under this approximation, we find:

$$\varepsilon_r'(\omega,\mathbf{k}) \approx 1 - \frac{\omega_p^2}{\omega^2} \qquad (3.44)$$

and

$$\varepsilon_r''(\omega,\mathbf{k}) \approx \frac{\omega_p^2}{\omega^3\tau} = \frac{\omega_p^2\gamma}{\omega^3} \qquad (3.45)$$

This approximation may break down in the far-infrared spectral region, where damping can be significant. Note that damping (γ) is absolutely necessary to have an imaginary part of $\varepsilon_r(\omega)$. The dispersion of the real part of the dielectric function $\varepsilon_r'(\omega,\mathbf{k})$ (Equation 3.44) is plotted in Figure 3.13. The Drude model predicts a monotonous decrease of $\varepsilon_r'(\omega,\mathbf{k})$ for decreasing frequency, and experiments confirm

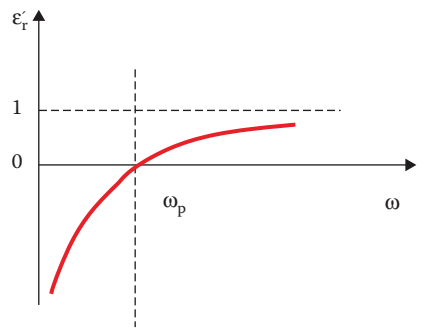


FIGURE 3.13 The dielectric function: ε_r' as a function of ω. The dielectric constant ε_r' becomes zero when $\omega = \omega_p$, and this supports free longitudinal collective modes for which all electrons oscillate in phase.

that $\varepsilon_r'(\omega,\mathbf{k})$ becomes zero when $\omega = \omega_p$. When $\varepsilon_r'(\omega)$ becomes zero at $\omega = \omega_p$, the material can support free longitudinal collective modes for which all electrons oscillate in phase. Such a collective charge oscillation is called a *plasmon*. Note that in the limit of γ going to 0, Equation 3.40 leads to the same result as shown in Equation 3.44 (no imaginary component to the dielectric constant at $\gamma = 0$). We come back to the field of plasmonics in Chapter 5 on photonics, and at that point we will not only analyze plots of $\varepsilon_r(\omega)$ for metals but also of $n(\omega)$ (refractive index), $\alpha(\omega)$ (absorption), and $R(\omega)$ (reflectance) for all types of materials.

With reference to Figure 3.12, we can derive the plasma frequency ω_p in yet another way. Let us look at the case where $\varepsilon_r'(\omega) = 0$ at $\omega = \omega_p$; because $\varepsilon_r'(\omega) = 1 + \chi_e'$, this means that at this frequency, $\chi_e' \approx -1$, and with the polarization, $\mathbf{P}(\omega,t) = \chi_e\varepsilon_0\mathbf{E}(\omega,t)$ **P** (Equation 3.36), this results in $\mathbf{P}(\omega,t) \approx -\varepsilon_0\mathbf{E}(\omega,t)$. The displacement δx of the electron gas of density n creates an electric field $\mathbf{E} = -\mathbf{P}/\varepsilon_0 = ne\delta x/\varepsilon_0$; this opposing field acts as the restoring force ($F_r = -eE$) on the electron gas. The equation of motion resulting from this restoring force F_r in such a simple electron gas oscillator is given as:

$$F_r = -eE = -\frac{ne^2\delta x}{\varepsilon_0} = m_e\frac{\partial^2(\delta x)}{\partial^2 t^2} \qquad (3.46)$$

This results in oscillations at the plasma frequency of $\omega_p = \sqrt{\dfrac{ne^2}{m_e\varepsilon_0}}$ (Equation 3.41).

The plasma frequency ω_p comes with a free space wavelength of:

$$\lambda_p = \frac{2\pi c}{\omega_p} \qquad (3.47)$$

In Table 3.2 we compare the plasma frequency and the free space wavelength for a generic metal and a generic semiconductor. Using Equation 3.41 we can now calculate ω_p for a variety of materials. For silver, for example, with a σ of 6.2×10^7 Ω-m, one obtains $\omega_p = 9.65 \times 10^{14}$ Hz (= 311 nm or 4 eV). The lower the resistivity (higher n), the higher the plasma frequency. Consequently, plasma frequencies ω_p are in the optical range (UV) for metals and in the THz to the infrared region for semiconductors and insulators. Metals reflect light in the visible region and are transparent at very high frequencies (UV and x-rays). With a free space wavelength less than λ_p, a wave propagates; when it is longer, the wave is reflected ($\omega > \omega_p$, or $\lambda < \lambda_p$).

Table 3.2 Plasma Frequency (ω_p) and Free Space Frequency (λ_p) for a Generic Semimetal and Semiconductor

Property	Metal	Semiconductor
Electron density: n, cm^{-3}	10^{22}	10^{18}
Plasma frequency: ω_p, Hz	5.7×10^{15}	5.7×10^{13}
Space frequency: λ_p, cm	3.3×10^{-5}	3.3×10^{-3}
Spectral range	UV	Infrared region

In a metal where both bound electrons [$\varepsilon_B(\omega)$] and conduction electrons [$\varepsilon_r(\omega)$] contribute an effective dielectric constant, $\varepsilon_{Eff}(\omega)$ must be defined as:

$$\varepsilon_{Eff}(\omega) = \varepsilon_B(\omega) + \varepsilon_r(\omega) \qquad (3.48)$$

where $\varepsilon_r(\omega)$ is the complex dielectric constant at $\omega\tau >> 1$, given by (see Equations 3.44 and 3.45):

$$\varepsilon_r(\omega) = \varepsilon_r'(\omega) + i\varepsilon_r''(\omega) = 1 - \frac{\omega_p^2}{\omega^2} + i\frac{\omega_p^2}{\omega^3\tau} \qquad (3.49)$$

Plasma oscillations can be excited in dielectric films as well. A dielectric material is a nonconducting substance whose bound charges are polarized under the influence of an externally applied electric field. In this case, although qualitatively the same effect as in a metal, the oscillations of the bound electrons are with respect to the immobile ions instead of with respect to the films boundary, and a restoring force related to the strength of the bond of the electrons to those ions must be introduced. In Chapter 5 on photonics, after introducing the Maxwell equations, the Drude plasmon model will be upgraded with the more complete Drude-Lorentz model that includes such a restoring force term. The restoring force pulls charges back in their equilibrium position. The Drude-Lorentz model without the restoring force reduces back to the Drude model, in which conduction electrons are not bound to atoms. There we will also calculate the effective dielectric constant $\varepsilon_{Eff}(\omega)$ (with bound and free electron contributions) of Equation 3.48.

Specific Heat Capacity of Metal and Insulators

Different materials require different amounts of heat to increase their temperature. Heat capacity per unit mass of a substance is known as specific heat, and as we will see, it is a measure of the number of degrees of freedom of the system. When using classical models, as in the case of electrical conductivity, things went wrong for calculations of the specific heat capacity of metals and insulators.

Based on the Drude assumption that electrons in a metal behave as a monatomic gas of N classical particles, they should be able to take up translational kinetic energy when the metal is heated. According to the equipartition principle of energy in a gas with N particles, the electron internal energy U, at temperature T, expressed per mole is:

$$\frac{U}{n} = \frac{3}{2}\frac{N}{n}k_B T = \frac{3}{2}N_A kT = \frac{3}{2}RT \qquad (3.50)$$

with n the number of moles in the system, N_A Avogadro's number, R the ideal gas constant, and k_B Boltzmann's constant. Remember also from thermodynamics that the first derivatives of the fundamental thermodynamic equations, in U (internal energy) or S (entropy), correspond to the intensive parameters T, P, and μ (chemical potential), whereas second derivatives correspond to important materials properties such as the molar heat capacity, or:

$$C_v = T\left(\frac{\partial s}{\partial T}\right)_v = \frac{T}{n}\left(\frac{\partial S}{\partial T}\right)_v = \frac{1}{n}\left(\frac{\partial Q}{\partial T}\right)_v \qquad (3.51)$$

This equation basically tells us that the molar heat capacity at constant volume is the quasistatic heat flux per mole required to produce a unit increase in the temperature for a system maintained at constant volume. Materials with high specific heat capacity (C_v) require more energy to reach a given temperature. In the case of a metal, one expects contributions to the heat capacity from both the lattice and the free electrons. The electronic contribution to the molar specific heat capacity at constant volume, C_v, from Equation 3.50, is:

$$C_{v,el} = \frac{\delta}{\delta T}\left(\frac{U}{n}\right)_v = \frac{3}{2}R \qquad (3.52)$$

or about 12.5 J/(mol·k).

Atoms in a lattice usually have no translational energy, and as temperature increases only vibrational energy increases. In the case of a monovalent metal, the lattice specific heat contribution, $C_{v,lat}$, equals $3R$. This follows from the fact that a classical atom oscillator has 3 degrees of freedom in vibration, or U = 3NkT. The vibration of a classical 1D harmonic oscillator is U = kT, with $\frac{1}{2}kT$ for kinetic and $\frac{1}{2}kT$ for potential energy (Figure 3.14). From the latter we expect $C_{v,lat}$ to be constant at $3R$ or 25 J/(mol·k), the so-called Dulong-Petit law. Experimentally, C_v, the sum of vibrational and electronic contributions ($C_v = C_{v,el} + C_{v,lat}$), is proportional to T^3 at low T and approaches a constant $3R$ at high T (see Figure 3.15). Diamond reaches the Dulong-Petit value 25 J/(mol·k) only at high temperature, whereas lead reaches the Dulong-Petit value at relatively low temperature.

From the above, at high temperature, electrons are expected to contribute to the lattice heat capacity for a total $C_v = C_{v,el} + C_{v,lat}$ or $\frac{9}{2}R$, which is obviously a different result from Dulong-Petit's $3R$! The total C_v for metals is found to be only slightly higher than that for insulators, which only feature lattice contributions. The question is then: where is the electronic contribution? The absence of a measurable contribution by electrons to the C_v was historically

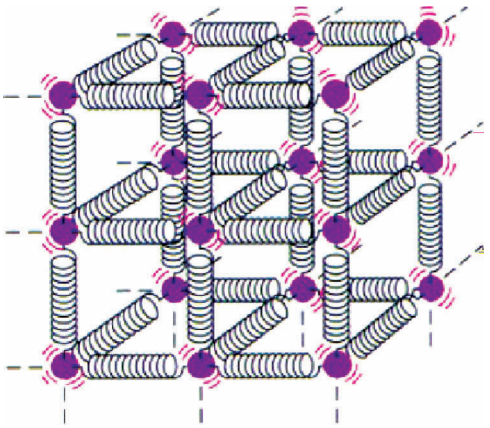


FIGURE 3.14 C_v: energy needed to raise T of 1 mol by 1 K at constant V, volume. The internal energy resides in vibrations of the solid constituents. The vibration of a classical 1D harmonic oscillator is U = kT. In a solid there are three perpendicular modes of vibration or three harmonic oscillators, and the total energy of a solid is thus U = 3NkT = 3RT.

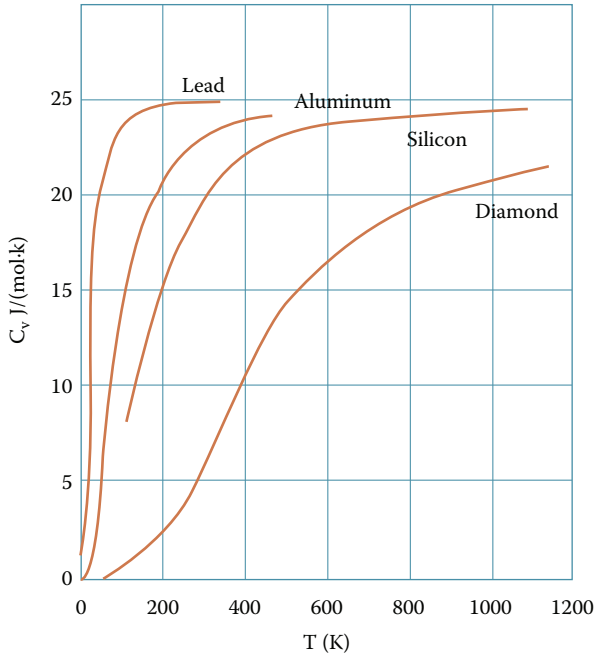


FIGURE 3.15 Dulong-Petit law with C_v = 3R or 25 J/(mol·k) at high temperatures.

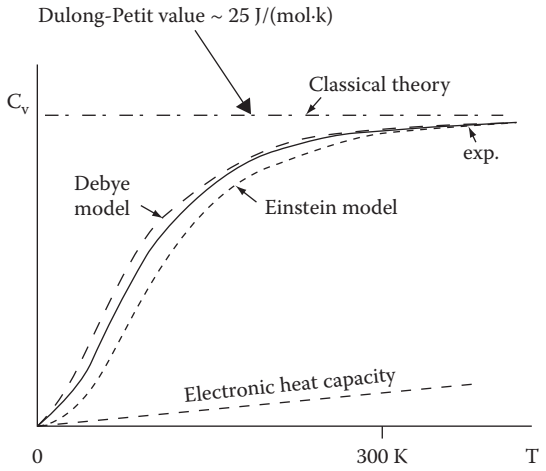


FIGURE 3.16 Schematic representation of the temperature dependence of the molar heat capacity, experimental and according to four models. For the Debye and Einstein models, see further below.

one of the major objections to the classical free electron model: because electrons are free to carry current, why would they not be free to absorb heat energy? Conduction electrons only contribute a small part of the heat capacity of metals, but, as we will see below, they are almost entirely responsible for the thermal conductivity.

In the above model, we used the principle of equipartition of the energy, which we borrowed from classical ideal gas theory. But, as Planck discovered in 1901, a vibrating system can only take up energy in quanta, the size of which is proportional to the vibration energy such that E = hν, where ν is frequency and *h* is the Planck constant. The chance that an atom vibrator can pick up an energy *h*ν is proportional to $e^{-h\nu/kT}$, and the average energy of a vibrating atom, at low temperature, may fall so low that the value of *kT* is now small compared with the size of the quantum *h*ν, and the probability factor $e^{-h\nu/kT}$ is dramatically decreased. Therefore, Dulong-Petit is incorrect at low temperatures as vibrations of successively lower frequencies fail to become excited. Atoms in a crystal do not obey Maxwellian statistics at low temperatures, and one needs to invoke the Einstein or Debye law instead. The reason that electrons do not contribute to the heat capacity at room temperature is, just like in the case of electrical conductivity, that they are fermions and cannot be treated as an ordinary Maxwell-Boltzmann gas. At room temperature, only the very few electrons in the so-called Maxwell-Boltzmann tail of the Fermi distribution can contribute to the heat capacity. Thus, at ordinary temperatures the electronic heat capacity is almost negligible, and it is the atomic vibration, i.e., the contribution of phonons, that dominates. In Figure 3.16 we show the experimental molar heat capacity as a function of temperature and the expectations for those values from classical theory and from the Einstein and Debye models, which are discussed at the end of this chapter. It is seen here that the electronic contribution $C_{v,el}$ varies linearly with temperature.

Thermal Conductivity

Heat conduction is the transfer of thermal energy from a hot body to a cold body. In a solid both electrons and phonons can move, and depending on the material involved, one or the other tends to dominate. We intuitively expect electrical and thermal conductivities somehow to be linked; from experience we know that, in general, good electrical conductors are also good thermal conductors. The connection between electrical and thermal conductivities for metals was first expressed in the Wiedemann-Franz law in 1853, suggesting that electrons carry thermal energy and electrical charge. Insulators are often transparent, and conducting

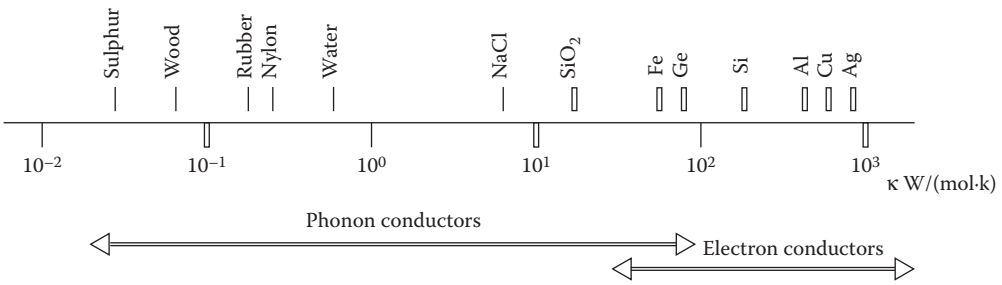


FIGURE 3.17 Phonon and electron conductors.

metals are reflecting and shiny when polished. Most metals are shiny because when light strikes a metal, the light is scattered by the moving electrons. For an exception, think about diamond, which is a transparent insulator but conducts heat better than aluminum or copper. For a glaring difference between heat and electrical conduction, consider that electrical conductivity of materials spans 25 orders of magnitude, whereas thermal conductivity only spans about four orders. In metals, heat is mostly transferred by electrons, but in electrical insulators, there are few free electrons, so the heat must be conducted in some other way, i.e., lattice vibrations or phonons. Thus, materials are divided into phonon conductors and electron conductors of heat as illustrated in Figure 3.17.

There is a major difference between heat conduction by electrons and by phonons; for phonons, the number changes with the temperature, but the energy is quantized, whereas for electrons, the number is fixed, but the energy varies. We will analyze now in more detail what that relation between electronic conductivity, σ, and thermal conductivity, κ is, and will discover that, when using classical models, the experimental results again cannot be properly explained.

Recall the electrical result $\mathbf{J} = \sigma\mathbf{E}$ (Equation 2.2) and also $\mathbf{J} = \sigma\mathbf{E} = \sigma\dfrac{dV}{dx}$. The thermal equivalent for this expression is Fourier's law for heat conduction, Q, which states that:

$$Q = \kappa A\frac{T_h - T_c}{L} = \kappa A\frac{dT}{dx} \tag{3.53}$$

with T_h the hot temperature and T_c the cold temperature, κ the thermal conductivity, L the thermal conduit path-length, and A its cross-sectional area. From the first law of thermodynamics (heat conservation), we know that the rate of heat conduction must equal the rate of change of energy storage:

$$\kappa\frac{\partial^2 T}{\partial x^2} = C_{v,el}\frac{\partial T}{\partial t} \tag{3.54}$$

The specific heat capacity, in general, is made up of a phonon and an electron term ($C_v = C_{v,el} + C_{v,lat}$). In most metals, the contribution of the electrons to heat conductivity greatly exceeds that of the phonons, and the phonon term can be neglected ($C_v = C_{v,el}$). Because of electrical neutrality, equal numbers of electrons move from hot to cold as the reverse, but their thermal energies are different.

Equations 3.53 and 3.54 only apply under the conditions that:

- t >> scattering mean free time of the energy carriers (τ)
- L >> scattering mean free path of the energy carriers (λ)

These conditions break down for applications involving thermal transport in small length/time scales, e.g., nanoelectronics, nanostructures, NEMS, ultrafast laser processing, etc. (see below and Volume III, Chapter 7 on scaling). For now, however, we are interested in the relation between electronic conductivity, σ, and thermal conductivity, κ, for somewhat larger systems.

To arrive at the expression for thermal conductivity, κ, we inspect Figure 3.18 and consider a metal bar [area A is a unit area (A = 1) here] with a temperature gradient dT/dx where we are interested in a small volume of material, with a length 2λ, with λ the mean free path between collisions. The idea is that an electron must have undergone a collision in this space and hence will have the energy/temperature of this location. To calculate the energy

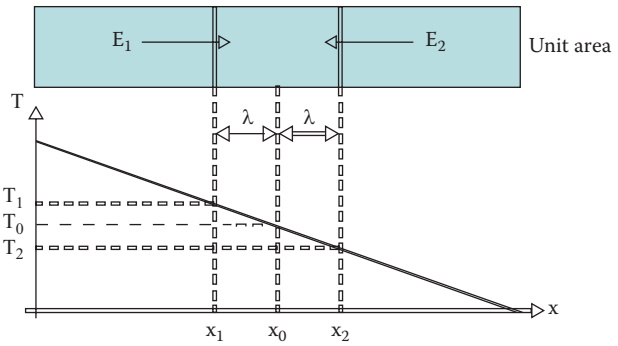


FIGURE 3.18 Fourier's Law for heat conduction: we are interested in a small volume of material, with a length 2λ, where λ is the mean free path between collisions with the lattice.

flowing per unit time per unit area from left to right (E_1), we multiply the number of electrons N crossing by the energy of one electron:

$$\text{energy} = \frac{1}{2}kT_1 = \frac{1}{2}k\left[T_0 + \lambda\left(-\frac{dT}{dx}\right)\right]$$

$$\text{number} = \frac{1}{6}nv_{dx} \qquad (3.55)$$

$$E_1 = \frac{nv_{dx}}{2}\frac{1}{2}k\left(T_0 - \lambda\frac{dT}{dx}\right)$$

with v_{dx} the drift velocity in the x-direction, and dT/dx a thermal gradient in the x-direction. We know that for charge neutrality the same number of electrons must flow in the opposite direction to maintain charge neutrality, but their energy per electron is lower.

Based on:

$$E_2 = \frac{nv_{dx}}{2}\frac{1}{2}k\left(T_0 + \lambda\frac{dT}{dx}\right) \qquad (3.56)$$

the thermal energy transferred per unit time per unit area is:

$$Q = E_1 - E_2 = -\frac{nv_{dx}}{2}k\lambda\frac{dT}{dx} \qquad (3.57)$$

Comparing this result with Equation 3.53 (with A = 1), we obtain:

$$\kappa = \frac{1}{2}nv_{dx}k\lambda \qquad (3.58)$$

This says that the thermal conductivity is larger if there are more electrons (n large), the electrons move faster (v_{dx} large), and they move more easily (large λ with fewer collisions). Because $N_Ak = R$ with R the ideal gas constant, $N/N_A = n$ (number of moles of electrons), and the electronic contribution to the molar specific heat capacity at constant volume, $C_{v,el}$, from Equation 3.52 is $3/2R$. The bulk heat conductivity of a solid per mole (n = 1) can then be rewritten as:

$$\kappa = \frac{1}{3}C_{v,el}v_{dx}\lambda \text{ or also } \frac{1}{3}C_{v,el}v_{dx}^2\tau \text{ sine } \lambda = v_{dx}\tau \quad (3.59)$$

As mentioned above, in most metals the conduction by electrons greatly exceeds that of the phonons; typically the phonon contribution at room temperature is only 1% of the electron contribution.

In 1853, long before Drude's time, Gustav Wiedemann and Rudolf Franz published a paper claiming that the ratio of thermal and electrical conductivities of all metals has almost the same value at a given temperature:

$$\frac{\kappa}{\sigma} = \frac{C_{v,el}mv_{dx}^2}{3ne^2} = \frac{3}{2}\left(\frac{k_B}{e}\right)^2 T \qquad (3.60)$$

calculated through Drude's application of classical gas law: $C_{v,el} = \frac{3}{2}nk_B$ and $\frac{1}{2}mv_{dx}^2 = \frac{3}{2}k_BT$. Ludwig Lorenz realized in 1872 that this ratio scaled linearly with temperature, and thus a Lorenz number, L, was defined as:

$$\frac{\kappa}{\sigma T} \equiv L \qquad (3.61)$$

which is very nearly constant for all metals (at room temperature and above). A typical value for L, say for Ag, is 2.31 10^{-8} WΩK^{-2} at 0°C and 2.37 10^{-8} WΩK^{-2} at 100°C (see Table 3.3).

Although Equation 3.61 is the correct relationship, we now know that the value for L calculated from it is wrong. However, Drude, using classical values for the electron velocity v_{dx} and heat capacity $C_{v,el}$, somehow got a number very close to the experimental value. But how lucky that Drude dude was: by a tremendous coincidence, the error in each term he made was about two orders of magnitude … in the opposite direction [the electronic $C_{v,el}$

TABLE 3.3 Some Typical Lorenz Numbers

Lorenz Number L in 10^{-8} WΩK^2		
Metal	273 K	373 K
Ag	2.31	2.37
Au	2.35	2.40
Cd	2.42	2.43
Cu	2.23	2.33
Ir	2.49	2.49
Mo	2.61	2.79
Pb	2.47	2.56
Pt	2.51	2.60
Sn	2.52	2.49
W	3.04	3.20
Zn	2.31	2.33

is 100 times smaller than the classical prediction, but $(v_{dx})^2$ is 100 times larger]. So the classical Drude model gives the prediction:

$$L_{Drude} = 1.12\,10^{-8}\,W\Omega K^{-2} \tag{3.62}$$

But in Drude's original paper, he inserted also an erroneous factor of two, as a result of a mistake in the calculation of the electrical conductivity. So he originally reported:

$$L = 2.24 \times 10^{-8}\,W\Omega K^{-2} \tag{3.63}$$

The correct value for the Lorenz number, L, derived from quantum mechanics is $\frac{\pi^2 k_B^2}{3e^2}$ or 2.45×10^{-8} WΩK^2 (see Equation 3.340 below). So although Drude's predicted electronic heat capacity was far too high (by a factor of 100!), his prediction of L made the free electron gas (FEG) model seem more impressive than it really was and led to a general acceptance of the model.

Hall Coefficients

Conductivity measurements do not yield information about the sign of charge carriers; for this we need the Hall effect. The Hall effect uses current and a magnetic field to determine mobility and the sign of charge carriers. The Hall effect, discovered in 1879 by American physics graduate student (!) Edwin Hall, is simple to understand.[1] With reference to Figure 3.19, consider a fairly strong magnetic field **B** (~2000 Gauss) applied perpendicular to a thin metal film (thickness is d and width is w) carrying a current I. The magnetic Lorentz force is

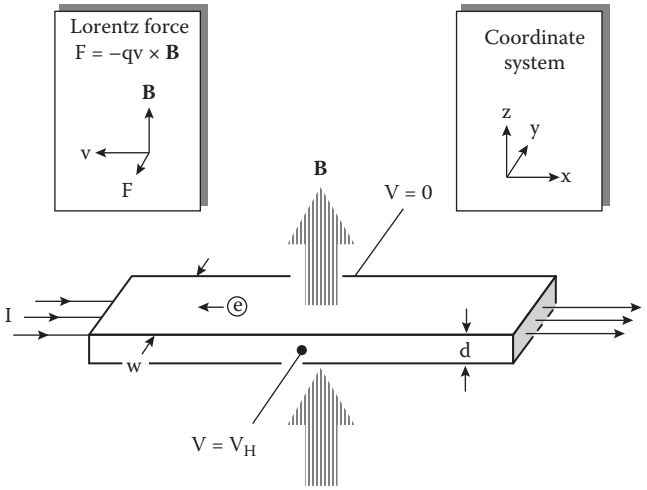


FIGURE 3.19 The Hall effect. Magnetic field **B** is applied perpendicular to a thin metal film sample carrying current I. V_H is the Hall voltage.

normal to both the direction of the electron motion and the magnetic field. The path of the charge carriers shifts as a result of this Lorentz* force, and, as a consequence, the Hall voltage, $V_H = E_y w$ (where E_y is the Hall field strength), builds up until it prevents any further transverse displacement of electrons. In other words, the Hall field increases until it is equal to and opposite of the Lorentz force. The orientation of the fields and sample is illustrated in Figure 3.19.

In mathematical terms, the above translates as follows. With both electric and magnetic fields present, a charge carrier with charge q experiences a Lorentz force in the lateral direction:

$$\mathbf{F}_L = q(\mathbf{v}_d \times \mathbf{B}_z) \tag{3.64}$$

We do use the drift velocity v_d of the carriers because the other velocities (and the forces caused by these components) cancel to zero on average. Note that instead of the usual term "electron," the term "charge carrier" is used here because in principle an electrical current could also be carried by charged particles other than electrons, e.g., positively charged ions or holes (missing electrons). The vector product in Equation 3.64 ensures that the Lorentz force is perpendicular to $\mathbf{v}_d$ and $\mathbf{B}_z$. For the geometry assumed in Figure 3.19, the Lorentz force $\mathbf{F}_L$ has only a component in the y-direction, and we can use a

* This is the Lorentz we encounter again in Chapter 5 (Figure 5.50), not the Lorenz from the Lorenz number L.

scalar equation for the Lorentz force: $F_L = -qv_dB_z$ or with Equation 3.11 (where we replaced μ_e by μ_q for generality):

$$\mathbf{F}_L = -q\mu_q E_x \mathbf{B}_z \tag{3.65}$$

As more and more carriers are deflected by the magnetic force, they accumulate on one side of the thin-film conductor, and this accumulation of charge carriers leads to the "Hall field" $\mathbf{E}_y$ that imparts a force opposite to the Lorentz force. The force from that electrical field $\mathbf{E}_y$ in the y-direction (which is of course $q \cdot \mathbf{E}_y$) must be equal to the Lorentz force with opposite signs. This way we obtain:

$$\begin{aligned} F_L &= -q\mu_q E_x B_z = qE_y \text{ or} \\ F_L &= -\mu_q E_x B_z = E_y \end{aligned} \tag{3.66}$$

The Hall voltage V_H now is simply the field in y-direction multiplied by the dimension w in the y-direction (=wE_y). It is clear then that the (easily measured) Hall voltage is a direct measure of the mobility μ of the carriers involved, and that its sign or polarity will change if the sign of the charges changes.

It is customary to define a Hall coefficient, R_H, for a given material as:

$$R_H \equiv \frac{E_y}{J_x B_z} \tag{3.67}$$

In other words, we expect that the Hall voltage, wE_y, will be proportional to the current density $\mathbf{J}_x$ and the magnetic field strength $\mathbf{B}$, which are, after all, the main experimental parameters (besides the trivial dimensions of the specimen):

$$\mathbf{E}_y = R_H \mathbf{B}_z J_x \tag{3.68}$$

Using Equations 3.10, 3.66, and 3.67, as well as $\mathbf{J}_x = \sigma \mathbf{E}_x$, we calculate R_H as:

$$R_H = -\frac{-\mu_q E_x B_z}{\sigma E_x B_z} = -\frac{\mu}{\sigma} = -\frac{\mu}{qn\mu} = -\frac{1}{qn} \tag{3.69}$$

When we calculate R_H from our measurements and assume $|q| = e$ (which Hall at the time did not know!), we can find n, the charge density. Also, the sign of V_H and thus of R_H tells us the sign of q! If R_H is negative, the predominant carriers are electrons; if positive, they are holes. If both types of carriers are present, the Hall field will change its polarity depending on the majority carrier (Figure 3.20). Also from Equation 3.69 one sees that the lower the carrier density, the higher R_H and thus the higher V_H; this is a key to some very sensitive magnetic field sensors.

For most metals the Hall constant is negative because electrons are majority carriers. But for Be and Zn, for example, there is a band overlap with dominant conduction by holes in the first band and fewer electrons in the second. As a result, R_H is positive for these metals. Of course energy bands were not heard of yet, and the results of a positive R_H were baffling at the time: how can we have $q > 0$ (even for metals!)? The Hall coefficient changes sign with the sign of the charge carrier and therefore provides an important method for investigating the electronic structure of the solid state. In particular, the positive Hall coefficients exhibited by metals such as magnesium and aluminum are a clear indication that a naive picture of a sea of conduction electrons is inappropriate because the majority carriers are clearly positively charged (and are, in fact, holes). The discrepancies between the FEG predictions and experiments nearly vanish when liquid metals are compared (see Table 3.4). This

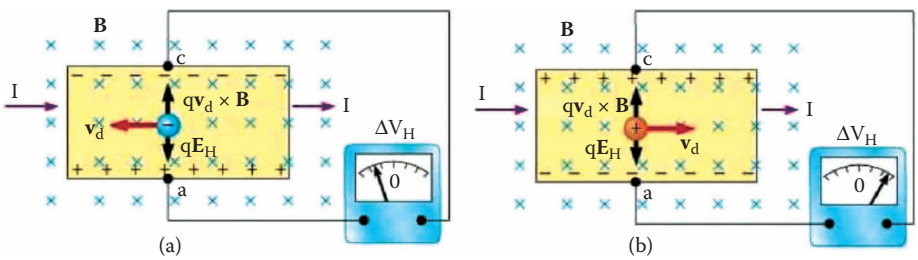


FIGURE 3.20 Electron and hole charge carriers in the Hall effect. When the charge carriers are negative, the upper edge of the conductor becomes negatively charged (a). When the charge carriers are positive, the upper edge becomes positively charged (b).

TABLE 3.4 Discrepancies between the FEG Predictions for R_H and Experiments Nearly Vanish When Liquid Metals Are Compared*

		R_H (10^{-11} m^3/As)		
Metal	n_0	Solid	Liquid	FEG Value
Na	1	–25	–25.5	–25.5
Cu	1	–5.5	–8.25	–8.25
Ag	1	–9.0	–12.0	–12.0
Au	1	–7.2	–11.8	–11.8
Be	2	+24.4	–2.6	–2.53
Zn	2	+3.3	–5	–5.1
Al	2	–3.5	–3.9	–3.9

*This reveals clearly that the source of these discrepancies lies in the electron-lattice interaction. Notice positive R_H for Be and Zn.

reveals clearly that the source of these discrepancies lies in the electron-lattice interaction of a solid.

The Hall effect "oddities" will be properly explained once we have introduced quantum physics. We will also see that under special conditions of extremely low temperature, high magnetic field, and two-dimensional electronic systems (2D electron gas, in which the electrons are confined to move in planes), the Hall resistance R_H is quantized and increases as a series of steps with increasing magnetic field. These discrete energy levels are called Landau levels.

Blackbody Radiation

Planck first discovered the discontinuous behavior that characterizes the atomic world in his analysis of the light spectra emitted from blackbodies in 1900. All substances, with thermal energy, radiate EM waves, and the radiation emanating from solids consists of a continuous spectrum of wavelengths. A blackbody (also cavity radiation) is a hypothetical object that is a perfect absorber or perfect emitter of radiation (Figure 3.21).

"Blackbody" is an unfortunate name as the ideal radiating/absorbing body does not have to be black; stars and planets, to a rough approximation, are blackbodies! When a blackbody is heated, it first feels warm and then glows red or white hot, depending on the temperature. A typical spectrum of the radiated light intensity, brightness, or emittance of a blackbody is shown in Figure 3.22. Intensity is a measure of how much energy is emitted from an object per unit surface area per unit time at a given wavelength

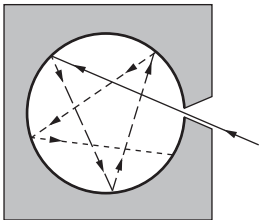


FIGURE 3.21 Ideal blackbody (also cavity radiation). "Blackbody radiation" or "cavity radiation" refers to an object or system that absorbs all radiation incident on it and reradiates energy that is characteristic of this radiating system only, not dependent on the type of radiation that is incident on it. The radiated energy can be considered to be produced by standing wave or resonant modes of the cavity, which is radiating.

and in a particular direction. A blackbody of temperature T emits a continuous spectrum peaking at λ_{max}. At very short and very long wavelengths there is little light intensity, with most energy radiated in some middle range frequencies. As the body gets hotter, the peak of the spectrum shifts toward shorter wavelengths (higher frequencies), but there is always a cutoff at very high frequencies.

It was experimentally observed that the brightness peak position shifted with temperature as:

$$T\lambda_{max} = \text{constant} = 2.898 \times 10^{-3}\ \text{m·K} \quad (3.70)$$

This is known as the Wien displacement law (see Figure 3.22). It was also known that the total energy, E, could be represented as the Stefan-Boltzmann law:

$$E = \sigma T^4 \quad (3.71)$$

where the constant $\sigma = 5.6704 \times 10^{-8}$ W/m²·K⁴. Classical interpretation predicted something

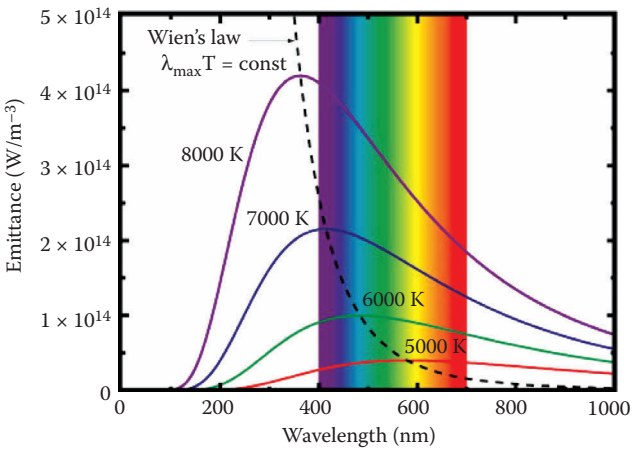


FIGURE 3.22 Blackbody radiation spectra at four different temperatures.

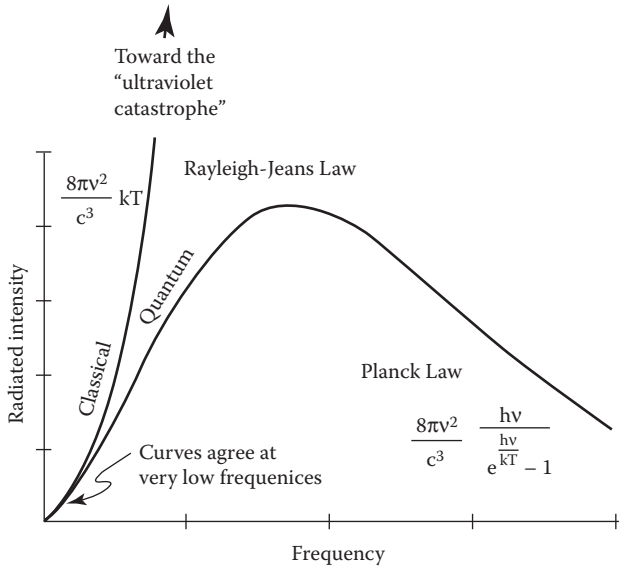


FIGURE 3.23 Planck and Rayleigh-Jeans models for blackbody radiation.

altogether different; in the classical Rayleigh-Jeans model, instead of a peak in the blackbody radiation and a falling away to zero at zero wavelength, the measurements were predicted to go off scale at the short wavelength end as shown in Figure 3.23.

Here is how the British physicists Lord Rayleigh and Jeans derived their model. They interpreted the blackbody radiation coming from a solid as electromagnetic radiation from oscillators that vibrate at every possible wavelength λ. In Figure 3.23 the radiated intensity, $E(\lambda)$, in $Js^{-1}\ m^{-3}$, is the energy distribution, i.e., the energy, E, at each λ. At equilibrium, the mean energy of all oscillators at temperature T is kT. The energy E_λ in a small interval $d\lambda$ is then given by:

$$E_\lambda\, d\lambda = kTdn \tag{3.72}$$

where dn is the fraction of oscillators at the "average energy" in the $d\lambda$ interval. From the Maxwell equations (see Chapter 5) one derives:

$$\frac{dn}{d\lambda} = \frac{8\pi}{\lambda^4} \tag{3.73}$$

Substituting this relation in Equation 3.72 leads to:

$$E_\lambda d\lambda = kT\left[\frac{8\pi}{\lambda^4}\right]d\lambda$$

or also since $\nu = c/\lambda$ (3.74)

$$E_\lambda d\lambda = kT\left[\frac{8\pi\nu^2}{c^3}\right]d\nu$$

The term in brackets $\left[\frac{8\pi\nu^2}{c^3}\right]$ is the number of modes per unit frequency per unit volume.

The amount of radiation emitted in a given frequency range should be proportional to the number of modes in that range. The best of classical physics suggested that all modes had an equal chance of being produced, and that the number of modes increased proportionally to the square of the frequency. Equation 3.74 works at long wavelengths (see Figure 3.23) but fails at short wavelengths. This failure at short wavelengths is called the ultraviolet catastrophe; as λ decreases, E_λ goes to $+\infty$. Even at very low T, the exponential curves in Figure 3.23 would have a huge value for visible light. Objects would be visible in the dark. The predicted continual increase in radiated energy with frequency (dubbed the "ultraviolet catastrophe") did not happen. Nature knew better.

The UV catastrophe attracted the attention of many physicists at the end of the nineteenth century, including Max Planck. Planck took the revolutionary step that led to quantum mechanics. He concluded that the available amount of energy could only be divided into a finite number of pieces among the atom oscillators in the cavity walls, and the energy of such a piece of radiation must be related to its frequency according to a new extremely important equation:

$$E = h\nu \tag{3.75}$$

where h was a new constant now called the Planck constant. The quantization implies that a photon of blue light of given frequency or wavelength will always have the same size quantum of energy. For example, a photon of blue light of wavelength 450 nm will always have 2.76 eV of energy. It occurs in quantized chunks of 2.76 eV, and you cannot have half a photon of blue light—it always occurs in precisely the same-sized energy chunks. Planck showed that the intensity I of radiation from a blackbody could be described by the function (now known as the Planck function):

$$I(\lambda,T) = \frac{2hc^2}{\lambda^5}\frac{1}{\left(e^{\frac{1}{\lambda kT}}-1\right)} \tag{3.76}$$

This Planck function was initially found empirically by trial and error but later derived by assuming the Planck constant and a Boltzmann distribution. Differentiating the Planck function with respect to wavelength one derives Wien's displacement law—the wavelength of the maximum:

$$\lambda_{max} = \frac{hc}{4.965kT} \qquad (3.77)$$

The constant hc/4.965kT agrees with the 2.898×10^{-3} m·K experimental value from Equation 3.70. Integration of the Planck function with respect to wavelength over all possible directions results in the total energy emitted per unit area per unit time from the surface of a blackbody (or absorbed per unit area per unit time by a body in a blackbody radiation field). This gives us back the Stefan-Boltzmann law (Equation 3.71):

$$E = \int_0^\infty E_\lambda(T) = \frac{2\pi^5 k^4}{15c^2h^3} T^4 = \sigma T^4 \qquad (3.78)$$

which also enables us to verify that the constant σ is indeed 5.6704×10^{-8} W/m²·K⁴.

Planck, interestingly, never appreciated how far removed from classical physics his work really was. He spent most of his life trying to reconcile the new ideas with classical thermodynamics. The Planck constant *h* was a bit like an "uninvited guest" at a dinner table; no one was comfortable with this new guest. But today we know that discontinuities in the nanoworld are meted out in units based on this Planck constant. This constant and its particular magnitude constitute one of the great mysteries in nature. It is the underlying reason for the perceived weirdness of the nanoworld, the existence of a "least thing that can happen" quantity—a quantum. The ubiquitous occurrence of discontinuities in the nanoworld constantly upsets our commonsense understanding of the apparent continuity of the macroscopic world.

Light as a Stream of Particles

Photoelectric Effect

The photoelectric effect, discovered by chance, by Heinrich Hertz and his student Hallwachs in 1888,

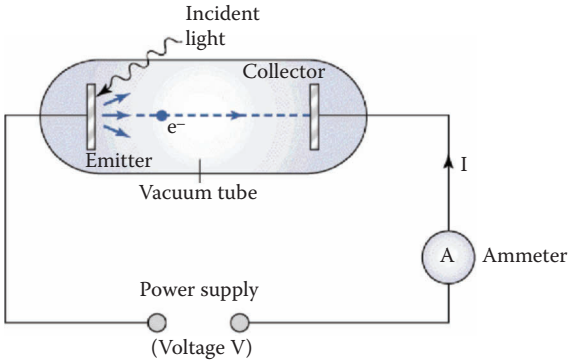


FIGURE 3.24 Experimental setup for studying the photoelectric effect.

is a process whereby light falling on negatively charged Zn, in an evacuated vessel, knocks electrons out of the surface as illustrated in Figure 3.24. The details of the photoelectric effect were in direct contradiction to the expectations of very well-developed classical physics of light waves, and the correct explanation by Einstein in 1905 marked one of the major steps toward quantum theory. Classical light wave theory predicts that the electrons in the metal will absorb radiation energy continuously. Once an electron has absorbed energy in excess of its binding energy, it will be ejected, and one would expect that a higher intensity would increase the chance that electrons are ejected.

As illustrated in Figure 3.24, to test this model we introduce a metal collector plate that collects electrons and a circuit that may be closed to measure the current *I*. In this setup, a retarding or stopping potential, V_0, may be applied to determine the kinetic energy of the electrons ($eV_0 = 1/2\ mv^2$). The classical wave model implies that the stopping voltage V_0 must be proportional to the intensity. Increasing frequency ν should not matter much, perhaps only causing a small decrease in current *I*, as a result of the rapid wave oscillations at high frequencies. With low intensity light, there should be a time delay to build up enough energy before current starts to flow.

The results were unexpected; no electrons were ejected, regardless of the intensity of the light, unless the frequency exceeded a certain threshold characteristic of the bombarded metal (red light did not cause the ejection of electrons, no matter what the intensity). The electrons were emitted

immediately—no time lag. Increasing the intensity of the light increased the number of photoelectrons but not their maximum kinetic energy (a weak violet light would eject only a few electrons, but their maximum kinetic energies were greater than those for intense light of longer wavelengths). The kinetic energy of the ejected electrons varied linearly with the frequency of the incident radiation but was independent of the intensity, or:

$$eV_0(\text{ = kinetic energy of the electrons}) = h\nu - \Phi \tag{3.79}$$

where Φ, the photoelectric work function, is the energy lost by a surface when an electron is freeing itself from its environment. This cannot be explained by the Maxwell equations.

If the charge of the electron is known, a plot of retarding or stopping voltage versus frequency of incident light, shown for three different metals in Figure 3.25, may yield a value for the Planck constant *h*. The electron charge was determined by Robert Millikan in 1909, and with that value and the slope of the lines in Figure 3.25, a value for *h* of 6.626×10^{-34} J·s was calculated, identical to the one derived from the hydrogen atom spectrum and blackbody radiation (see above). The intercept with the frequency axis (at kinetic energy zero) is the threshold frequency, ν_0, and the stop or retarding voltage axis intercept is the binding energy ($-\Phi$) of the electron.

The photoelectric phenomenon could not be understood without the concept of a light particle, i.e., a quantum amount of light energy for a particular frequency. Einstein's paper explaining the photoelectric effect was one of the earliest applications of quantum theory and a major step in its establishment. The remarkable fact that the ejection energy was independent of the total energy of illumination showed that the interaction must be like that of a particle that gave all of its energy to the electron! This fit in well with Planck's hypothesis that light in the blackbody radiation experiment could exist only in discrete bundles with energy. In quantum theory, the frequency, ν, of the light determines the energy, *E*, of the photons and $E = h\nu$, where *h* is Planck's constant ($h = 6.626069 \times 10^{-34}$ J·s) (Figure 3.26).

This assumption explains quantitatively all the observations associated with the photoelectric effects. A photon hits an electron and is absorbed. The energy of the emitted electron is given by the energy of the photon minus the energy needed to release the electron from the surface. This explains the observance of a threshold value below which no electrons are emitted. Thus, it depends on the frequency of light falling on the surface but not on its intensity. It also explains why there is no time lag; a photon hits an electron, is absorbed by the electron, and the electron leaves. Higher intensity light contains more photons, and so it will knock out more electrons. However, if the frequency of the light is such that a single photon is not energetic enough to release an electron from the surface, then none will be ejected no matter how intense the light. Gilbert N. Lewis in 1926 called Einstein's light particles photons. Just as the word *photon* highlights the particle aspect of an electron, the word *graviton* emphasizes

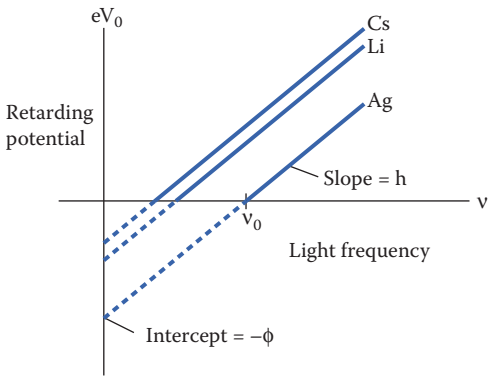


FIGURE 3.25 A plot of retarding or stopping voltage versus frequency of incident light. Slope is the Planck constant *h*. The intercept with the frequency axis (at kinetic energy zero) is the threshold frequency, ν_0, and the stop voltage axis intercept is the binding energy.

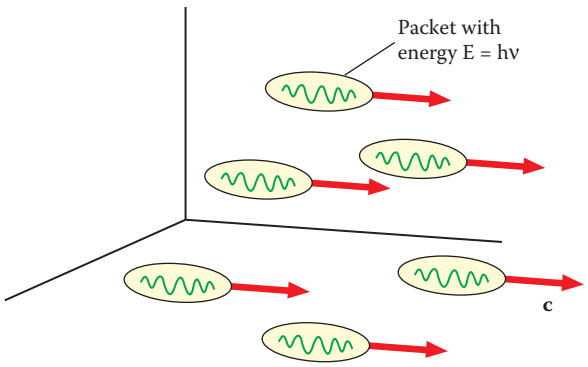


FIGURE 3.26 Light as energy packets with an energy $E = h\nu$.

the particle aspect of gravity, and *gluon* that of the strong nuclear force.

The young Einstein, in 1905, was the first scientist to interpret Planck's work as more than a mathematical trick and took the quantization of light ($E = h\nu$) for physical reality. He gave the uninvited dinner guest—the Planck constant h—a place at the quantum mechanics dinner table. What Einstein proposed here was much more audacious than the mathematical derivations by Planck to explain away the UV catastrophe. For a long time the science of optics had hesitated between Newton's corpuscular hypothesis and Huygens' wave theory. By the beginning of the nineteenth century the wave theory had become the dominant theory, largely because of the persistence of Augustin Fresnel (1788–1827), who described diffraction mathematically, and James Clerk Maxwell (1831–1879), who introduced the famous Maxwell equations in 1864 (Chapter 5). No wonder Einstein wrote at that time to one of his friends: "I have just published a paper about light, but I am sure nobody will understand it." Einstein reintroduced a modified form of the old corpuscular theory of light, which had been supported by Newton but which was long abandoned.

The particle nature of light was hard to swallow at the time, and it indeed still is. Remember the diffraction of x-rays on a crystal described in Chapter 2. Diffraction is something that happens with waves, not with particles. Einstein's light particles also negated the issue of "ether," a medium required for wave propagation as in the case of sound waves (sound waves cannot propagate in a vacuum); light is perfectly happy traveling in a vacuum.

How can we reconcile this duality of a photon as both wave- and particle-like in nature? We will have to learn to think of the wave as the probability of finding the particle. For example, if we know the momentum of the particle exactly, we cannot say where it is, only where it is likely to be [the Heisenberg uncertainty principle (HUP)].

Compton Scattering

The first strong support for the quantum nature of light came from monochromatic x-ray scattering on a graphite block. In 1922, Arthur Compton

FIGURE 3.27 Arthur Harry Compton (1892–1962).

(Figure 3.27), at Washington University in St. Louis, saw that the wavelength of x-rays increases on scatterings off graphite, depending on the angle (Figure 3.28). This effect cannot be explained using wave theory of x-rays.

As first explained by Compton in 1923, a photon can lose part of its energy and momentum on scatterings with electrons in the graphite, and the resulting energy loss (or change in wavelength, $\Delta\lambda$) can be calculated from the scattering angle θ. The result is that some of the scattered radiation has a smaller frequency (longer wavelength) than the incident

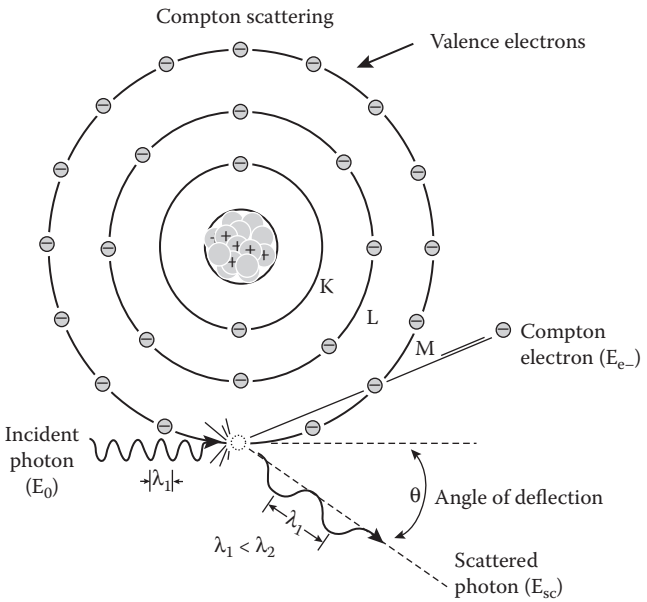


FIGURE 3.28 Compton effect. An incoming photon (E_0) can inelastically scatter from an electron and lose energy, resulting in an outgoing scattered photon (E_{sc}) with lower energy ($E_{sc} < E_0$).

radiation. This change in wavelength depends on the angle through which the radiation is scattered, and Compton concluded that an x-ray photon—or light in general—possesses momentum and thus behaves as a particle. Obviously, this was a big boost for the theory of light as composed of quanta. The three experiments that made the quantum revolution—blackbody radiation, the photoelectric effect, and the Compton effect—all indicate that light consists of particles.

Quantum Mechanics to the Rescue

Introduction

Invoking quantum mechanics we can solve the many problems with classical theories we exposed in the previous sections. Central to quantum mechanics is Schrödinger's equation, but before introducing Schrödinger's equation, we must put down some more foundations. We start with Kelvin's and Thomson's plum pudding atom model, then review Rutherford's and Bohr's improved orbital atom models, emphasize the importance of the Balmer's hydrogen emission lines in the discovery of the inner structure of an atom, and then we get baffled by de Broglie's matter waves and Heisenberg's uncertainty principle (HUP). There are four principal representations of quantum mechanics: Dirac's Hamiltonian and quantum algebra representation; the matrix representation of Born, Heisenberg, and Jordan; Schrödinger's wave equations; and Feynman's sum-over-histories approach; the latter constitutes a fundamentally new way of looking at quantum theory. With the Schrödinger formalism of quantum mechanics we tackle the band model and revisit electrical and thermal conductivity and heat capacity.

Bohr's and Rutherford's Atom

The Greek philosopher Leucippus of Miletus, who lived around 400 BC, first proposed atomic theory of matter. His disciple, Democritus of Abdera, concluded that infinite divisibility of a substance belongs only in the imaginary world of mathematics and further developed his mentor's atomic theory (Figure 3.29).

FIGURE 3.29 Democritus.

Democritus suggested that all things are "composed of minute, invisible indestructible particles of pure matter." According to the ancient Greeks, "atomos"* or atoms were all made of the same basic material, but atoms of different elements had different sizes and shapes. The sizes, shapes, and arrangements of a material's atoms determined the material's properties. For example, sour-tasting substances were believed to contain atoms with jagged edges, whereas round atoms made substances oily. It was further believed that there were four elements that all things were made of: earth, air, fire, and water. This basic theory remained unchanged until the nineteenth century, when it first became possible to test the theory with more sophisticated experiments. Lord Kelvin and J.J. Thomson developed a "raisin cake" model of the atom (also called the plum pudding model, 1897), which incorporated Thomson's negatively charged electrons as the raisins in a positively charged cake (Figure 3.30).

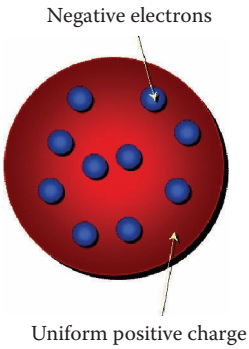


FIGURE 3.30 Kelvin's and Thomson's raisin cake.

* Atomos in Greek means "unbreakable."

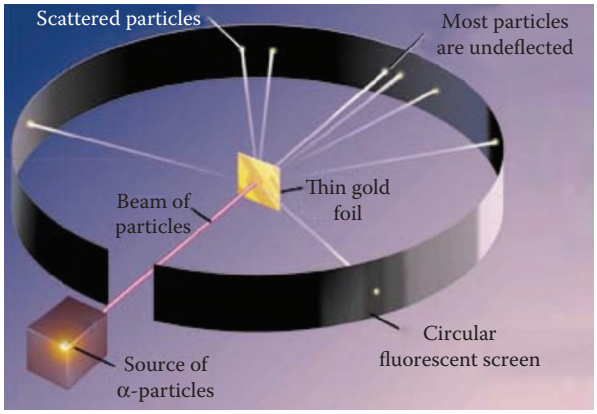


FIGURE 3.31 The Rutherford experiment. Structure of atom experiments.

To check the Kelvin and Thomson model, Geiger and Marsden, working in Rutherford's lab in 1911, directed a narrow beam of alpha particles, double-ionized helium atoms, of known energy onto a thin gold foil. A ZnS scintillation screen was used to record the striking alpha particles. Most particles went through the gold film undeflected, and some were deflected at small angles; however, they found that, once in a while, for about 1% of particles, the α-particles were scattered backward by the target (Figure 3.31).

For a moving alpha particle to be scattered through a large angle, a considerable repulsive force must be exerted. To explain the backscattering, Rutherford proposed that the positive charge in a gold atom must be concentrated in a small region. Rutherford proposed in this way the first realistic model of the atom: he concentrated 99.99% of the mass of the atom in the nucleus, which is only 10^{-15} m across (Figure 3.32). In other words, the atom (and therefore matter in general) is composed of 99.9999999999999% empty space. The proportion of nucleus to total atom size is obviously not drawn to scale in Figure 3.32.

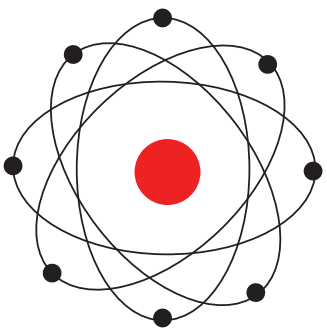


FIGURE 3.32 Rutherford's atom. Matter is mostly empty space.

The problem is that according to classical models the Rutherford model cannot lead to stable atoms. Rutherford's electrons are undergoing a centripetal acceleration and should radiate electromagnetic waves of the same frequency, so-called *bremsstrahlung* or "braking" radiation, leading to an electron "falling on a nucleus" in about 10^{-12} s! In the real world we have stable atoms, and atoms emit certain discrete characteristic frequencies of electromagnetic radiation. The Rutherford model is unable to explain these phenomena.

It was the analysis of light emitted or absorbed by atoms and molecules that led the way to better and better atom models. Cold, dilute gases absorb light at characteristic and discrete wavelengths (absorption spectra), and hot gases emit light at discrete wavelengths while continuous light spectra result when hot solids, liquids, very dense gases, or blackbodies (see Figure 3.33) emit light at all wavelengths (emission spectra).

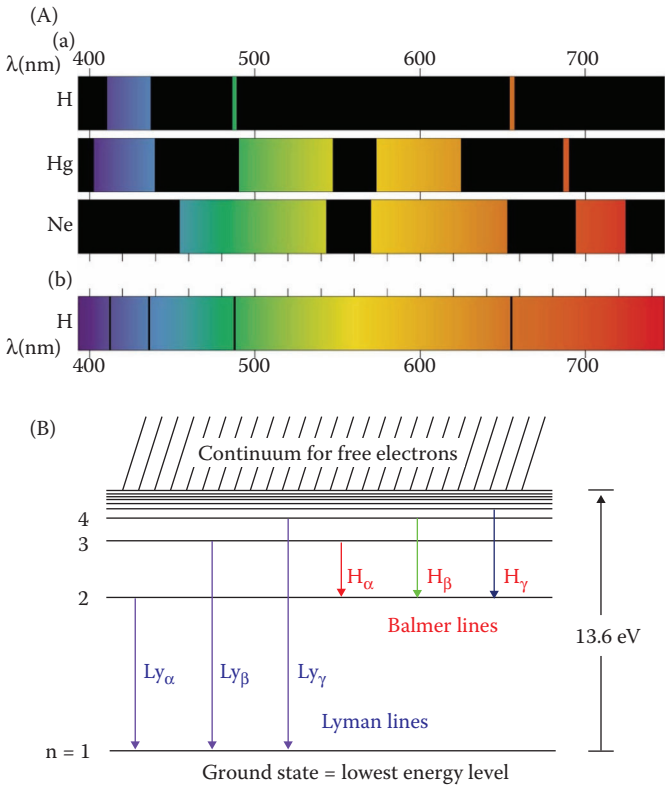


FIGURE 3.33 (Aa) Emission lines for H, Hg, and Ne; (Ab) Spectrum of sunlight with the hydrogen lines adsorbed out. (B) The hydrogen atom emission spectra with Balmer and Lyman lines for hydrogen.

When the light emitted from a hot gas is analyzed with a spectrometer, a series of discrete bright lines is observed. Each line has a different wavelength and color. This series of lines is the emission spectrum (Figure 3.33A). The absorption spectrum consists of a series of dark lines superimposed on the otherwise continuous spectrum. The dark lines of the absorption spectrum coincide with the bright lines of the emission spectrum. The continuous spectrum emitted by the sun passes through the cooler gases of the sun's atmosphere. The various absorption lines can be used to identify elements in the solar atmosphere, and this led, for example, to the discovery of helium.

Atomic hydrogen (one electron and one proton) in a gas discharge tube emits strongly at visible wavelengths H_α, H_β, and H_γ. More lines are found in the ultraviolet region, and the lines get closer and closer until a limit is reached (Figure 3.33B). These same lines are also seen in stellar spectra from absorption in the outer layers of the stellar gas.

There are four bright lines in the hydrogen emission spectrum, and in 1885, a Swiss teacher, Johann Balmer (1825–1898), guessed the following formula for the wavelength of these four lines:

$$\lambda = 364.5 \times 10^{-7} \frac{n^2}{n^2 - 4} (m) \qquad (3.80)$$

where n = 3, 4, 5, and 6, which are now called the *Balmer series*. Equation 3.80 may be rewritten as the so-called Balmer formula as:

$$\frac{1}{\lambda} = R_H \left(\frac{1}{n^2} - \frac{1}{k^2} \right) \qquad (3.81)$$

with positive n and k integers and $n > k$ and R_H the Rydberg constant for hydrogen with a measured value of $1.096776 \times 10^7\ m^{-1}$. With n = 1, we obtain the Lyman series in the UV; visible light is emitted in the Balmer series (n = 2), and with n ≥ 3 the infrared series are obtained (Paschen with n = 3, Brackett with n = 4, and Pfund with n = 5). The discrete emissions in Figure 3.33B suggest discrete energy levels, and Balmer's formula suggests that the allowed energies are given by (R_H/n^2) (in case of hydrogen) (Figure 3.34).

The "Great Dane," Niels Bohr, explained the above results for the H emission spectra by introducing four quantum postulates in a model halfway

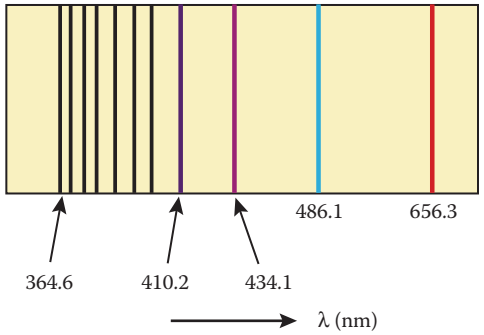


FIGURE 3.34 Some of the hydrogen emission lines.

between classical physics and quantum theory. He reasoned that if electrons orbit the nucleus in circles with radii restricted to certain values, then the energy also can only take on certain discrete values, i.e., if it is quantized and there is a lowest energy orbit, and the electron is not allowed to fall to a lower energy. The allowed states are called *stationary states*. When in these permitted orbits, contrary to classical theory, the electrons do not radiate (Postulate 1). Bohr also assumed that classical mechanics applies to electrons in those stationary states (Postulate 2). He recognized that there might be a link between stable orbits and the Planck's and Einstein's quantum relation between the quantized energy of a photon and its frequency, so he proposed that radiation absorption or emission corresponds to electrons moving from one stable orbit to another, i.e., $\Delta E = h\nu$ (where h is the Planck constant = 6.6×10^{-34} J·s) (Postulate 3). In classical physics, angular momentum (L) of an object in circular motion is defined as mass (m_e) multiplied by velocity (v), multiplied by the radius (r) of the orbit, i.e., $L = m_e v r$ (or $L = pr$) (Figure 3.35). Bohr argued that allowed orbits are determined by the quantization of that angular momentum, $L = nh/2\pi$.

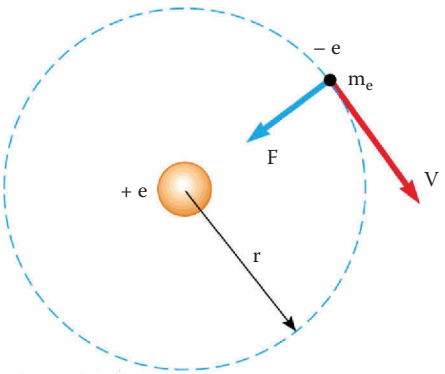


FIGURE 3.35 Momentum **p** for circular motion is $m_e vr$.

In this expression n is called the principal quantum number, and the ground state corresponds to n = 1. So here Bohr postulated that it was not the atom that determined the Planck constant h, but h that determined the properties of atoms (Postulate 4)!

Although this planetary kind of model has been shown to be mostly wrong, it makes for a very nice transition to full-fledged quantum mechanics. Bohr's theory, with slight modifications, is used, for example, to estimate the binding energies of dopant atoms in Chapter 4 and to explain the energy of excitons in Chapter 5.

Mathematically we can express Postulate 4 in vector notation as $|\mathbf{L}| = |\mathbf{p} \times \mathbf{r}| = m_e vr$, with v the tangential velocity and $\mathbf{p}$ the momentum of the electron (Figure 3.35). Bohr now combined his quantum model of the atom with Newton's classical model of planetary orbits to calculate the radius of the hydrogen atom. The attractive force between the electron and the proton in a hydrogen atom is caused by the Coulomb force and is given as:

$$F = \frac{e^2}{4\pi\varepsilon_0 r^2} \tag{3.82}$$

Based on Newton (Postulate 2), we may write:

$$F = \frac{m_e v^2}{r} = \frac{e^2}{4\pi\varepsilon_0 r^2} \tag{3.83}$$

Rutherford could not explain the spectral lines in and Figures 3.33 and 3.34, but he did set up the energy balance for a H atom correctly:

$$E_n \text{ (total energy)} = \text{KE (kinetic energy)} + \text{PE (potential energy)} \tag{3.84}$$

where the KE term depends on the velocity v and the PE term on the system (e.g., positional or electrostatic), or in the current case:

$$E = \frac{m_e v^2}{2} - \frac{e^2}{4\pi\varepsilon_0 r} \tag{3.85}$$

The radial acceleration is $a = v^2/r$, and with Newton's Law (F = ma) this yields $F = e^2/(4\pi\varepsilon_0 r^2) = mv^2/r$, which, solving for r, results in:

$$r = e^2/4\pi\varepsilon_0 m_e v^2 \tag{3.86}$$

And substituting Equation 3.86 in Equation 3.85 results in:

$$E = \frac{m_e v^2}{2} - m_e v^2 = -\frac{m_e v^2}{2} < 0 \tag{3.87}$$

meaning that, because E < 0, the motion of the electron is not free: it is bound by the attractive force of the nucleus as illustrated in Figure 3.36. To free the electron (and ionize the atom), the electron must receive an amount of energy, called

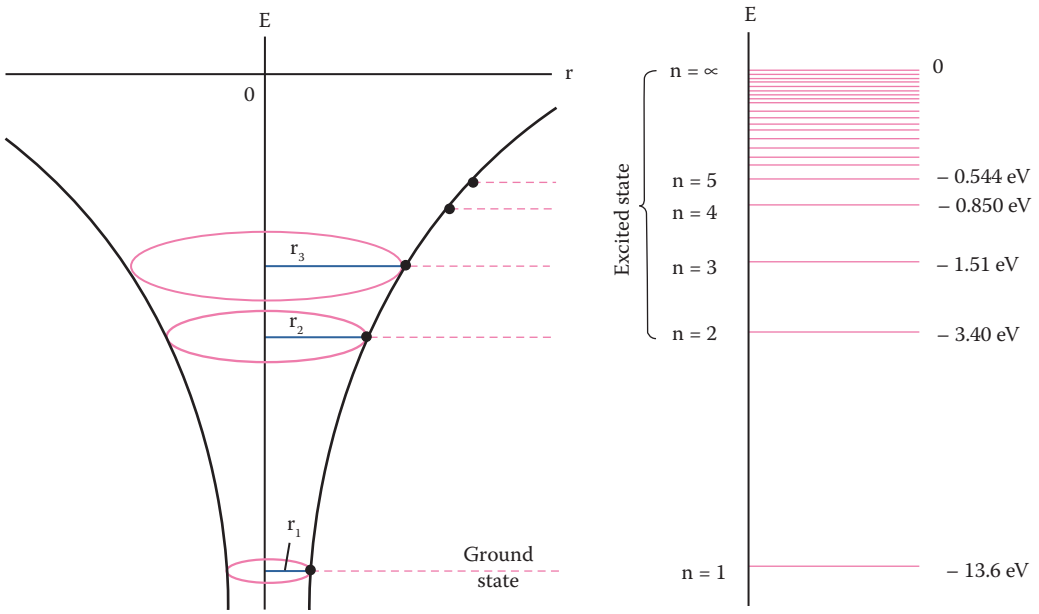


FIGURE 3.36 The electron is bound by the attractive force of the nucleus $E < 0$.

ionization energy, that will bring its total energy up to zero. Although this concept of a negative energy is somewhat counterintuitive, it does make sense; if we say that an electron has zero energy when far removed from the nucleus, then the electrons that are attached to an atom have a negative amount of energy.

The Quantum Numbers

First or Principal Quantum Number n for the Level of Energy E_n

Now we recall Bohr's fourth postulate about the quantization of angular momentum, $L_n = m_e v r_n = \frac{nh}{2\pi}$ or also $m_e v r_n = n\hbar$ with $\hbar = h/2\pi$. In the latter expression, n = 1, 2, 3, 4… and is called the principal quantum number; the ground state corresponds to n = 1. Thus, L_n is not only conserved but also constrained to discrete values by the quantum number *n*. This quantization of angular momentum is a crucial result and can be used in determining the Bohr orbit radii and Bohr energies. From the expression for *r* (Equation 3.86), we derive:

$$r_n = \frac{e^2}{4\pi\varepsilon_0 m_e (n\hbar/m_e r_n)^2} = \frac{e^2 r_n^2 m_e}{4\pi\varepsilon_0 n^2\hbar^2} \qquad (3.88)$$

Solving this expression for r_n results in:

$$r_n = \frac{n^2\hbar^2 4\pi\varepsilon_0}{m_e e^2} = n^2 a_0 \qquad (3.89)$$

where a_0 is the Bohr radius, and permitted orbits have radii $r_1 = a_0$, $r_2 = 4a_0$, $r_3 = 9a_0$ … for n = 1, 2, 3, … as shown in Figure 3.37.

The Bohr radius, $a_0 = \frac{\hbar^2 4\pi\varepsilon_0}{m_e e^2}$, has a value of 0.53×10^{-10} m.

We are now also in a position to calculate Bohr orbit speeds v_n:

$$v_n = \frac{n\hbar}{m r_n} = \frac{\hbar}{n m a_0} = \frac{e^2}{4\pi n\hbar\varepsilon_0} \qquad (3.90)$$

With the latter two expressions, we are able to calculate the total energy of the electron, E_n, associated

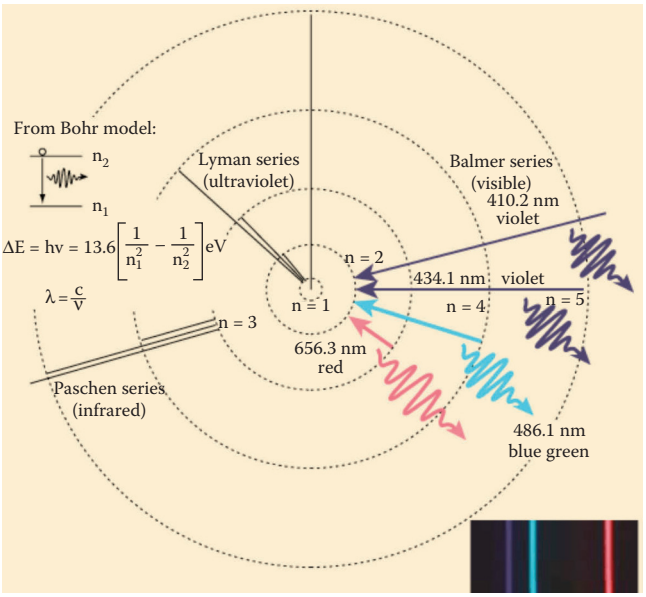


FIGURE 3.37 The Bohr atom with permitted radii. This Bohr model picture of the orbits has some usefulness for visualization, as long as it is realized that the "orbits" and the "orbit radius" just represent the most probable values of a considerable range of values.

with any integer value for *n*; by substituting the last two expressions in Equation 3.85, we obtain:

$$\begin{aligned} E_n &= \frac{m_e v^2}{2} - \frac{e^2}{4\pi\varepsilon_0 r_n} \\ &= \frac{1}{2} m_e \left[\frac{e^2}{n\hbar 4\pi\varepsilon_0}\right]^2 - \frac{e^2}{4\pi\varepsilon_0 n^2 a_0} \\ E_n &= \frac{m_e e^4}{32\hbar^2\pi^2\varepsilon_0^2 n^2} - \frac{m_e e^4}{16\hbar^2\pi^2\varepsilon_0^2 n^2} \\ &= -\frac{m_e e^4}{32\hbar^2\pi^2\varepsilon_0^2 n^2} = -\frac{E_0}{n^2} \end{aligned} \qquad (3.91)$$

and $E_0 = \frac{e^2}{8\pi\varepsilon_0 a_0} = \frac{me^4}{2(\hbar 4\pi\varepsilon_0)^2}$ with a value of 13.6 eV = 21.8×10^{-19} J, sometimes called the Rydberg energy. The 13.6-eV energy value for an electron in a hydrogen atom (with n = 1) is the energy required to remove an electron from that atom (see also Figure 3.33). The possible energy levels of the hydrogen atom are labeled by the values of the quantum number *n*. The lowest energy level occurs for n = 1; this is the most negative energy level and the ground state. As *n* progressively increases, the energy increases

(becomes less negative) for the excited states of the hydrogen atoms, as is clear from Figure 3.36.

Using the expression $E_n = -\frac{E_0}{n^2}$, we can now invoke Bohr's second postulate, which says that the photon energy is the difference in E_n values, and we derive Balmer's formula:

$$E_n = h\nu = \frac{hc}{\lambda} = -E_0\left[\frac{1}{k^2} - \frac{1}{n^2}\right]$$

$$\frac{1}{\lambda} = \frac{E_0}{hc}\left[\frac{1}{n^2} - \frac{1}{k^2}\right] = \frac{e^2}{8\pi\varepsilon_0 a_0 hc}\left[\frac{1}{n^2} - \frac{1}{k^2}\right] \qquad (3.92)$$

$$= R_H\left[\frac{1}{n^2} - \frac{1}{k^2}\right]$$

It was considered a big success for the Bohr model that it accounted for Balmer and other series and that the calculated R_H agreed almost exactly (to within 0.1%) with the phenomenological value for the Rydberg constant for hydrogen, i.e., $1.096776 \times 10^7\ m^{-1}$. The model also gives an expression for the radius of the atom that increases with n^2, the orbit speed that decreases with $1/n$, and predicts the energy levels of a hydrogen atom that increase with $-1/n^2$, and it can be extended to "hydrogen-like" atoms.

Second Quantum Number l for Orbital Angular Momentum

By 1914, Bohr had combined the quantum models of Planck and Einstein with the experimental work of Rutherford to provide a quantum model of the hydrogen atom, which fully explained the bright line spectra of hydrogen. To explain the spectra of more complicated atoms, other quantum numbers in addition to the principal quantum number n, defined by Bohr, had to be introduced. More detailed analysis of light spectra again led the way. Arnold Sommerfeld's (1868–1951) first major contribution was to extend Bohr's model to quantize all types of motion. He conceived elliptical atomic orbits by analogy to Johannes Kepler's elliptical planetary orbits and described mathematically orbits with different elliptical shapes but the same value of the principal quantum number n. This gave a number of different stationary states, some with slightly smaller and some with slightly larger energies—and hence multiple spectral lines—just as observed. The orbital angular momentum for an atomic electron can be visualized in terms of a vector model where the angular momentum vector is seen as precessing about a direction in space. Whereas the angular momentum vector has the magnitude shown in Figure 3.38, only a maximum of l units of $\hbar$ can be measured along a given direction, where l is the orbital quantum number. One of Sommerfeld's other important contributions was to link quantum theory to relativity. If one calculates the speed of an electron moving around the nucleus of an atom a value of the order of 1000 km/s is found. This is small compared with the velocity of light but large enough to have to use relativistic mechanics (see Chapter 5). Using this approach Sommerfeld was able to work out the fine structure of the hydrogen spectrum—a result that was regarded as a triumph both for relativity and quantum theory.

Third Quantum Number m, the Magnetic Number

In the 1890s, Pieter Zeeman (1865–1943) had noticed that if a magnetic field **B** was applied to a hot gas, the emission lines were further split into yet a finer structure. Sommerfeld, in 1916, showed that this was because of the direction (i.e., the orientation) of the orbiting of the electron with respect to the magnetic field. Sommerfeld was able to account for this orientation effect with the addition of a magnetic quantum number, m, which was an integer (Figure 3.39). Although called a "vector," the

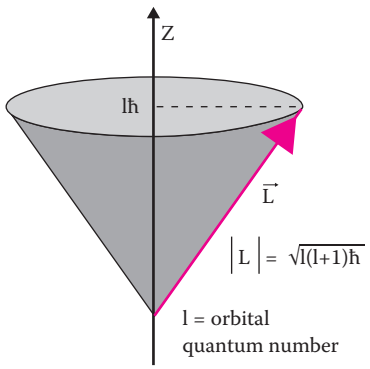


FIGURE 3.38 The Bohr-Sommerfeld atom. Sommerfeld introduced a second quantum number, *l* for elliptical orbitals. Spin-up and spin-down. Pauli's fourth quantum number: the electron spin *s* (see text later this chapter and Figure 3.39).

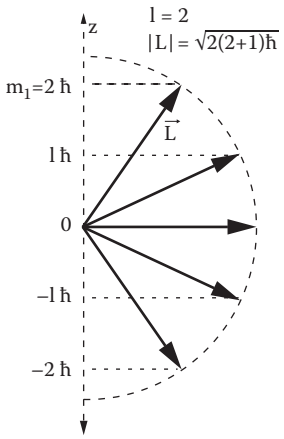


FIGURE 3.39 The quantum number m describes how the direction of the angular momentum is quantized with respect to the direction of the magnetic field.

orbital angular momentum in quantum mechanics is a special kind of vector because its projection along a direction in space is quantized to values one unit of angular momentum ($\hbar$) apart. In Figure 3.39 we show that the possible values that the "magnetic quantum number," m_l (for l = 2), can take are the values m = –2, –1, 0, +1, +2.

Fourth Quantum Number s for Electron Spin

More detailed measurements of the effects of a magnetic field on spectral line splitting showed yet further fine double splitting of the spectral lines. Wolfgang Pauli (1900–1958) explained this a little later, in 1921, by introducing yet another quantum number. While still a student, he proposed a fourth quantum number that he took to represent the electron spinning, not just in its orbit, but around its own axis. Because there are only two directions of spin, clockwise or spin-up and counterclockwise or spin-down, there are only two values for Pauli's electron spin quantum number. No two electrons in an atom can have identical quantum numbers. This is an example of a general principle that applies not only to electrons but also to other particles of half-integer spin (fermions). It does not apply to particles of integer spin (bosons). The two lowest-energy electron shells have an almost identical shape. Of the two, one shell is occupied or "filled" first with an electron that has an intrinsic magnetic direction that is opposite to the intrinsic magnetic field caused by the nucleus. The next shell has an electron with the opposite magnetic direction. The Dutch-American physicists Samuel Goudsmit and George Uhlenbeck discovered the intrinsic "spin" magnetism of the electron in the 1920s. It had been discovered years earlier that in a magnetic field, a beam of electrons splits into two beams (Stern-Gerlach experiment). This fourth quantum number explains this phenomenon. It is believed to be caused by some internal circulation of the electron matter, in addition to its wave flow around the equator of the shell. The wave flow around the equator of the atom also produces atomic orbital magnetic effects. Some shells have no net orbital circulation, which is explained as the result of two equal and opposite counter-rotating orbital waves. The magnetism of the nucleus itself is the result of the fundamental internal spin of the proton.

Surprisingly, it turned out that the fourth quantum number, spin, *s*, had only half of the usual quantization value of $h/2\pi$, i.e., $s = \pm 1/2\ h/2\pi$. Each quantum state, characterized by *n*, *l*, and *m*, is restricted to one electron of s = +1/2 and one electron of s = –1/2. Here classical analogies break down completely—a spin of a half implies the electron has to turn round twice to get back to where it started!

The implications of Pauli's exclusion principle are extremely profound. It is the restrictions imposed by Pauli's exclusion principle—i.e., the fact no two electrons can be in the same quantum state—that prevents all the electrons from piling up into the lowest (n = 1) energy state, and hence stops all matter from collapsing! Pauli's exclusion principle also implies that there is some sort of connectivity between the electron states in an atom: one electron must "know" which states all the other electrons are occupying to choose its own state! It is part of one of our most basic observations of nature: particles of half-integer spin must have antisymmetric wave functions, and particles of integer spin must have symmetric wave functions.

In Table 3.5 the different quantum numbers with their properties are summarized. The quantum numbers associated with the atomic electrons along with Pauli's exclusion principle provide insight into the building up of atomic structures and the periodic properties observed. For a given principal number *n*, there are $2n^2$ different possible states.

Table 3.5 Quantum Numbers and Their Properties

The principal quantum number:	
n	The principal quantum number. Quantization of angular momentum: must be a positive integer (1, 2, 3, 4 … etc.).
The angular momentum quantum number:	
l	Related to the ellipticity of the orbit: must again be an integer but for a particular orbit can be no bigger than n (l = 0, 1, 2, 3 … n – 1).
The magnetic quantum number:	
m	Quantization of the orientation of the orbit with respect to a magnetic field: can be a positive or negative integer (m = –l, –l + 1, 0, 1, 2, … l) but must be no larger than $-l \le m_l \le 1$.
The spin quantum number:	
s	The electron spin quantum number: must be +1/2 or –1/2.

n	Possible l	Possible m	Possible s	Spectroscopic Notation	Total States	Shell/Maximum Number of Electrons
1	0	0	±1/2	1s	2	K or 1st/2
2	0 1	0 –1,0,+1	±1/2 ±1/2	2s 2p	2 6	L or 2nd/8
3	0 1 2	0 –1, 0, +1 –2, –1, 0, +1, +2	±1/2 ±1/2 ±1/2	3s 3p 3d	2 6 10	M or 3rd/10
4	0 1 2 3	0 –1, 0, +1 –2, –1, 0, +1, +2 –3, –2, –1, 0, +1, +2, +3	±1/2 ±1/2 ±1/2 ±1/2	4s 4p 4d 4f	2 6 10 14	N or 4th/32

The Periodic Table of Dmitri Mendeleev (1834–1907)

Most of the qualities of an atom are derived from the structure of its electron cloud with the nucleus in the hinterland. This includes most chemical, material, optical, and electronic properties. Not only did the quantum model of the atom proposed by Bohr, Sommerfeld, and Pauli explain atomic spectra, it also explained the periodic table itself (inside front cover plate)! Dmitri Mendeleev (1834–1907) had devised the periodic table by grouping together elements with similar chemical properties.

The order of filling of electron energy states in an atom is dictated by energy, with the lowest available state consistent with Pauli's principle being the next to be filled. The labeling of the levels follows the scheme of the spectroscopic notation. For each value of principal quantum number, one refers to a different shell: n = 1 for the *K* shell, n = 2 for the *L* shell, n = 3 for the *M* shell, and n = 4 for the *N* shell. Different values of *l* correspond to different subshells. For example, for n = 2 (*K* shell) we have two subshells, namely, 2*s* and 2*p*. For historical reasons, the orbital quantum numbers *l* were given names associated with their appearance in spectroscopic emission and absorption patterns: l = 0 is given the letter *s* (for sharp), l = 1 is given the letter *p* (for principal), l = 2 is given the letter *d* (for diffuse), and l = 3 is given the letter *f* (for fundamental). Therefore, we can write the electronic occupation of an atom's shell as nl^e, where *n* is the principal quantum number, *l* is the appropriate letter, and *e* is the number of electrons in the orbit. So, for example, nitrogen (N) with seven electrons has the configuration: $1s^2 2s^2 2p^3$. In chemistry, this is called the Aufbau (build-up) principle, from the German word for *structure*.

As the periodic table of the elements is built up by adding the necessary electrons to match the atomic number, the electrons will take the lowest energy consistent with Pauli's exclusion principle. The maximum population of each shell is determined by the quantum numbers, and the diagram in Figure 3.40 is a convenient way to illustrate the order of filling of the electron energy states. For a single electron, the energy is determined by the principal quantum number *n*, and that quantum

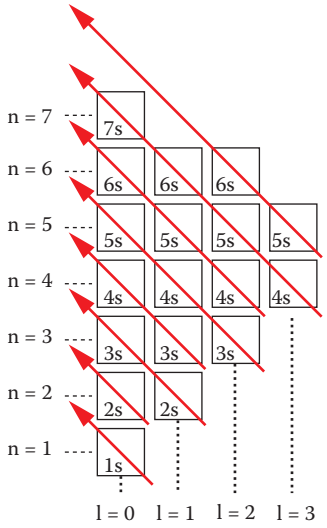


FIGURE 3.40 The periodic table of the elements is built up by adding the necessary electrons to match the atomic number.

number is used to indicate the "shell" in which the electrons reside. For a given shell in multielectron atoms, those electrons with lower orbital quantum number l will be lower in energy because of greater penetration of the shielding cloud of electrons in inner shells. These energy levels are specified by the principal and orbital quantum numbers using the spectroscopic notation. When you reach the 4*s* level, the dependence on orbital quantum number is so large that the 4*s* is lower than the 3*d*. Although there are minor exceptions, the level crossing follows the scheme indicated in the diagram, with the arrows indicating the points at which one moves to the next shell rather than proceeding to a higher orbital quantum number in the same shell.

The quantum scheme gives a firm scientific basis for Mendeleev's grouping: the chemical properties of the elements are defined by how nearly full or nearly empty a shell is. For example, full shells are associated with chemical stability (e.g., helium, neon, argon). Shells with a single electron or with one electron short of a filled shell are associated with chemical activity (e.g., sodium, potassium, chlorine, bromine). Most metals are formed from atoms with partially filled atomic orbitals, e.g., Na and Cu, which have the electronic structure:

Na $1s^2\,2s^2\,2p^6\,3s^1$
Cu $1s^2\,2s^2\,2p^6\,3s^2\,3d^{10}\,4s^1$

In the simplest picture, a metal has core electrons that are bound to the nuclei and valence electrons that can move through the metal. Insulators are formed from atoms with closed (totally filled) shells, e.g., inert gases:

He $1s^2$
Ne $1s^2\,2s^2\,2p^6$

or they form closed shells by covalent bonding as in the case of diamond.

With principal quantum number n = 4, the 4*f* orbitals make their capricious appearance. Because electrons in these orbitals are less strongly held by the nucleus, it is easier to excite them, and they can exhibit a myriad of distinct energy states. This gives rise to the unusual optical and magnetic properties of the rare-earth elements.

Most atoms with odd atomic numbers (1, 3, 5…) have a very slight overall atomic magnetism because of the one electron spin (and some orbital magnetism in some elements), whereas most even atomic number (2, 4, 6…) atoms have no net electron spin magnetism, and thus approximately zero resulting atomic magnetism.

Bohr's Correspondence Principle

Classical physics works for large systems; it is only at the atomic level that it fails. In 1923, Bohr proposed that any satisfactory quantum theory, then still being sought, should be in agreement with classical physics when the energy differences between quantized levels are very small or the quantum numbers are very large. This is the so-called Bohr correspondence principle. Let us consider the two main differences between the quantum theory and classical physics. The first difference is that whereas classical theory deals with continuously varying quantities, quantum theory deals with discontinuous or indivisible processes (e.g., the unit of energy packed in a quantum). The second difference is that whereas classical theory completely determines the relationship between variables at an earlier time and those at a later time, quantum laws determine only probabilities of future events in terms of given conditions in the past. The Bohr correspondence principle states that the laws of quantum physics must be so chosen that in the classical limit, where many quanta

are involved (e.g., n is a large integer in E = nh), the quantum laws lead to the classical equations as an average. This requirement combined with indivisibility and incomplete determinism defines the quantum theory in an almost unique manner. In this chapter we will on a few occasions show how classical theory can be derived from an extension of quantum mechanics. An example of this principle that we may appreciate already is that of Newtonian mechanics as a special case of relativistic mechanics when the speed is much less than the speed of light ($v \ll c$) (see Chapter 5, Equation 5.180). Similarly, Schrödinger's equation agrees with Bohr's ideas in general, except that the electron wave function is not at a given radius but spread out over a range of radii.

Summary of Bohr's Model

Although Bohr's model, in which r (space), energy, and momentum are quantized, was a major step toward understanding the quantum theory of the atom, it is not in fact a correct description of the nature of electron orbits. Some of the shortcomings of the model are:

1. It fails to provide any understanding of why certain spectral lines are brighter than others.
2. There is no mechanism for the calculation of electron transition probabilities.
3. Bohr's model treats the electron as if it were a miniature planet, with definite radius and momentum. This is in direct violation of the uncertainty principle (see below), which dictates that position and momentum cannot be simultaneously determined.

The precise details of spectra and charge distribution must be left to quantum mechanical calculations. Wave functions derived from Schrödinger's equation, the model that corrects all of Bohr's errors, are determined by the values of the above reviewed quantum numbers. For each energy level E_n, there is more than one distinct state with the same energy but different quantum numbers; this is called *degeneracy*. Degeneracy has no counterpart in Bohr's model. The solution of Schrödinger's equation, introduced below, for the case of the hydrogen atom is achieved by using spherical polar coordinates and by separating the variables so that the wave function is represented by the product of three terms. The separation leads to three equations for the three spatial variables, and their solutions give rise to three quantum numbers associated with the hydrogen energy levels. The solution to these equations can exist only when a few constants, which arise in the solution, are restricted to integer values. This gives the same hydrogen atom quantum numbers: n, l, and m as introduced. The fourth quantum number is, of course, again the electron spin s = 1/2, an intrinsic property of electrons.

de Broglie Matter Waves

Matter Waves

To launch Schrödinger's equation, we revisit de Broglie's concept of "matter waves." Louis-Victor, seventh duc de Broglie (1892–1987), discovered that the secret of Planck's and Einstein's quanta could be found in a general law of nature, i.e., the dual character of waves and particles.

Einstein's special relativity* (Chapter 5) allows one to calculate the momentum, **p**, of a photon starting from:

$$E^2 = (\mathbf{p}c)^2 + (m_0c^2)^2 \tag{3.93}$$

The rest mass, m_0, is zero for a photon, so one obtains for its momentum:

$$E^2 = (\mathbf{p}c)^2 + 0 \text{ or } E = \mathbf{p}c \tag{3.94}$$

And also, from Einstein and Planck, E = hν (the photoelectric effect), or:

$$\mathbf{p} = \frac{E}{c} = \frac{h\nu}{c} = \frac{h}{\lambda} \tag{3.95}$$

de Broglie, while studying for his PhD in Paris in 1924, postulated that Equation 3.95 also applied to a moving particle such as an electron, in which case λ is the wavelength of the wave associated with the moving particle, i.e., a "matter wave." Einstein commented about this fantastic insight: "de Broglie has lifted the great veil." From Equation 3.95, $\mathbf{p}\lambda = h$,

* For p = 0 Equation 3.93 leads to the familiar $E = m_0c^2$. This amounts to 511 keV for an electron and 930 GeV for a proton. Where atomic physics deals in eV to keV, nuclear physics deals in keV to MeV and particle physics involves GeV to TeV.

or momentum multiplied by wavelength, gives the Planck constant; the smaller the wavelength, the bigger the momentum of the particle! Thus, electrons, with their small mass and correspondingly small momentum, are very "wavy" particles. The de Broglie wavelength of a particle is then given as:

$$\lambda = \frac{h}{mv} \tag{3.96}$$

where v is its velocity. This simple equation proves to be one of the most useful, and famous, equations in quantum mechanics. It accounts for both waves (characterized by wavelength and frequency) and particles (characterized by position, mass, and velocity), incorporating momentum (particle aspect) and the wavelength (wave aspect).

The critical reader could very legitimately ask here: "But how can photons have a momentum if they do not have any mass?" The answer is that photons do not have a mass, but they do have energy, and as Einstein famously proved, mass and energy are the same thing. For photons, the wavelength and frequency, from Equation 3.95, are related because $\lambda\nu = c$ is fixed, but for matter particles $\lambda\nu$ is not a constant because they do not travel at speed c. For nonzero rest mass particles, such as electrons or protons, neither $\lambda\nu = c$ nor $E = \mathbf{p}c$ applies.

We almost have come full circle here; Einstein gave what we had come to think of as a wave (light) a particle character, and de Broglie gave what we thought of as a particle (electrons) a wave character. Radiation has wave character and particle character, and matter has particle and wave character, or at the nanoscale, nature presents itself with a wave-particle duality. This duality for material things is only significant at the very small scale. With its mass m_e of 9.11×10^{-31} kg and a velocity v of 3×10^6 m/s (i.e., c/100), an electron has a wavelength of 2.5×10^{-10} m. Larger objects, such as a car (1000 kg), moving at highway speeds (60 mph ≈ 100 km/h) have a de Broglie wavelength as well, but there is no measurement capable of resolving them; the oscillations cannot be seen. In other words, the car does not behave like a wave, with a $\lambda = \frac{h}{p} = \frac{h}{mv} = \frac{6.6 \times 10^{-34}\,\text{J·s}}{1000\ \text{kg} \times 10^5\ \text{m/h} \times 1\,\text{h}/36{,}000\,\text{s}} = 2.4 \times 10^{-38}$ m. A man (60 kg) running at 20 km/h (5 m/s) has a wavelength of 2.2×10^{-36} m (for some other matter wave examples see Table 3.6).

Wave Packets

Because we associate a de Broglie wave with a moving body, it is reasonable to expect that this wave moves at the same velocity as that of the particle—this turns out not to be the case. de Broglie hypothesized that the particle itself was not a wave but always had with it a pilot wave, or a wave that helps guide the particle through space and time. He also pointed out that in wave theory there is a difference between the speed of waves of some uniquely defined frequency and the speed of a localized pulse. Waves have both a phase velocity and a group velocity, and de Broglie proposed to associate the localized pulse, wave packet, or group velocity with the moving particle.

To understand *matter* waves and their link with wave packets or group velocity and phase velocity better, we reconsider here for a moment a sinusoidal wave function $\psi(x) = Ae^{ikx} = A\cos kx + iA\sin kx$

Table 3.6 de Broglie Wavelengths of Various Moving Objects

	Mass	Kinetic Energy (eV)	de Broglie Wavelength (m)
Electron	9.11×10^{-31}	1	1.23×10^{-9}
		100	1.23×10^{-10}
		10^4	1.23×10^{-11}
Neutron	1.67×10^{-27}	1.00×10^3	9.05×10^{-13}
		1.00×10^6	2.86×10^{-14}
		1.00×10^9	9.05×10^{-16}
Proton	1.67×10^{-27}	1.00	2.86×10^{-11}
		1.00×10^2	2.86×10^{-12}
		1.00×10^3	9.04×10^{-13}
Thermal neutrons (300 K)		2.50×10^{-2}	1.81×10^{-10}

that extends all the way from $x = -\infty$ to $x = +\infty$ with an amplitude A. To make this wave more localized or "lumpy," we add together two different stationary-state waves with slightly different wave vectors $\mathbf{k}$ (e.g., $\mathbf{k}_1$ and $\mathbf{k}_2$) and different amplitudes A (e.g., A_1 and A_2) and consider the result at one instance of time, e.g., at $t = 0$. For $t = 0$, the time factor of the superposed waves, $e^{-i\omega t}$ (e.g., ω_1 and ω_2), is equal for both superposed waves, i.e., $e^0 = 1$, so the wave function is represented by:

$$\begin{aligned}\psi(x) &= A_1 e^{ik_1 x} + A_2 e^{ik_2 x} \\ &= [A_1 \cos(k_1 x) + iA_1 \sin(k_1 x)] \\ &\quad + [A_2 \cos(k_2 x) + iA_2 \sin(k_2 x)] \\ &= [A_1 \cos(k_1 x) + iA_2 \cos(k_2 x)] \\ &\quad + i[A_1 \sin(k_1 x) + A_2 \sin(k_2 x)]\end{aligned} \tag{3.97}$$

This is represented in Figure 3.41a, where we graph the real parts of the individual waves for the case that $A_2 = -A_1$. The real part of the combined wave function is shown in Figure 3.41b. With two additional sinusoidal waves, we can choose to emphasize every other lump and suppress the in-between ones as illustrated. Going on like this, we can, in the extreme, superpose a very large number of waves with different $\mathbf{k}$ and A values to construct a wave with only one lump, as illustrated in Figure 3.41c. The latter wave pulse is called a *wave packet*. The particle represented by the lump in Figure 3.41c is more likely to be found in the region of the lump than in other regions; in other words, it has become more localized. However, the particle now does have a less definite momentum as we introduced a very large number of different waves. The above procedure of adding waves is not new to us: in Chapter 2 on Fourier transforms (Equations 2.15 and 2.16) we saw how a localized disturbance or pulse can be synthesized from a very large number of pure sinusoidal waves of different frequencies and amplitudes. Such a pulse, a wave function with a lump, is an entity with both particle and wave characters. It behaves like a particle in the sense that it is localized in space, but it also has periodicity, a property of a wave. The Fourier integral representing the synthesized pulse or wave packet is:

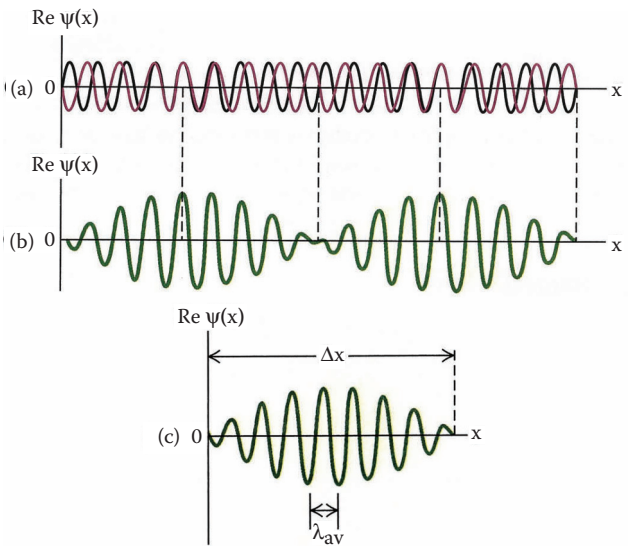


FIGURE 3.41 (a) Two sinusoidal waves with slightly different wave numbers at one instant of time ($A_2 = -A_1$) in Equation 3.97. (b) The superposition of these two waves has a wave number equal to the average of the two individual wave numbers. The amplitude varies, giving the total wave a more lumpy character. (c) Superimposing a large number of sinusoidal waves with different wave numbers and appropriate amplitudes can produce a wave pulse.

$$\psi(x) = \int_{-\infty}^{\infty} A(\mathbf{k}) e^{ikx} d\mathbf{k} \tag{3.98}$$

This Fourier integral superimposes wave functions with a continuous distribution of $\mathbf{k}$ values and amplitudes $A(\mathbf{k})$ that depend on $\mathbf{k}$. Below we will see how Heisenberg's uncertainty principle (HUP) follows directly from the existence of the de Broglie waves and this Fourier integral (see Figure 3.47 further below).

The synthesized pulse travels with a group velocity (V_g) that may be different from the characteristic phase velocity (V_p) of the individual waves. Let us consider the Fourier superposition of just two waves, $\Psi_1 = \sin[kx - \omega t]$ and $\Psi_2 = \sin[(k + \Delta k)x - (\omega + \Delta\omega)t]$, as represented in Figure 3.42a, where we can write:

$$\Psi_1 + \Psi_2 = 2\cos\left[\frac{\Delta\omega}{2}t - \frac{\Delta k}{2}x\right] \times \sin\left[\left(k + \frac{\Delta k}{2}\right)x - \left(\omega + \frac{\Delta\omega}{2}\right)t\right] \tag{3.99}$$

Modulation — Sine wave

The phase velocity is directly proportional to the angular frequency, but inversely proportional to the wave number, or the phase velocity v_p is given as:

$$v_p = \frac{\omega + \frac{\Delta\omega}{2}}{\mathbf{k} + \frac{\Delta\mathbf{k}}{2}} = \frac{\omega'}{\mathbf{k}'} \tag{3.100}$$

from the sine wave in Equation 3.99. For the group velocity v_g and thus the particle velocity:

$$v_g = \frac{\frac{\Delta\omega}{2}}{\frac{\Delta k}{2}} = \frac{\Delta\omega}{\Delta\mathbf{k}} = \frac{d\omega}{dk} \tag{3.101}$$

which is based on the modulation amplitude in Equation 3.99. This group velocity, the derivative of the angular frequency with respect to the wave vector, is the velocity at which the energy of the wave propagates. Group and phase velocities are illustrated in Figure 3.42b.

The group velocity is the derivative of the phase velocity, and it is often the case that the phase velocity is larger than the group velocity. For any wave that is not electromagnetic, the phase velocity will be larger than *c*, or the speed of light.

de Broglie was trying to find a velocity that was fit for all particles (not just photons), so he associated the group velocity v_g with any particle's velocity. He equated the following two expressions:

$$v_g = \frac{\partial\omega}{\partial\mathbf{k}} = \frac{\partial E}{\partial\mathbf{p}} \tag{3.102}$$

because $v_p = E/\mathbf{p}$ and $v_g = \partial E/\partial\mathbf{p}$ and (as we will see below) $E = \hbar\omega$ and $p = \hbar\mathbf{k}$, where $\hbar = h/2\pi$. In the case of light:

$$v_p = \frac{E}{\mathbf{p}} = \frac{mc^2}{mc} = c \text{ and } v_g = \frac{\partial E}{\partial p} = \frac{\partial pc}{\partial p} = c \tag{3.103}$$

We get the same result for phase and group velocity. For any other particle, from Equation 3.93 with $m_0 = (m \neq 0)$, we obtain:

$$v_g = \frac{\partial E}{\partial p} = \frac{pc^2}{\sqrt{(\mathbf{pc})^2 + \left(mc^2\right)^2}} = \frac{pc^2}{E} \tag{3.104}$$

with $m = m_0$, $v_g = c$ again.

de Broglie Vindicated

de Broglie's doctoral dissertation exam committee was distinctly unsure as what to do with the thesis of

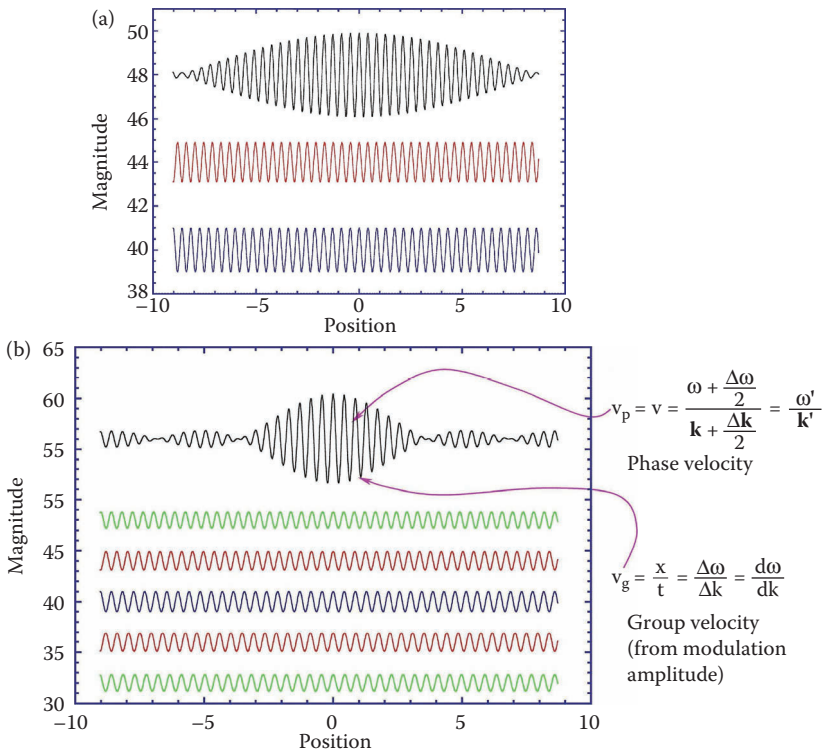


FIGURE 3.42 (a) Using Fourier's theorem to construct wave packets. (b) Phase velocity v_p from the sine wave and group velocity v_g from the modulation amplitude.

the prince. Luckily, Langevin, who was on the committee, asked Einstein's opinion on the work, and, as mentioned above, Einstein was duly impressed. Experimental evidence was soon to follow.

J.J. Thompson received the Nobel Prize in 1906 for his work demonstrating the particle nature of an electron. J.J. Thompson's son, G.P. Thompson, received the Nobel Prize in 1937 for experiments that tested Equation 3.96, demonstrating that electrons also exhibit wave-like properties. In his 1928 tests, G.P. Thompson and Reid at the University of Aberdeen observed interference patterns from electrons reflecting from a thin polycrystalline metal foil surface. Working independently, Clinton Davisson and Lester Germer, at Bell Laboratories, in 1927 found the same experimental evidence: a beam of electrons scattered from a single-crystal of nickel resulted in a diffraction pattern fitting Bragg's diffraction law (Equation 2.20 and Figure 3.43). This established the wave character of electrons, forming the basis of analytical techniques for determining the structures of molecules, solids, and surfaces, such as in low energy electron diffraction (LEED) and SEM described in Volume III, Chapter 6 on metrology. Whereas optical microscopes cannot resolve details smaller than the wavelength of visible light, with an SEM the wavelength depends on the electron's momentum in accordance with de Broglie's relation. In this case we can increase the resolution by simply accelerating the electron more.

Other particles, including protons, neutrons, and He atoms, were subsequently found to possess wave properties, including diffraction. In 1999, researchers from the University of Vienna demonstrated that the wave-particle duality even applied to molecules such as fullerene. One of the coauthors of this research, Julian Voss-Andreae, became an artist and has since created several sculptures symbolizing wave-particle duality in buckminsterfullerenes.[2]

Bohr's Model and de Broglie Matter Waves

de Broglie's idea of matter waves fits nicely with Bohr's model; the de Broglie wavelength associated with the electron is what causes the quantization that Bohr had assumed before matter waves were even known. The angular momentum of the electron is restricted to certain values because an integral number of electron wavelengths must fit into the circumference of the circular orbit. Mathematically we expressed Bohr's Postulate 4, in vector notation, as $|\mathbf{L}| = |\mathbf{p} \times \mathbf{r}| = m_e vr$, with v the tangential velocity and $\mathbf{p}$ the momentum of the electron. Bohr's model, as we saw above, assumed $\mathbf{L} = 2\pi r$, or:

$$\mathbf{L} = 2\pi r = n\lambda = \frac{nh}{\mathbf{p}} \text{ and } r = \frac{n\hbar}{p} (\text{with } \hbar = \frac{h}{2\pi}) \tag{3.105}$$

or

$$L = \mathbf{p}r = \frac{nh}{2\pi} = n\hbar$$

The latter is illustrated in Figure 3.44, where we show that only an integer number of wavelengths fits into a circular orbit or, i.e., $L = nh/2\pi$ (where L

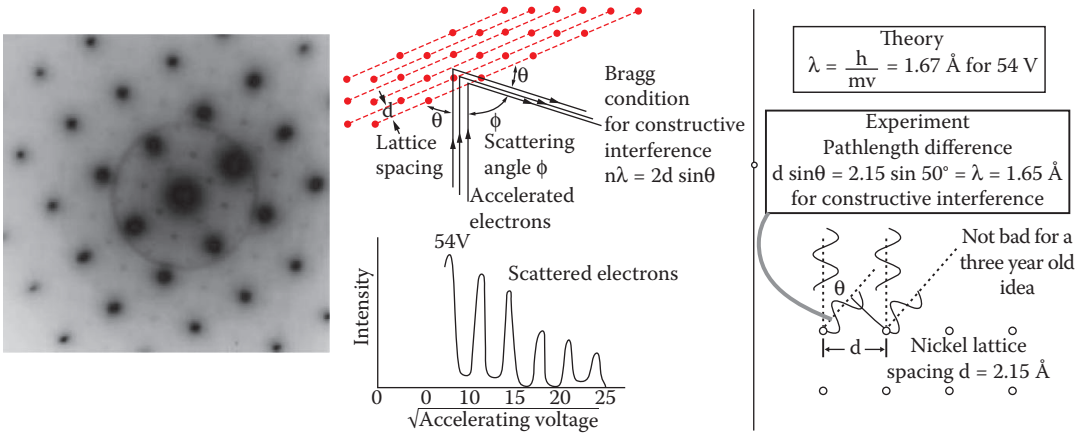


FIGURE 3.43 Davisson-Germer experiment (1927). A graph of the intensity of the scattered electron beam demonstrates that the angular position of the maxima depended on the accelerating voltage used to produce the electron beam. Using the de Broglie equation, the de Broglie wavelength of the electron could be calculated.

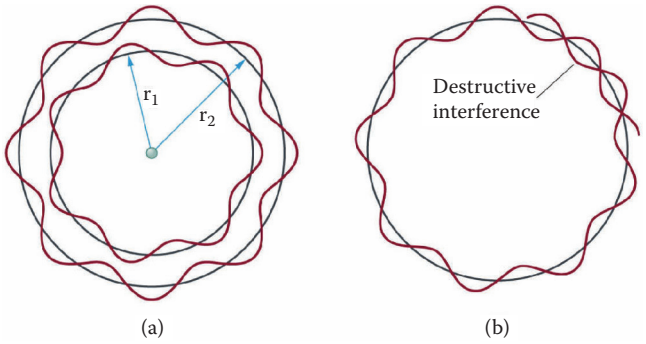


FIGURE 3.44 Bohr's atom model: only an integer number of wavelengths fits into a circular orbit, $L = nh/2\pi$ (where *L* is the length of an orbit and *h* is the Planck constant).

is the length of an orbit). Thus, a wave-mechanical picture leads naturally to the quantization of the angular momentum.

de Broglie's relations imply the quantization of orbits of electrons in atoms, $L = n\hbar$ with *n* an integer, but although this explains the root of quantization—the wave nature of matter—it does not provide a complete dynamic that would be able to describe the motion of particles that are waves at the same time. One needs a wave equation for that, and Maxwell's equations are inappropriate for this purpose because they predict that $\lambda\nu = c$, which is not valid for matter particles, where $\nu\lambda = v$. Interestingly, we will see next that Bohr's model violates Heisenberg's uncertainty principle: an electron may not move in a circle with exact radius *r*!

Heisenberg's Uncertainty Principle

The wave-particle duality introduced in the previous section forced physicists to reconsider their description of the position and momentum of very small particles, and it is at the core of Heisenberg's uncertainty principle (HUP). In Newtonian mechanics, we can always describe a particle in terms of its spatial coordinates and the three components of its velocity. However, in the nanoworld, Heisenberg's principle states that there are physical parameters in quantum physics whose values cannot be known accurately simultaneously. For example, the momentum, $\mathbf{p}_x$, and position, *x*, of an electron cannot be known simultaneously. In classical physics, knowing $\mathbf{p}_x$ would not be difficult because light is considered to have continuously varying energies and would hence cause minimal disturbance on an electron that is being observed. In quantum physics, light consists of photons that have discrete (quantized) energies, and a photon bouncing off an electron being observed gives that electron a kick, disturbing its momentum (Figure 3.45). To obtain an accurate measurement of the position of the electron, one must use a very short probing wavelength because a long wavelength would not have enough resolution to locate the electron. Thus, a short wavelength is expected to give good resolution or a small uncertainty in position. Nevertheless, a short wavelength

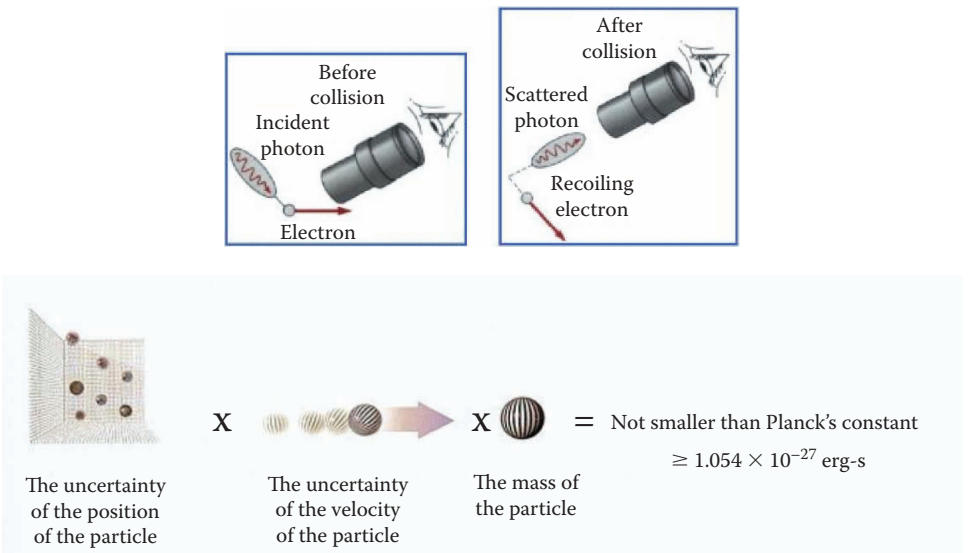


FIGURE 3.45 Looking is disturbing! Imagine trying to see an electron with a powerful optical microscope. At least one photon must scatter off the electron and enter the microscope, but in doing so it will transfer some of its momentum to the electron.

also means a large energy and hence a large momentum ($\mathbf{p}_x = \mathbf{E}/c$), which gives the observed electron too large a kick and too large an uncertainty about its momentum $\Delta\mathbf{p}_x$ (Compton effect).

A more quantitative analysis shows that the product of the two uncertainties, Δx and $\Delta\mathbf{p}_x$, is a constant:

$$\Delta\mathbf{p}_x\Delta x \geq \frac{h}{2\pi} = \hbar \qquad (3.106)$$

where h is again the Planck constant, and the constant "h-bar" has the approximate value of 10^{-34} J·s. There is an uncertainty relation as shown in Equation 3.106 for each coordinate and its corresponding momentum component. Equation 3.106 is graphically represented in Figure 3.46. Note that there is no restriction on the precision in simultaneously knowing/measuring the position along a given direction (x) and the momentum along another, perpendicular direction (y or z).

For a particle in circular motion, the equivalent Heisenberg expression to Equation 3.106 is $\Delta\mathbf{p}_r\Delta\mathbf{r} \geq \hbar$. Applying this to Bohr's model, in which an electron moves in a circle of *exact* radius r, we obtain $\Delta r = 0$ and $\Delta P_r = \infty$, or the model violates the uncertainty principle. We will see further below that despite this obvious problem, the energy level predictions of Bohr's model remain valid.

Similar ideas lead to the expression of uncertainty for other pairs of observables such as time and energy:

$$\Delta E\Delta t \geq \frac{h}{2\pi} = \hbar \qquad (3.107)$$

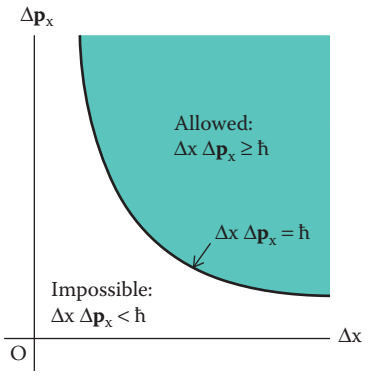


FIGURE 3.46 A graphic presentation of the Heisenberg principle from Equation 3.106 for position and momentum.

This says that if an energy state only lasts for a limited time, its energy will be uncertain. The uncertainty about the energy of a particle depends on the time interval Δt that the system remains in a given energy state. Importantly, this also means that conservation of energy can be violated if the time is short enough. From the uncertainty principles, it is possible that empty space locally does not have zero energy but may have sufficient ΔE for a very short time Δt to create particles and their antiparticles. This can be demonstrated through the Casimir effect (see below). Equation 3.107 is responsible for "lifetime broadening" of spectral lines. Short-lived excited states (small Δt) possess large uncertainty in the energy of the state (large ΔE). As a consequence, shorter laser pulses (e.g., femto- and attosecond lasers) have broader energy (therefore, wavelength) bandwidths.

The existence of a zero-point energy—vibrational energy cannot be zero even at T = 0 K (see below)—is also a consequence of Heisenberg's uncertainty principle. If the vibration would cease at T = 0 K, then the position and momentum would both be 0, violating the HUP.

One might object that perhaps light was a poor choice to measure the position and momentum of the electron and that some other method might avoid these uncertainties. No such luck: it turns out that this is the absolute best that can be achieved independently of the measuring method. Note that in the last two equations all references to light have dropped out; the result does not depend on λ, n, or c. The Heisenberg uncertainty principle follows solely from the wave-particle duality and has nothing to do with the unavoidable disturbance of the system by the measurement. Quantum mechanics tells us there are limits to measurement—not because of the limits of our instruments, but inherently.

Actually, the uncertainty principle is an inevitable consequence of de Broglie's relation and the Fourier integral representing the synthesized pulse or wave packet (Equation 3.98). In Figure 3.47 we illustrate qualitatively how $\psi(x)$ depends on $A(\mathbf{k})$. In Figure 3.47a, we show a sharp peak in $A(\mathbf{k})$ for a narrow range of wave numbers $\mathbf{k}$. The resulting real part of the wave pulse, shown in Figure 3.47b, is relatively broad: a narrow range of $\mathbf{k}$ means a narrow range of $\mathbf{p}_x = \hbar\mathbf{k}$ and thus a small $\Delta\mathbf{p}_x$, and the

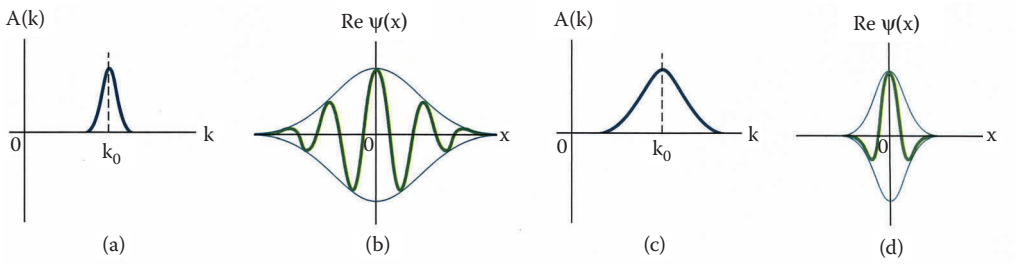


FIGURE 3.47 (a) A sharp peaked $A(\mathbf{k})$ function leads to a wave function $\psi(x)$ with a broad spatial extent (Δx) (b); (c) a broad peaked $A(\mathbf{k})$ function leads to a wave function $\psi(x)$ with a narrow spatial extent (Δx) (d).

result is a relatively large Δx. Broadening the $A(\mathbf{k})$ function in Figure 3.47c results in a more localized wave pulse with a smaller Δx, as clear from Figure 3.47d. In other words, this is Heisenberg's uncertainty principle, $\Delta \mathbf{p}_x \Delta x \geq \hbar$, in action.

Heisenberg's relations are of no practical importance in the macroworld, and in classical physics they can be ignored completely.

The Launchpad: Classical Mechanics Revisited

By 1924, the quantum concepts of Planck, Einstein, Bohr, and de Broglie were widely accepted, and between 1925 and 1927 three distinct, independent, and very different integrating theories of quantum theory were proposed: Dirac's Hamiltonian and quantum algebra representation; the matrix representation of Born, Heisenberg, and Jordan; and Schrödinger's wave equations. Much later, in the late 1940s, Feynman formulated his sum-over-histories approach.

Schrödinger, after attending a seminar on Einstein's and de Broglie's ideas that wave-like entities can behave like particles, and vice versa, thought that there must be a wave equation, $\Psi(x,t)$, to describe particles. Schrödinger almost completely dispensed with the concept of a particle and instead focused on the wave-like properties of matter. His picture of the atom has the electron standing waves vibrating in their orbitals much like the vibrations on a string—but in three dimensions instead of one. In Figure 3.48, a 2D representation of Schrödinger waves, like vibrations on a drum skin, is shown.

A concept that plays an important role in both classical and quantum theory is that of the Hamiltonian of a system. Consider an isolated system composed of one or more particles, and assume that the total energy of this system remains constant. The Hamiltonian of this system is merely its total energy:

$$\mathrm{H} = \mathrm{E_n}(\text{total energy}) = \mathrm{KE}\ (\text{kinetic energy, depends on v}) + \mathrm{PE}\ (\text{potential energy, depends on position}) \qquad (3.84)$$

with the kinetic energy arising from motion and the potential energy arising from the position in a force field, F. The Hamiltonian in quantum mechanics is an example of an operator, a mathematical object that tells us what operation to perform on the function that follows. We rewrite Equation 3.84 here as:

$$\mathrm{H} = \mathrm{E} = \mathrm{E_k} + \mathrm{V(x)} \qquad (3.108)$$

The linear momentum $\mathbf{p}$ equals mv so that $\mathrm{E_k} = \mathrm{mv}^2/2$ or $\mathbf{p}^2/2\mathrm{m}$. When working with the Hamiltonian, the kinetic energy of a particle is expressed as $\mathbf{p}^2/2m$, not as $mv^2/2$. Thus, Equation 3.84 may be rewritten as:

$$\mathrm{H} = \mathrm{E} = \frac{\mathbf{p}^2}{2\mathrm{m}} + \mathrm{V(x)} \qquad (3.109)$$

In case the potential is time varying, the last term in Equation 3.109 must be written out as $V(x,t)$. The total energy, E, in the absence of a potential energy

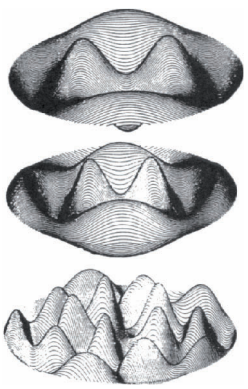


FIGURE 3.48 A two-dimensional representation of Schrödinger waves. Notice the nodes of the vibration.

equals E_k. Newton's second law relates potential energy to change in momentum. According to this law of mechanics, forces give rise to a change in the momentum of particles [(F = d(mv)/dt = dp/dt], and each force, *F*, has an associated potential energy *V*(*x*) (F = −dV/dx). The direction of the force is toward decreasing potential energy:

$$F = \frac{dp}{dt} = ma = m\frac{d^2x}{dt^2}$$
$$F = -\frac{dV(x)}{dx} \qquad (3.110)$$

or:

$$F = -\frac{dV(x)}{dx} = m\frac{d^2x}{dt^2}$$

i.e., given *V*(*x*) one can solve for *x*(*t*) or *v*(*x*,*t*). We illustrate the use of the above deterministic expressions with two important examples so we might better appreciate the analogies and the differences between classical and quantum physics.

Example 3.1: Free particle [no force, V(x) = 0] moving along *x* (1D problem):

$$E = E_k = \frac{mv^2}{2} = \frac{p^2}{2m} \qquad (3.111)$$

v = dx/dt, and:

$$v = \frac{dx}{dt} = \sqrt{\frac{2E_k}{m}} \qquad (3.112)$$

Therefore, we can formulate the differential equation:

$$dx = \sqrt{\frac{2E_k}{m}}dt \qquad (3.113)$$

Integration from x_0 to x_t and from t = 0 to t = t and using Equation 3.112 results in:

$$x_t - x_0 = \sqrt{\frac{2E_k}{m}}t = \sqrt{\frac{2p^2}{2m^2}}t = \frac{p}{m}t = vt \qquad (3.114)$$

With V = 0, knowing x_0 and *p*, we can predict x_t at *t*; in other words, we can predict the trajectory at all times later.

Example 3.2: Harmonic oscillators [F = −kx with *x* the displacement]. The pendulum, which is just a mass on a spring, is an example of a harmonic oscillator, but its physical description also encompasses many objects that oscillate, from tuning forks to oscillating bridges and oscillating skyscrapers (Figure 3.49).

Let us first calculate *V*(*x*) in Equation 3.109 for the given problem. The force F = −dV(x)/dx = −kx (Hooke's law), with *k* the force or spring constant, or:

$$dV(x) = kxdx \qquad (3.115)$$

Integrating, we obtain:

$$V - 0 = k\frac{1}{2}(x^2 - 0^2) \text{ since } V = 0 \text{ at } x = 0$$

or

$$V = \frac{kx^2}{2} \qquad (3.116)$$

The force F = ma = −kx, and we obtain the differential equation:

$$\frac{d^2x}{dt^2} = -\frac{k}{m}x \qquad (3.117)$$

with a solution:

$$x(t) = A\sin\left(\sqrt{\frac{k}{m}}t\right) \qquad (3.118)$$

where *A* is the maximum displacement (x_{max} = A). For a simple harmonic oscillator ω = 2πf = 2π/T and:

$$\omega = \sqrt{\frac{k}{m}} \quad \text{and} \quad T = 2\pi\sqrt{\frac{m}{k}} \qquad (3.119)$$

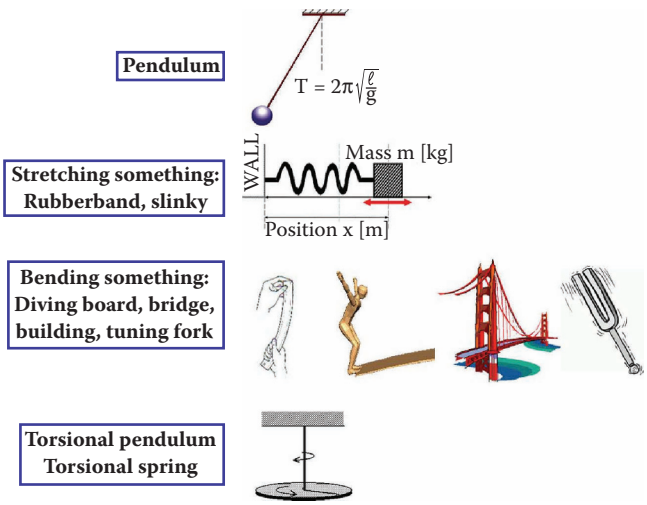


FIGURE 3.49 A wide variety of harmonic oscillator examples.

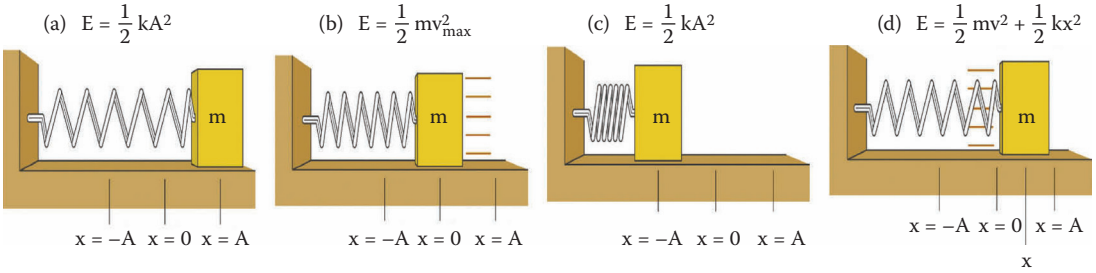


FIGURE 3.50 Spring/mass system with values for kinetic and potential energy as a function of the position of the mass on the *x*-axis. (a) x = A; (b) x = 0; (c) x = –A; and (d) x = x.

Observe that frequency only depends on characteristics of the system (*m*,*k*) and not on the amplitude *A*. The first and second derivatives of Equation 3.118 are given as:

$$v(t) = \frac{dx}{dt} = A\sqrt{\frac{k}{m}}\cos\left(\sqrt{\frac{k}{m}}t\right)$$

and

$$a(t) = \frac{d^2x}{dt^2} = -A\frac{k}{m}\sin\left(\sqrt{\frac{k}{m}}t\right) = -\frac{k}{m}x(t) \qquad (3.120)$$

with $v_{max} = A\omega$, and $a_{max} = A\omega^2$. For the momentum, $p(t) = mv$, we obtain:

$$p(t) = mv = m\frac{dx}{dt} = mA\sqrt{\frac{k}{m}}\cos\left(\sqrt{\frac{k}{m}}t\right) \qquad (3.121)$$

We can now solve for *E* exactly:

$$E = \frac{p^2}{2m} + \frac{kx^2}{2} \text{ or also } E = \frac{mv^2}{2} + \frac{kx^2}{2}$$

$$= \frac{\left[mA\sqrt{\frac{k}{m}}\cos\left(\sqrt{\frac{k}{m}}t\right)\right]^2}{2m} + \frac{k\left[A\sin\sqrt{\frac{k}{m}}t\right]^2}{2} \qquad (3.122)$$

Thus, the total mechanical energy of a simple oscillator is proportional to the square of the amplitude. As the amplitude (*A*) can take any value, this means that the energy (*E*) can also take any value—i.e., energy is continuous. Any energy value is allowed by simply changing the force constant *k*. At x = –A and x = A, $E = kA^2/2$, and at x = 0, $E = mv^2_{max}/2$ (see Figure 3.50).

In Figure 3.51 we show a typical nonquantized oscillator with a parabolic curve for the potential energy as a function of position $x - x_0$ of the mass. The potential sketched here is very important because it describes the potential for many systems, including vibrational and electronic states in molecules. It approximates the potential in many more systems for small departures from equilibrium. As a typical application, in Figure 3.52, we illustrate the potential energy for two hydrogen atoms approaching each other. Clearly over a considerable range of energies, a parabolic curve, typifying a simple harmonic oscillator, can represent the real situation. The horizontal lines cutting through the parabolic part of the curve in Figure 3.52 represent

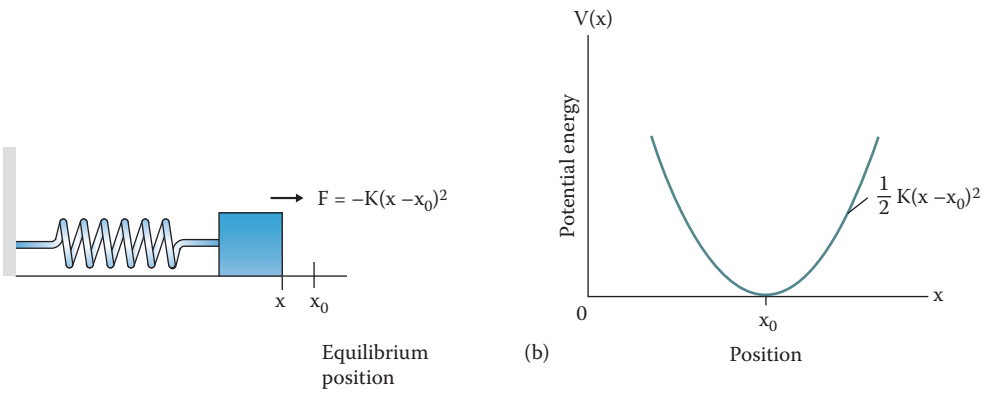


FIGURE 3.51 Typical nonquantized oscillator with a parabolic curve for the potential energy as a function of position *x* of the mass.

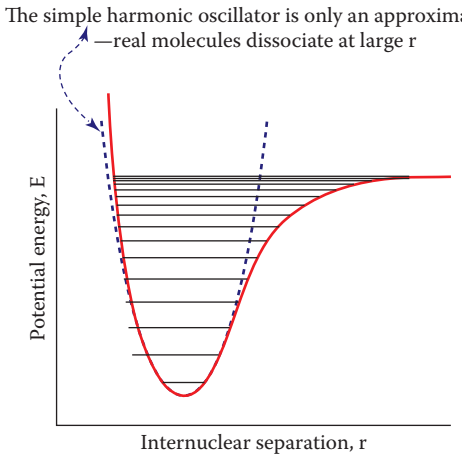


FIGURE 3.52 Application of simple oscillator approximation: two hydrogen atoms in a hydrogen molecule. For an explanation of the horizontal lines in this figure, see quantized oscillator below.

the quantization of the vibration states as introduced next.

The above examples demonstrate the essence of determinism in classical physics, i.e., given $V(x)$, one can solve for $x(t)$ or $v(x,t)$, or also at any time (t), the position $[x(t)]$ and velocity $[v(t)]$ can be determined exactly—i.e., the particle trajectory can be specified precisely. In other words, given the force, the motion can be found. In the late eighteenth century the mathematician Pierre Simon de Laplace (1749–1827) encapsulated classical determinism as follows: "…if at one time we knew the positions and motion of all the particles in the Universe, then we could calculate their behavior at any other time, in the past or the future." In classical physics, particles and trajectories are real entities, and it is assumed that the universe exists independently from the observer, that it is predictable, and that for every effect there is a cause so experiments are reproducible. Heisenberg's uncertainty principle destroyed all this. In quantum physics, measured and unmeasured particles are described differently. The measured particle has definite attributes such as position and momentum, but the unmeasured particle does not have one but all possible attribute values, as Nick Herbert describes it in his book *Quantum Reality* … somewhat like a broken television that displays all its channels at the same time.[3]

We shall see that these ideas of classical mechanics fail when we go to the atomic regime (where E and m are very small). Classical mechanics also fails when velocity is very large (as $v \rightarrow c$) because of relativistic effects.

Schrödinger's Equation

Plausibility of Schrödinger's Equation

From the evidence presented above, an atomic or subatomic particle cannot be described anymore as a simple Newtonian point, and *matter waves* must be taken into account. For matter waves, from $p = h/\lambda$ (Equation 3.95), the smaller the wavelength, the bigger the momentum, and as we calculated, electrons, for example, are very "wavy" particles. The book-keeping term for the energy of a system or the Hamiltonian, H, for a Newtonian particle with mass m is:

$$H = E = \frac{p^2}{2m} + V(x) \qquad (3.109)$$

An equivalent to Newton's equation is needed to calculate how forces (described by a potential energy) affect λ or $\mathbf{p}$ in the case of a "waving" particle. Once that differential wave equation is found, we may solve it for a wave, which has an amplitude for each value of position (x) and time (t).

Erwin Schrödinger (1887–1961), in 1926, encouraged by Debye, who remarked that there should be a wave equation to describe the de Broglie waves, proposed a wave equation that can be applied to any physical system in which it is possible to describe the energy mathematically. In one dimension he postulated:

$$\frac{\partial^2 \Psi(x,t)}{\partial x^2} + \frac{8\pi^2 m}{h^2}[E - V(x,t)]\Psi(x,t) = 0 \qquad (3.123)$$

with $\Psi(x,t)$ the wave function, a wave representing the spatial distribution of a "particle," and m the characteristic mass of the particle. The first term on the left is the rate of change of the wave function with distance x. The energy of the particle is E, and the potential energy function to describe the forces acting on the particle is represented by $V(x,t)$. Schrödinger's equation has the same central role in quantum mechanics that Newton's laws have in

mechanics and Maxwell's equations have in electromagnetism. Solutions to Newton's equations are of the form $v = f(x,t)$ (see Examples 3.1 and 3.2 above), whereas solutions to the wave equation (Equation 3.118) are called wave functions $\Psi(x,t)$. Schrödinger's equation is more difficult to solve than Newton's equation, but it is just as well-defined, and like Newton's equation, it describes the relation between kinetic energy, potential energy, and total energy. If one knows the forces involved, one can calculate the potential energy V and solve the equation to find Ψ. Solving Schrödinger's equation specifies $\Psi(x,t)$ completely, except for a constant; if $\Psi(x,t)$ is a solution, then $A\Psi(x,t)$ is a solution as well. Remember that Equation 3.123, like Newton's law, cannot be derived—it is a plausibility argument. Einstein called Ψ a *Gespensterfeld* or ghost field. Because it carries no energy, the wave function is also referred to as an empty wave. In France, the Ψ wave is sometimes called *densité de présence*, or *presence density*.

Here is how Schrödinger, using a rather sophisticated analogy with classical mechanics, came to derive Equation 3.123. He assumed a sinusoid wave of wavelength λ and frequency ν and hence a velocity $v = \lambda\nu$. The equation of such a traveling wave, as we saw in Chapter 2, is $\Psi(x,t) = A\cos(\mathbf{k}_x x - \omega t) + B\sin(\mathbf{k}_x x - \omega t)$, where $\mathbf{k}_x$ (the wave number) $= 2\pi/\lambda$, and the period is $T = 2\pi/\omega = 1/\nu$. Rewriting this expression in terms of the complex variable form, with $B = iA$ and using Euler's formula, one obtains as the simplest form of a wave:

$$\Psi(x,t) = Ae^{i(k_x x-\omega t)} \tag{3.124}$$

This is classical so far; indeed, we will see in the section "Solution of the Wave Equation for Free Particles", this chapter, and in Chapter 5 that Equation 3.124 is also a solution of the Maxwell equations.

Now apply Einstein's photon formula, i.e., $E = h\nu$, to the particle, with ν the frequency of the "waving" of the particle—de Broglie's idea! Because $E = h\nu = (h/2\pi)(2\pi\nu) = \hbar\omega$, we find $\omega = E/\hbar$, where we have introduced the definition of "h-bar," $\hbar = h/2\pi$. From de Broglie, $p = h/\lambda$ and $p = (2\pi/\lambda)(h/2\pi) = \mathbf{k}_x\hbar$ and $\mathbf{k}_x = p/\hbar$. Substitute the results for ω and $\mathbf{k}$ into Equation 3.124 and obtain:

$$\Psi(x,t) = Ae^{\frac{i(px-Et)}{\hbar}} \tag{3.125}$$

Schrödinger now assumed that the total energy E could be expressed in terms of the kinetic energy (KE) and the potential energy (PE) of the particle:

$$H = E = E_k + V(x) = \frac{p^2}{2m} + V(x,t) \tag{3.126}$$

The derivatives of Equation 3.125 are:

$$\frac{\partial\Psi(x,t)}{\partial x} = \frac{ip}{\hbar}\Psi(x,t)$$

so that

$$\frac{\partial^2\Psi(x,t)}{\partial^2 x} = \frac{-p^2}{\hbar^2}\Psi(x,t) \tag{3.127}$$

and

$$\frac{\partial\Psi(x,t)}{\partial t} = \frac{-iE}{\hbar}\Psi(x,t)$$

Note that Ψ cannot be canceled. Making use of the correspondence of the operators on $\psi(x,t)$ we can write: $\frac{\partial^2}{\partial^2 x} = \frac{-p^2}{h^2}$ and $\frac{\partial}{\partial t} = \frac{-iE}{h}$, and substituting these into the total energy equation (Equation 3.109) one derives:

$$E = i\hbar\frac{\partial}{\partial t} = -\left(\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\right) + V(x,t) \tag{3.128}$$

Applying these operators to $\Psi(x,t)$, we obtain:

$$E\Psi(x,t) = i\hbar\frac{\partial\Psi(x,t)}{\partial t} = -\left(\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2}\right) + V(x,t)\Psi(x,t) \tag{3.129}$$

which, the reader can easily demonstrate, is the famous Schrödinger's wave equation (Equation 3.123) for the 1D case. We may rewrite Equation 3.129 also in the form:

$$\left[-\frac{h^2}{8\pi^2 m}\frac{\partial^2}{\partial x^2} + V(x,t)\right]\Psi(x,t) = E\Psi(x,t) \tag{3.130}$$

This equation shows that H, the Hamiltonian operator, is given by:

$$H = -\frac{h^2}{8\pi^2 m}\frac{\partial^2}{\partial x^2} + V(x,t) \tag{3.131}$$

Thus, Equation 3.130 may simply be formulated in operator form as $H\Psi$ (operator Ψ acting on function

Ψ, an eigenfunction) = EΨ(function Ψ multiplied by a number E, an eigenvalue), where H is the 1D Hamiltonian operator and in which the energy E of the particles is called the eigenvalue, and Ψ the eigenfunction. Note again that Ψ cannot be canceled. The latter merely represents the wave associated with the particle. Expressed yet another way, kinetic and potential energies are transformed into the Hamiltonian, which acts on the wave function to generate the evolution of the wave function in time and space. Schrödinger's equation gives the quantized energies of the system and gives the form of the wave function so that other properties may be calculated.

The above does not represent a derivation of the wave equation; it is just a description of Schrödinger's thought process to make his postulate more plausible.

Wave Function Interpretation

Schrödinger did not have a clear idea of the meaning of a matter wave. However, in 1926 Max Born (1882–1970) presented a new, vivid interpretation of the particle wave function. The so-called "Copenhagen interpretation" of Schrödinger's equation is that the Ψ(x,t) function is not some physical representation of a physical substance as in classical physics (e.g., the amplitude of a water wave), but rather a "probability amplitude" of the particle, which, when squared, gives the probability of finding the particle at a given place at a given time: $|\Psi(x,t)|^2dx$ = probability the particle will be found between x and $x + dx$ at time t and the wave function itself has no physical meaning. Because Ψ(x,y,z,t) is complex and can be positive or negative, it cannot be the probability directly. The Born interpretation of Ψ places restrictions on the form of the wave function:

1. Ψ must be continuous (no breaks).
2. The gradient of Ψ($d\Psi/dx$) must be continuous (no kinks).
3. Ψ must have a single value at any point in space.
4. Ψ must be finite everywhere.
5. Ψ cannot be zero everywhere.

The Copenhagen interpretation also holds that an unmeasured particle is not real: its attributes are created or realized by the measuring act. Another way of saying this is that a wave function *collapses* up on measurement; before measurement a particle is described by a wave function described by Schrödinger's equation, but on measuring that particle's wave suddenly and discontinuously *collapses*. We will come back to this *mystic* interpretation of quantum reality when introducing *Schrödinger's cat* at the end of this chapter. Because the probability that the particle is somewhere must equal one, it holds that one can normalize this probability function as:

$$\int_{-\infty}^{+\infty} |\Psi(x,t)|^2 dx = 1 \qquad (3.132)$$

Note that the probability is a real number; although Ψ is complex, $|\Psi(x,t)|^2$ is real. In this case, Ψ is said to be a normalized wave function. Electrons do not fly around the nucleus like the Earth around the sun (Rutherford, Bohr), but depending on which energy level it is in, the electron can take one of a number of stationary probability cloud configurations (Schrödinger).

The boundary conditions imposed on Ψ mean that only certain wave functions and thus only certain energies of the system are allowed. Quantization of the wave function leads to quantization of the energy.

As we will learn below, solutions of Schrödinger's equation for an atom are spherical Bessel functions. In Figure 3.53 we show, as an example of the type

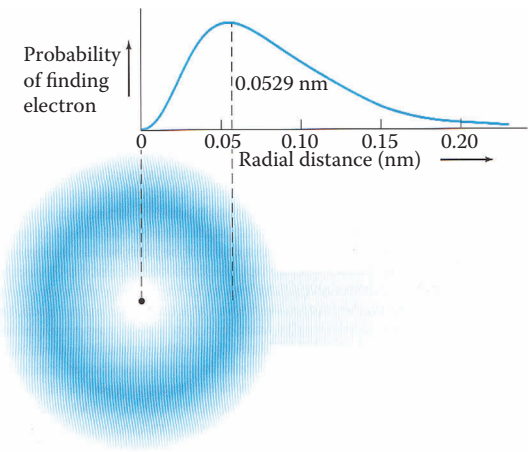


FIGURE 3.53 The probability of finding the ground state hydrogen electron (n = 1) as a function of the radial distance from the proton. The value of $|\Psi(x,t)|^2$ at some location is proportional to the probability of finding the particle at that location at that time.

of solutions obtained, the probability of finding an electron around the nucleus of a hydrogen atom. The potential that must be used in Schrödinger's equation for this case is $V(r) \propto 1/r$. Where we assume that the Coulomb force between the electron and the nucleus is the force responsible for binding the electron in the atom, this is the so-called central force or inverse square law $[F(r) \propto \frac{1}{r^2}]$.

Within Schrödinger's model the atom is regarded as a sort of vibrating balloon that extends to infinity and whose vibrations are in tune with Bohr's frequencies. The quantum numbers of Bohr and Sommerfeld are related to the number of nodes in this vibrating 3D system. This theory of matter waves also reproduces the Balmer series for the bright lines in the hydrogen atom. Schrödinger removed the mysterious discontinuous jumps between electron orbitals, replacing them with "beats" between the vibration frequencies of different quantum states. Schrödinger's elegant wave theory explains many things but lacks elsewhere, e.g., it cannot explain quantum processes such as the photoelectric effect.

The Time-Independent Schrödinger Equation (TISE) for Stationary States

The value of $|\Psi(x,t)|^2$ in Equation 3.132 at a particular point is in general a function of time. However, if the particle under consideration has a definite energy—think about an electron in a specific energy level in an atom—then the value of $|\Psi(x,t)|^2$ at each point becomes independent of time. It follows from quantum mechanics that for a particle in such a state of definite energy E, Ψ can be factored into a time-dependent component and a space-dependent component: $\Psi(x,t) = \psi(x)\psi(t)$:

$$\Psi(x,t) = \underset{\text{spatial}}{\psi(x)} \underset{\text{temporal}}{e^{-\frac{iEt}{\hbar}}} \qquad (3.133)$$

This is the time-dependent wave function for a stationary state where we use Ψ if the wave is a function of all coordinates and time, and ψ if it is a function of the space coordinates only. For simplicity we just deal with the x-coordinate here. The probability distribution function $|\Psi|^2$ for this state is the product of Ψ and its complex conjugate Ψ^*, or:

$$\begin{aligned} \left|\Psi(x,t)\right|^2 &= \Psi^*(x,t)\Psi(x,t) \\ &= \psi^*(x)\psi(x)e^{+\frac{iEt}{\hbar}}e^{-\frac{iEt}{\hbar}} \\ &= \psi^*(x)\psi(x)e^0 \\ &= |\Psi(x)|^2 \end{aligned} \qquad (3.134)$$

As $\psi(x)$ does not depend on time, we see from Equation 3.134 that the probability function does not depend on time either. As soon as one can define a state with a definite energy, one can define a stationary state.

We can now substitute the result from Equation 3.133 into the time-dependent Schrödinger's equation (Equation 3.129) to get:

$$i\hbar\psi(x)\frac{\partial\psi(t)}{\partial t} = -\left(\frac{\hbar^2}{2m}\psi(t)\frac{\partial^2\psi(x)}{\partial x^2}\right) + V(x,t)\psi(x)\psi(t) \qquad (3.135)$$

Dividing both sides by $\psi(x)\psi(t)$ results in:

$$i\hbar\frac{1}{\psi(t)}\frac{\partial\psi(t)}{\partial t} = -\left(\frac{\hbar^2}{2m}\frac{1}{\psi(x)}\frac{\partial^2\psi(x)}{\partial x^2}\right) + V(x,t) \qquad (3.136)$$

If the potential V is independent of t, then $V(x,t) = V(x)$, and the left side in Equation 3.136 depends on t only and the right side on x only. The only way these two sides can be equal to each other is if they are both equal to a constant, i.e., E. In the case $V(x,t)$ is independent of time, Equation 3.130 can thus be converted into a time-independent Schrödinger's equation (TISE). Hence we obtain the time-independent form of Schrödinger's equation as:

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)\right]\psi(x) = E\psi(x) \qquad (3.137)$$

Solving this equation, say for an electron acted on by a fixed nucleus, we will see that this results in standing waves.

The more general Schrödinger equation features a time-dependent potential $V = V(x,t)$ and must be used, for example, when trying to find the wave function of an atom in an oscillating magnetic field or other time-dependent phenomena such as photon emission and absorption.

Applications of Schrödinger's Equation

Solution of the Wave Equation for Free Particles

The wave function of free particles, such as photons, phonons, plasmons, and "nearly free" particles, such as conduction electrons in metals, should all be solutions of Schrödinger's equation. In the case of freely traveling photons, the expression for planar light waves, $\Psi(x,t) = Ae^{i(k_x x-\omega t)}$ (1D case), derived from Maxwell's equations in Chapter 5, according to Bohr's correspondence principle, should also be a solution of Schrödinger's equation. To solve this first problem, we start from Equation 3.137, with V(x) = 0. This represents a free particle that experiences no force, has a definite energy E, and is moving in the x-direction with a momentum $\mathbf{p}_x$ (see Figure 3.54), say a photon:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = E\psi \text{ or } \frac{d^2\psi}{dx^2} = -\frac{2mE}{\hbar^2}\psi \qquad (3.138)$$

This is a second-order differential equation whose solutions are functions that, when differentiated twice, yield back the same functions multiplied by a constant. Such solutions include sines, cosines, and exponentials. Specifically, in this case, solving Equation 3.138 leads to the general solution, consisting of the superposition of two spatial wave functions of the form:

$$\psi(x) = Ae^{ik_x x} \qquad (3.139)$$

The two traveling waves, one traveling in the $+x$-direction and one traveling in the $-x$-direction, can be rewritten (using Euler) as:

$$\psi(x) = A_1[\cos(\mathbf{k}_x x) + i\sin(\mathbf{k}_x x)] + A_2[\cos(-\mathbf{k}_x x) + i\sin(-\mathbf{k}_x x)] \qquad (3.140)$$

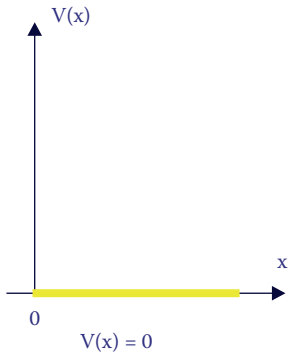


FIGURE 3.54 Photon, no boundary conditions, except for V(x) = 0.

where A_1 and A_2 are constants. We expected this result, of course, because in the preceding section we saw how Schrödinger built up his equation with the requirement that it would, at minimum, yield the same results that Maxwell's equations provide. This wave function for a free particle has a definite momentum $\mathbf{p}_x$ in the x-direction, or $\Delta\mathbf{p}_x = 0$. From the Heisenberg uncertainty principle (Equation 3.106), it then follows that Δx must be infinite; we have no idea where the particle is located in space. This makes sense as we are dealing with traveling waves. To confirm this further, we calculate the probability distribution function $|\Psi|^2$ for a free photon, which is the product of Ψ and its complex conjugate Ψ^*, or:

$$\begin{aligned}|\Psi(x,t)|^2 &= \Psi^*(x,t)\Psi(x,t)\\ &= (A^*e^{-ik_x x}e^{+i\omega t})(Ae^{ik_x x}e^{-i\omega t})\\ &= A^*Ae^0 = |A|^2\end{aligned} \qquad (3.141)$$

This result is independent of space and time, something we again expect for a sinusoidal wave function that extends all the way from $x = -\infty$ to $x = +\infty$ with an amplitude A. Normalization is not possible here as the wave extends to infinity; integration of Equation 3.139 over all space is infinite for any value of A.

By substituting $\psi(x) = Ae^{ik_x x}$ into Equation 3.138, we derive:

$$E = \mathbf{p}^2/2m \qquad (3.142)$$

and also:

$$\mathbf{k}_x = \frac{\sqrt{2mE}}{\hbar} \qquad (3.143)$$

With a positive $\mathbf{k}_x$, the wave function represents a free particle moving in the positive x-direction. If $\mathbf{k}_x$ is negative, the motion is in the negative x-direction. For a free particle, there is no restriction on the value of $\mathbf{k}_x$, and the associated, unquantized energy from Equation 3.143 is given as:

$$E(\mathbf{k}) = \frac{\hbar^2\mathbf{k}_x^2}{2m} \qquad (3.144)$$

Any positive value of energy is allowed for such a free wave/particle; there is nothing that restricts the values of E. Equation 3.144 is the so-called dispersion relation for free particles. When we plot the

energy E versus the wave vector $\mathbf{k}_x$ for free particles, we obtain the parabola shown in Figure 3.55. It is appreciated from this plot that mass, m, is inversely proportional to curvature.

Free Electrons in an Infinite Piece of Metal

For free electrons in an infinitely large 3D piece of metal, the allowed electron states are solutions of an expanded version of Schrödinger's equation in Equation 3.123, or:

$$\nabla^2\psi(\mathbf{r}) + \frac{8\pi^2 m}{h^2}[E - V(\mathbf{r})]\psi(\mathbf{r}) = 0 \quad (3.145)$$

For electrons swarming around freely in this infinite metal, the potential energy $V(r)$ is zero inside the conductor, and the solutions inside the metal are plane waves moving in the direction of $\mathbf{r}$:

$$\Psi_k(\mathbf{r}) = A\exp i\mathbf{k}\cdot\mathbf{r} \quad (3.146)$$

where $\mathbf{r}$ is any vector in real space, and $\mathbf{k}$ is any wave vector. As with a freely moving particle, normalization is impossible as the wave extends to infinity.

Plotting the energy E versus the wave number $\mathbf{k}_x$ for a free electron gas in one direction leads to the same parabolic dispersion relation shown in Figure 3.55. The density of states (DOS) function $G(E)$ is the number of possible energy states per unit volume (also degeneracy), and we will derive it further below for free electrons in a bulk piece of solid. We find that, just as in the case of the density of states of an ideal gas (Equation 3.19), it increases smoothly with the square root of the energy. Importantly, we will also find that when reducing the dimensionality of the solid, say to a plane, a line, or a dot, the density of state function changes dramatically and acquires more density of states at specific energy values; i.e., $G(E)$ is not a smooth function of E anymore.

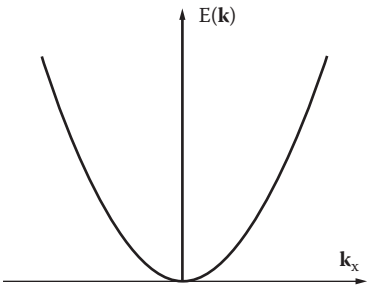


FIGURE 3.55 Plot of energy versus wave number $\mathbf{k}_x$ from Equation 3.144 for free particles. Dispersion relation $E(\mathbf{k}_x)$.

From classical theory we could not appreciate the occurrence of long electronic mean free paths; indeed, Drude used the interatomic distance a for the mean free path λ. But from experiments with very pure materials and at low temperatures it is clear the mean free path may be much longer; actually it may be as long as 10^8 or 10^9 interatomic spacings or more than 1 cm. The quantum physics answer is that the conduction electrons are not deflected by ion cores arranged in a periodic lattice because matter waves propagate freely through a periodic structure just as predicted by Equation 3.146.

Confining electrons by limiting their propagation in certain directions in a crystal introduces a varying $V(r)$ in Schrödinger's equation, and this may lead to an electronic bandgap as we will introduce below.

Particles in Infinitely Deep Potential Wells of Finite Size

The Born and von Karman's Periodic Boundary Condition According to Pauli's exclusion principle, in an atom, no two electrons can have all four quantum numbers the same. We ask ourselves now if there are similar restrictions for electrons in a larger structure. Instead of the infinitely large piece of metal, considered above, we limit the size of the metal chunk to say a 1-cm^3 piece of metal, and we find that restrictions for electron energies do indeed materialize. Discrete energy levels inevitably arise whenever a small particle such as a photon or electron is confined to a region in space. Sommerfeld, in 1928, was the first to show this. He adopted Drude's free electron or Fermi gas (FEG) and added the restriction that the electrons must behave in accordance with the rules of quantum mechanics. In his Fermi gas, electrons are free, except for their confinement within a cubic piece of crystalline conductor with a volume of $V = L^3$, and they follow Fermi-Dirac statistics instead of Maxwell-Boltzmann rules. The choice of a cube shape is a matter of mathematical convenience; a periodic boundary condition ensures that the free electron form of the wave function is NOT modified by the shape of the conductor or its boundary. This can be interpreted as follows: an electron coming to the surface is not reflected back in but reenters the metal piece from the opposite surface. This excludes the surfaces from playing any

role in transport phenomena. The value L is set by the Born and von Karman's periodic boundary condition, i.e., that the wave functions must obey the following rule:

$$\psi(x + L, y + L, z + L) = \psi(x,y,z) \qquad (3.147)$$

For the derivation of the possible energies, we assume N electrons (one for each metal ion) in a cube of solid conductor with sides of length L.

Outside the 3D cube of solid the potential $V \rightarrow \infty$, and the wave function ψ is zero anywhere outside the solid with $x,y,z \leq 0$ and $x,y,z \geq L$. The situation applies, for example, to totally free electrons in a metal where the ion cores do not influence their movement. Sommerfeld assumed that $V(x)$ outside the conductor equaled the work function Φ. The work function is the amount of energy required to remove an electron from the surface of a metal, i.e., the height of the wall electrons would have to scale to escape the solid, but $V = \infty$ is a good enough approximation for electrons in low-energy levels. Before we apply the Born and von Karman's periodic boundary condition to a finite-size 3D box that is infinitely deep ($V = \infty$), we consider finite-size infinitely deep 1D and 2D wells.

Infinitely Deep, Finite-Sized 1D Potential Wells—Quantum Wells We apply Equation 3.137 now to a finite-sized 1D box with infinitely high potential walls (Figure 3.56). In a 1D box, $V(x)$ in Equation 3.137—the time-independent Schrödinger equation (TISE)—is 0 everywhere inside the conductor (region II) and is infinite outside the conductor (regions I and III).

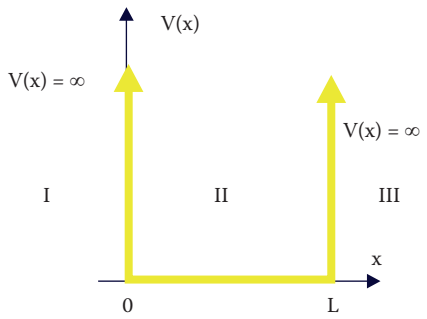


FIGURE 3.56 Electron in a box. L is not the real physical boundary of the conductor, as the surfaces of the conductor are not determining the physical properties. The value of L is set by the Born and von Karman periodic boundary condition (see Equation 3.147).

Outside the well $V \rightarrow \infty$ and $\psi = 0$ for $x \leq 0$ and $x \geq L$. For the Schrödinger equation inside the well ($0 < x < L$), we write:

$$-\frac{\hbar^2}{2m_e}\frac{d^2\psi(x)}{dx^2} = E\psi(x) \qquad (3.148)$$

or:

$$\frac{d^2\psi}{dx^2} = -k_x^2\,\psi \qquad (3.149)$$

with:

$$k_x = \sqrt{\frac{2m_eE}{\hbar^2}} \qquad (3.150)$$

We are interested in the quantum mechanically allowed energy levels of the electrons in this 1D box; because $V(x) = 0$, these will be kinetic energy levels (from Equation 3.109, $E = p^2/2m_e$). From Equation 3.148 the kinetic energy is proportional to the curvature of ψ (the curvature of a function is its second derivative).

In Sommerfeld's mathematical model, for $x = 0$ and at $x = L$, the wave function must be zero. The general stationary state solution for Equation 3.149 consists of the superposition of two spatial wave functions $\psi(x)=Ae^{ik_xx}$, as shown in Equation 3.140, which we rewrite here as:

$$\psi(x) = (A_1 + A_2)\cos k_x x + i(A_1 - A_2)\sin k_x x \qquad (3.151)$$

with one traveling in the $+x$-direction and one traveling in the $-x$-direction. The boundary conditions are set as follows: $x = 0$, $\psi(x) = A_1 + A_2 = 0$ or $A_2 = -A_1$, and we rewrite Equation 3.151 as:

$$\psi(x) = 2iA_1 \sin k_x x = \psi_0 \sin k_x x \qquad (3.152)$$

From the second boundary condition that $\Psi_0 \sin(k_xL)$ must be 0, it follows that either $\psi_0 = 0$ or $\sin(k_xL) = 0$. If $\psi_0 = 0$, the wave function is zero everywhere, or there is no probability of finding the particle in the box anywhere; therefore, this solution must be disregarded. Hence $k_xL = n\pi$, where $n = 1, 2, 3,\ldots$ or:

$$\psi(x) = \psi_0 \sin\left(\frac{n\pi x}{L}\right) \qquad (3.153)$$

This is a standing wave that acts like a string clamped down at both ends. With n = 1 one has the basic harmonic, and with n = 2, 3,... higher harmonics result. Note that the constant, ψ_0, cannot be determined from solving Schrödinger's equation because we are dealing with a linear equation. The constant must be recovered from the normalization condition, which is required to give a probability interpretation to the wave function. The constant ψ_0 is calculated as:

$$\int_{-\infty}^{+\infty} |\Psi(x,t)|^2 \, dx = 1 = |\psi_0|^2 \int_0^L \sin^2\left(\frac{n\pi x}{L}\right) dx = |\psi_0|^2 \frac{L}{2} \tag{3.154}$$

or we find:

$$\Psi_n(x,t) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \tag{3.155}$$

with n = 1,2,3,…. These are the eigenfunctions of an electron in a box. The discrete energies—energy eigenvalues—that the electron can adopt are:

$$E_n(\mathbf{k}) = \frac{\hbar^2}{2m_e}\left(\frac{n\pi}{L}\right)^2 \tag{3.156}$$

The values of $E_n(\mathbf{k})$ for different quantum numbers *n* represent the various allowed energy levels in a 1D box, and the gap between two successive levels describes the effect of quantization (discreteness). At the lowest energy (n = 1), the *ground state*, the energy remains finite despite the fact that V = 0 inside the region. According to quantum mechanics, an electron cannot be inside the box and have zero energy. This is called the zero-point energy, an important consequence of Heisenberg's uncertainty principle (we elaborate on this point on p. 114 "Heisenberg's Uncertainty Priniciple"). If V(x) ≠ 0, everywhere in the box, all energies are shifted by V ($E_n = \frac{n^2h^2}{8m_eL^2} + V$). For the same value of quantum number *n*, the energy is inversely proportional to the mass of the particle and to the square of the length of the box. For a heavier particle and a larger box, the energy levels become more closely spaced. Only when m_eL^2 is of the same order as h^2 do quantized energy levels become important in experimental measurements (with L = 1 nm, $E_1 = \frac{h^2}{8m_eL^2} = 0.36$ eV). With a 1-cm^3 piece of metal (instead of 1 nm^3), the energy levels become so closely spaced that they seem to be continuous ($E_1 = \frac{h^2}{8m_eL^2} = 3.6 \times 10^{-15}$eV); in other words, the quantum mechanics formula gives the classical result for dimension such that $m_eL^2 >> h^2$. This is another nice illustration of Bohr's correspondence principle: for large dimensions, Schrödinger's equation yields the classical results back. Because the kinetic energy is proportional to the curvature of the wave function, a higher kinetic energy E_n, caused by a higher value of *n* and/or a smaller value for *L*, corresponds to a more curved wave function (shorter wavelength). On a molecular scale, because L^2 appears in the denominator of Equation 3.156, increasing *L*, the size of the box, stabilizes the particle. Chemists are familiar with this effect in the case of electrons where delocalization is a stabilizing factor. As a consequence, allyl and benzyl carbonium ions, radicals, and carbanions are relatively stable.

The wave number $\mathbf{k}_x$ given as:

$$\mathbf{k}_x = \frac{2\pi p}{h} = \frac{p}{\hbar} = \frac{\sqrt{2m_eE_n}}{\hbar} \tag{3.150}$$

is now also quantized because E_n is quantized so that $\mathbf{k}_x = n\pi/L_x$ where *n* can take on only values 1, 2, 3….

Quantization occurs because of boundary conditions and the requirement for ψ to be physically reasonable (Born interpretation). The quantum number *n* labels each allowed state (ψ_n) of the system and determines its energy (E_n). Knowing *n*, we can calculate ψ_n and E_n. The wave number is related to the momentum $\mathbf{p}_x$ of the electron, viewed as a particle, by $\mathbf{k}_x = 2\pi\mathbf{p}_x/h$, and to the wavelength (λ) of the electron, viewed as a wave, by $\mathbf{k}_x = 2\pi/\lambda$. The dispersion function $E(\mathbf{k}_x)$ for an electron in an infinite deep well of finite size is shown in Figure 3.57. The dispersion or energy spectrum $E(\mathbf{k}_x)$ is now discrete rather than continuous; the allowed wave vectors are uniformly spaced in **k**-space with a separation of π/L; and the sample size *L* determines the spacing of allowed wave vectors and single-particle energies, with a smaller box giving

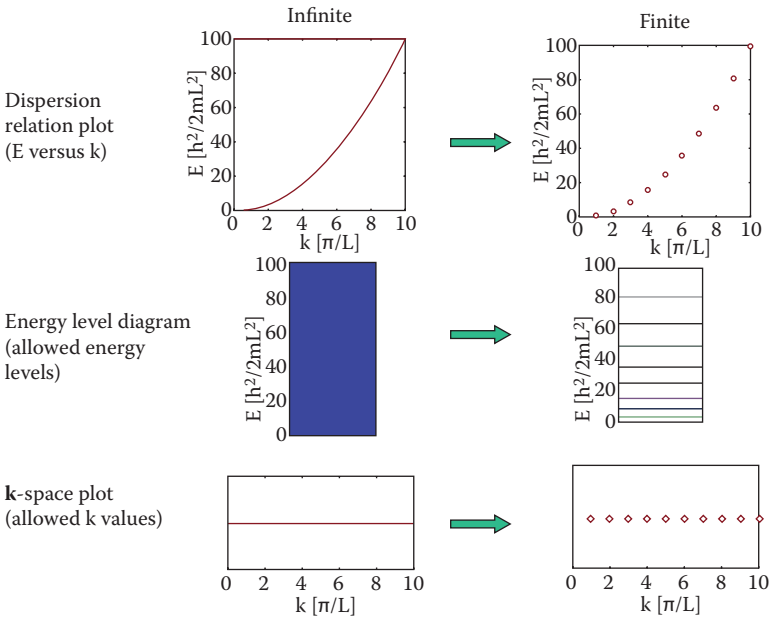


FIGURE 3.57 Dispersion $E(\mathbf{k}_x)$ function for an electron in an infinite box compared with the dispersion of the electron in a finite-size box. Finite size drastically alters allowed energy levels.

the larger spacings. For the free electron, no such values exist (see Figure 3.55).

The ground state (n = 1) of the wave function ψ_1 (Equation 3.150) has no "zero crossings." The first excited state (n = 2) has one zero crossing, and so on. Successive functions possess one more half-wave (they have a shorter wavelength). Nodes in the wave function are points at which $\psi_n = 0$ (excluding the ends, which are constrained to be zero), and the number of nodes = n − 1. The important point to notice is that the imposition of the boundary conditions has restricted the energy to discrete values. The allowed wave functions, a family of standing waves, and energies of an electron in a 1D box are summarized in Figure 3.58.

As apparent from Figure 3.58, the probability distribution of the particle in a box is not uniform; however; for a very large n, the probability is almost uniform throughout the box, as dictated again by classical physics. The probability of finding the particle at some point varies with the energy of the particle. A particle with energy E_1 is most likely to be found in the middle of the box. A particle with energy E_2 will never be found at that spot. As the energy of the particle increases, so does the number of nodes in the eigenfunction. Increasing the number of nodes (decreasing λ) corresponds to increasing kinetic energy.

Devices that come with a length L in one direction comparable with the size of the de Broglie wavelength of an electron are known as quantum wells (1D confinement). With at least one dimension between 1 and 100 nm, the excess electrons in this confined direction have no room to move in a Newtonian fashion. Instead their positions and velocities take on a probabilistic nature as their wave nature now dominates. Charge carriers are still free to move in the other two unconstrained directions though and

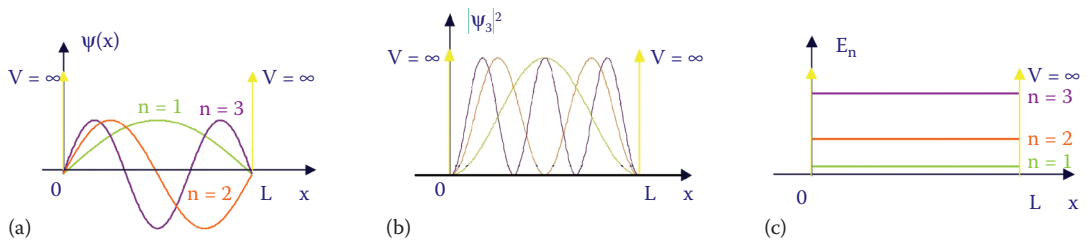


FIGURE 3.58 Particle in a box: overview. (a) Standing waves in the box. (b) The probability distribution $|\psi_3|^2$. (c) The energy is proportional to n^2.

form a 2D carrier gas in a plane perpendicular to the confinement direction (the *x*-*y* plane if the confinement is in the *z*-direction). Quantum wells are relatively easy and cheap to fabricate. By using two different types of semiconductor materials, one of these having a wider bandgap, quantum wells are formed on sandwiching a thin layer of a narrow bandgap material between wider bandgap semiconductor layers. This is realized, for example, by sandwiching a thin epitaxial layer of GaAs between two layers of $Al_xGa_{1-x}As$, in a so-called heterostructure as depicted in Figure 3.59. The narrower bandgap GaAs is enclosed by $Al_xGa_{1-x}As$, a material with a considerably larger bandgap to establish a potential barrier for electrons at the surface of the confined material. The motion of electrons and holes is thereby restricted in the direction perpendicular to the thin layer of GaAs (*z*-direction). In this structure, electrons are still free to move unrestricted in the *x*-*y* plane.

These quantum wells (QWs) were developed in the early 1970s and constituted the first lower dimensional heterostructures. The foremost advantage of such a design involves their improved optical properties. Reduced dimensionality leads to marked improved optical performances because of the change in the density of state or DOS function (see p. 161 "Fermi Surfaces, Brillouin Zones, Density of State Functions, and Conductivity as a Function of Quantum Confinement" for more details) for such a device compared with its 3D counterpart. In contrast to a bulk semiconductor, in a QW there are no allowed electron states at the very lowest energies (an electron in a box with energy = 0 does not exist; see p. 125 "Infinitely Deep, Finite-Sized 1D Potential Wells—Quantum Wells"), but there are many more available states (higher DOS) in the lowest conduction state in a QW so that many more electrons can be accommodated. Similarly, the top of the valence band has plenty more states available for holes. This means that it is possible for many more holes and electrons to combine and produce photons with identical energy for enhanced probability of stimulated emission (lasing) (see Chapter 5). Optical properties can be tuned by changing the structural parameters of a QW, principally its thickness and composition; this is known as bandgap engineering.

Today QWs form the basis of most optoelectronic devices. QWs are used, for example, in high electron mobility transistors (HEMTs) and solar cells with high efficiency. HEMTs are in use in all types of high frequency electronics such as cell phone and satellite television. The reason for the higher speed of a HEMT over a classical transistor is that the mean free path of the charge carriers in the 2D GaAs layer of a HEMT can be made larger than the gate length of the transistor, resulting in ballistic transport, i.e., charge transport without any intervening collisions with other charge carriers. Another mature application of QWs involves solid-state lasers. With the right voltages applied over a single QW, very large numbers of holes and electrons can be brought together in a tiny physical space and narrow energy range, leading to powerful surface emitting lasers as used in $5 laser pointers, compact disc players, and laser printers.

Almost 30 years after quantum wells were first developed, the 2000 Nobel Prize in Physics was awarded to Zhores Alferov and Herbert Kroemer (Figure 3.60) for their contributions in the field of semiconductor heterostructures and their high-speed and optoelectronics applications.

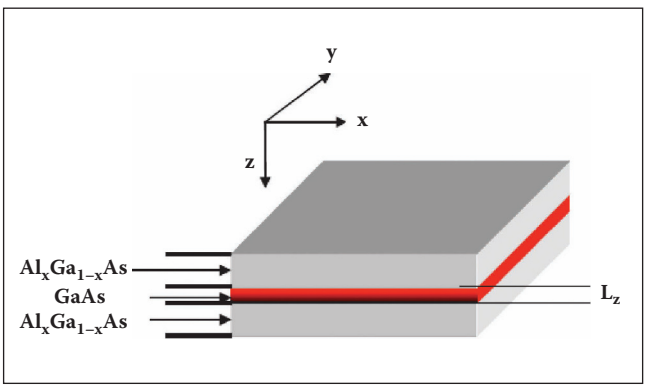


FIGURE 3.59 Quantum well with *x*,*y* dimensions infinite and L_z finite.[4]

FIGURE 3.60 Zhores Alferov (a) and Herbert Kroemer (b) were awarded the Nobel Prize in Physics in 2000.

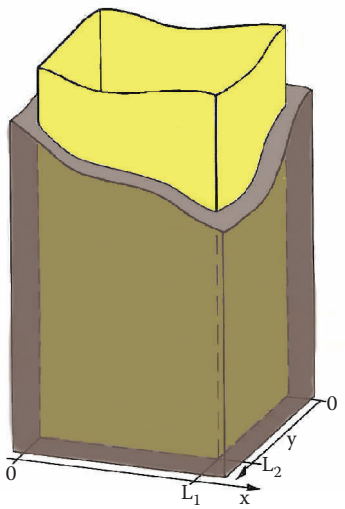


FIGURE 3.61 A 2D square well. A quantum wire.

Led by the insights garnered on QWs, scientists investigated the possibility of reducing the dimensionality of heterostructures even further to create 1D (quantum wire) and 0D (quantum dot) structures (see following sections).

Infinitely Deep, Finite-Size 2D Potential Wells—Quantum Wires For a particle in a finite-sized, 2D infinitely deep potential well, we define a wave function similar to the 1D potential well, but now we obtain $\psi(x,y)$ solutions that are defined by two quantum numbers, one associated with each confined dimension. The pertinent wave functions and energies for a 2D infinitely deep potential well, as shown in Figure 3.61, are:

$$\Psi_{n_1n_2}(x,y) = \left(\frac{4}{L_1L_2}\right)^{\frac{1}{2}} \sin\left(\frac{n_1\pi x}{L_1}\right)\sin\left(\frac{n_2\pi y}{L_2}\right) \quad (3.157)$$

and:

$$E_{n_1n_2}(\mathbf{k}) = E^x_{n_1} + E^y_{n_2} = \frac{h^2}{8m}\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2}\right) \quad (3.158)$$

with $n_1 = 1, 2, 3$ and $n_2 = 1, 2, 3$ and where L_1 is along the x-axis and L_2 along the y-axis, making the z-direction the unconstrained axis. The energy is quantized now along the x- and y-axes as a consequence of the boundary conditions, and most of the features of the 1D well are reproduced. Some of the low-energy wave functions together with their contour maps are illustrated in Figure 3.62.

One new feature in 2D potential wells, not found in 1D potential wells, is degeneracy, i.e., the occurrence of several quantum states at the same energy level. This is best understood through an inspection of Equations 3.157 and 3.158: for a square well instead of a rectangular one, $L_1 = L_2 = L$, and

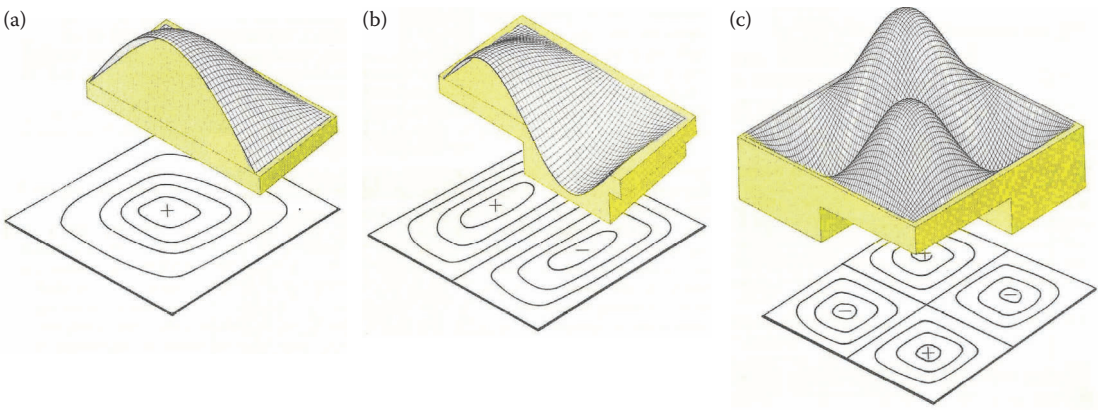


FIGURE 3.62 (a) ψ_{11}, (b) ψ_{21}, and (c) ψ_{22} for a 2D square potential well. A quantum wire.

Equation 3.158 yields now the same energy of $E_{n_1n_2}(\mathbf{k}) = E^x_{n_1} + E^y_{n_2} = \frac{h^2}{8mL^2}(n_1^2 + n_2^2)$. A state with $n_1 = b$ and $n_2 = a$ has the same energy as a state with $n_2 = a$ and $n_2 = b$ (even if $a \neq b$); these different states that correspond to a same energy are called degenerate. That these are indeed different quantum states becomes obvious when substituting the quantum numbers for such states in Equation 3.157. In the case of two degenerate states with $a = 1$ and $b = 2$, both corresponding to an energy of $\frac{5h^2}{8mL^2}$, we obtain two different wave functions:

$$\Psi_{1,2}(x,y) = \left(\frac{2}{L}\right)\sin\left(\frac{\pi x}{L}\right)\sin\left(\frac{2\pi y}{L}\right) \quad \text{and}$$

$$\Psi_{2,1}(x,y) = \left(\frac{2}{L}\right)\sin\left(\frac{2\pi x}{L}\right)\sin\left(\frac{\pi y}{L}\right).$$

Degeneracy is obviously connected to the degree of symmetry of a system; in the case we make an asymmetric box with $L_1 \neq L_2$ (rectangle instead of a square), the degeneracy disappears.

When *L* is very small along two directions (2D confinement)—of the order of the de Broglie wavelength of an electron—one obtains a quantum wire where electrons can only move freely in one direction, i.e., along the length of the quantum wire as illustrated in Figure 3.63.

Examples of 2D quantum confinement comprise nanowires and carbon nanotubes. Quantum wires represent the smallest dimension feasible for efficient transport of electrons and are thus aimed at as the ultimate interconnects in nanoelectronics and nanooptoelectronics. We mentioned above that quantum well lasers are superior over traditional-bulk solid-state lasers. Structures with yet lower dimensionality, such as nanowires and quantum dots (see next section), are even better, coming with a lower threshold current and switching on and off faster than quantum well lasers (~40 GHz and higher). Quantum wires also have been made into transistors (bipolar and FET), inverters, LEDs, and memory structures.

Compared with the fabrication of quantum wells, the realization of nanoscale quantum wires requires more difficult and precise growth control in the lateral dimension, and, as a result, quantum wire applications are only in the development stage. Quantum wire fabrication techniques include nanoscale lithography, self-organization, selective growth, and chemical and electrochemical synthesis (see Volume III, Chapter 3). In the top-down approach, taking advantage of well-developed quantum well fabrication technologies, using molecular beam epitaxy (MBE) and metalorganic vapor deposition (MOCVD), the most straightforward method to realize 1D nanostructures is etching (and regrowth) through wire-defining masks placed above a quantum well. This way, GaAs quantum wires are fabricated starting from the same thin layer of GaAs sandwiched between two layers of Al_xGa_{1-x} As we encountered in the manufacture of quantum wells. In the bottom-up approach, quantum wires are formed via direct growth in the form of semiconductor or metal nanowires and carbon nanotubes. Sumio Iijima (Figure 3.64) discovered multiwall carbon nanotubes (MWNTs) in 1991, after experimenting with

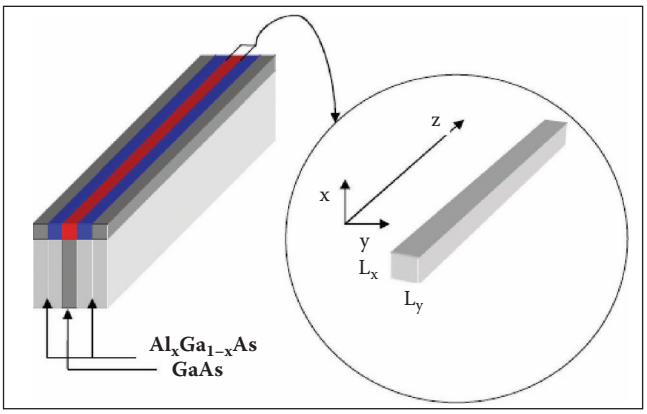


FIGURE 3.63 Quantum wire with dimensions *z* infinite and L_x, L_y finite.[4]

FIGURE 3.64 NEC's Sumio Iijima discovered multiwall carbon nanotubes (MWNTs) in 1991.

an arc-discharge technique similar to the one used by Richard Smalley and his team at Rice University in the discovery of buckminsterfullerene (C60). During arc discharge between two closely spaced graphite rods, carbon vaporizes and after condensing yields a "sooty" mass. When Iijima looked at the soot with an SEM, he noticed hollow nanotubes of carbon. One may speculate that this finding, like that of C60 "buckyballs," could have been achieved earlier if better microscopy techniques had been available to see the nano-sized products hidden in the soot.

During the past decade, there has been major progress reported in the chemical synthesis (i.e., bottom-up manufacture) of all types of nanoscale semiconductor wires. As originally proposed by R.S. Wagner and W.C. Ellis, from Bell Labs, for the Au-catalyzed Si whisker growth, a vapor-liquid-solid mechanism is still mostly used.[5] The field got a shot in the arm (a rebirth, so to speak) with efforts by Charles Lieber (Harvard) (Figure 3.65), Peidong Yang (http://www.cchem.berkeley.edu/pdygrp/main.html), James Heath (http://www.its.caltech.edu/~heathgrp), and Hongkun Park (http://www.people.fas.harvard.edu/~hpark). Lieber's group at Harvard (http://cmliris.harvard.edu) reported arranging indium phosphide semiconducting nanowires into a simple configuration that resembled the lines in a tick-tack-toe board.

The team used electron beam lithography to place electrical contacts at the ends of the nanowires to show that the array was electronically active. The tiny arrangement is not yet a circuit, but it is a first step, showing that separate nanowires can be contacted to one another.

Infinitely Deep, Finite-Sized 3D Potential Wells—Quantum Dots The solution of Schrödinger's equation for electrons in a cube of metal with side L is given as:

$$\Psi_k(\mathbf{r}) = V^{-\frac{1}{2}}\exp(i\mathbf{k}\cdot\mathbf{r}) \qquad (3.159)$$

—a 3D generalization of Equation 3.157 (with $V = L^3$). The wave vector $\mathbf{k}$, if you recall, points in the direction of wave propagation and has a magnitude given by $2\pi/\lambda$ for a plane wave. Generally, if a wave propagates along a displacement vector $\mathbf{r}$, the

FIGURE 3.65 Harvard's Charles Lieber reinvigorated the nanowire field.

amount of phase accumulated by the wave is given by $\mathbf{k}\cdot\mathbf{r}$. From the generalization of Equation 3.158, electrons in a cubic box of side L ($L_1 = L_2 = L_3$) have allowed energy levels specified by three quantum numbers n_1, n_2, and n_3, or:

$$E_{n_1, n_2, n_3} = \frac{h^2}{8m}\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2} + \frac{n_3^2}{L_3^2}\right)$$

$$= \frac{h^2}{8mL^2}(n_1^2 + n_2^2 + n_3^2) \qquad (3.160)$$

and they come with the following allowed wave vectors:

$$\mathbf{k}_x = \frac{2\pi n_x}{L}, \mathbf{k}_y = \frac{2\pi n_y}{L}, \mathbf{k}_z = \frac{2\pi n_z}{L},$$

$$n_x, n_y, n_z = 1, 2, 3, \ldots \text{(integers)} \qquad (3.161)$$

The allowed $\mathbf{k}$-states are uniformly spaced along the axes with one state per length $2\pi/L$. This is shown for a 2D square array and a 3D cubic array in Figure 3.66. One consequence of confining a quantum particle in a cubic 3D is again "degeneracy." As we

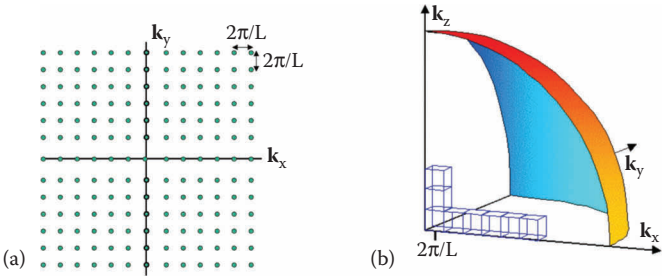


FIGURE 3.66 A 2D array of allowed $\mathbf{k}$-state where we are looking in "$\mathbf{k}$-space" or "reciprocal space" (so-called because $\mathbf{k}$ has units of $1/L$) (a) and a cubic array of allowed states in 3D $\mathbf{k}$-space; in three dimensions, we have states described by $\mathbf{k} = (\mathbf{k}_x; \mathbf{k}_y; \mathbf{k}_z)$ (b). The area per point is $(2\pi/L)^2$, and the volume per point is $(2\pi/L)^3$.

saw for the 2D case, degeneracy reflects an underlying symmetry in the $V(x,y,z)$ potential and can be removed if certain symmetry is broken (e.g., by making $L_1 \neq L_2 \neq L_3$). Each dot in Figure 3.66 represents an allowed **k**-state. An important feature here is that the larger the box, the closer the spacing in energy of single particle states, and vice versa. The **k**-states are discrete, but for a normal-sized conductor there are $\sim 10^{26}$ states, so we can treat them as a continuum.

From Figure 3.66a, each value of **k** occupies an area $A = (2\pi/L)^2$, and from Figure 3.66b each point occupies a volume $V = (2\pi/L)^3$. The number of states per unit volume of **k**-space is $1/V$ (or $L/2\pi)^3$.

When L becomes very small, of the order of the de Broglie wavelength of an electron in all three directions, the electrons lose all capacity to move (3D confinement). These 0D structures, with 3D quantum confinement, are called quantum dots or QDs, as illustrated in Figure 3.67. Quantum dots, also known as nanocrystal semiconductors, ranging from 2 to 10 nm (10–50 atoms) in diameter, are typically composed of materials from periodic groups II–VI, III–V, or IV–VI (e.g., CdS, CdSe, PbS, PbSe, PbTd, CuCl…). The trapped electrons in these dots behave as de Broglie standing waves. The confinement of the waves leads to a blue energy shift, and by varying the particle size one can produce any color in the visible spectrum, say from deep (almost infra-) reds to screaming (almost ultra-) violets as illustrated in Figure 1.31. A quantum dot (QD) is an atom-like state of matter sometimes referred to as an "artificial atom." What is so interesting about a QD is that electrons trapped in them arrange themselves as if they were part of an atom, although there is no nucleus for the electrons to surround here. The type of atom the dot emulates depends on the number of atoms in the well and the geometry of the potential well $V(x)$ that surrounds them.

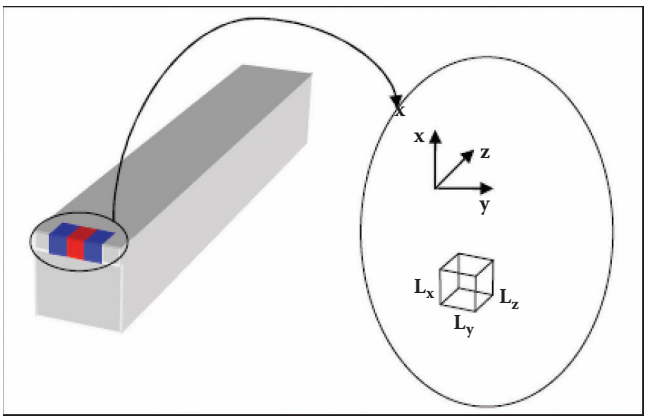


FIGURE 3.67 Quantum dot with dimensions L_x, L_y, and L_z that are finite.[4]

An important consequence of decreasing the dimensionality even further than in quantum wells and wires is that the density of state function for quantum dots features an even sharper and yet more discrete density of states. As a consequence, quantum dot lasers exhibit a yet lower threshold current than lasers based on quantum wires and quantum wells, and because of the more widely separated discrete quantum states, they are also less temperature sensitive. However, because the active lasing material volume is very small in quantum wires and dots, a large array of them has to be made to reach a high enough overall intensity. Making quantum wires and dot arrays with a very narrow size distribution to reduce inhomogeneous broadening remains a real manufacturing challenge, and as a result only quantum well lasers are commercially mature.

However, quantum dots already form an important alternative to organic dye molecules. Unlike fluorescent dyes, which tend to decompose and lose their ability to fluoresce, quantum dots maintain their integrity, withstanding many more cycles of excitation and light emission (they do not bleach as easily!). Combining a number of quantum dots in a bead conjugated to a biomolecule is used as a spectroscopic signature—like a bar code on a commercial product—for tagging those biomolecules (see Chapter 7, "Fluorophores, Quantum Dots, and Fluorescent Proteins").

In the early 1980s, Dr. Ekimov discovered quantum dots with his colleague, Dr. Efros, while working at the Ioffe Institute in St. Petersburg (then Leningrad), Russia.[6,7] This team's discovery of quantum dots occurred at nearly the same time as Dr. Louis E. Brus (Figure 3.68), a physical chemist then working at AT&T Bell Labs (he is now at Columbia), found out how to grow CdSe nanocrystals in a controlled manner.[8,9] Experimenting with CdSe nanocrystal semiconductor material, Dr. Brus and his collaborators observed solutions of astonishingly different colors made from this same substance. Their observation led to the recognition that there is a very clear transition in material behavior when

FIGURE 3.68 Columbia's Louis Brus.

a chunk of it becomes smaller than a fundamental scale, intrinsic to the substance.

Even though it was predicted in 1982 that QDs could be used as the active region of lasers, providing reduced threshold current and lower temperature dependence, it took nearly a decade to develop reliable growth techniques to produce QDs of a quality suitable for commercial applications. Major contributions to making QDs a reality were made by two Bell Labs scientists, Dr. Moungi Bawendi and Dr. Paul Alivisatos. They have since moved to MIT and University of California, Berkeley, respectively, where they continue their investigation of optical properties of quantum dots. Before 1993, QDs were prepared in aqueous solution with added stabilizing agents to avoid colloid precipitation (see colloid stability in Chapter 7, p. 518 "Intermolecular Forces"). In 1993, Bawendi and coworkers synthesized better luminescent CdSe QDs by using a high-temperature organometallic procedure instead.[10,11] Crystallites from 12 to –115 Å in diameter with consistent crystal structure, surface derivatization, and a high degree of monodispersity were prepared in a single reaction based on the pyrolysis of organometallic reagents by injection into a hot coordinating solvent. At that time, even if one could make QDs in a narrow size range, they still came with two major problems: 1) poor fluorescence and 2) hydrophobicity, making them useless in biology. The addition of semiconductor caps, such as ZnS or CdS caps over a CdSe core, was found to dramatically increase the fluorescence quantum yield to 45% or higher.[12–14] The core determines the nanocrystal color, and the shell of the higher bandgap material (ZnS or CdS) dramatically enhances not only the brightness but also the chemical stability. A few different methods for making nanocrystals water soluble are also available today.[15] Methods for water solubilization of QDs include derivatizing the surface with mercaptoacetic acid or dithiothreitol, and Alivisatos and coworkers use a silica/siloxane coating.[16,17] The addition of these coats makes the particles water soluble, and therefore more useful in biology experiments.

The quest to find new energy solutions stimulated the development of modern quantum dot technology even further. The advantage of the surface area-to-volume ratio of nanocrystal particles for energy conversion was realized, and photoelectrochemistry research (e.g., solar energy conversion) tapped the semiconductor/liquid interface to exploit this (also visit http://www.technologyreview.com/Energy/19256).[18,19]

Quantum dots, we will learn in Chapter 7 and Volume III, Chapter 3, are manufacturable through a very wide variety of methods, including self-assembly in colloidal wet chemical synthesis, template chemistry (zeolite, alumina templates), sol-gel methods, micelle methods, organometallic synthesis, pyrolysis, lithography and etching, electrostatic confinement, scanning tunneling microscope (STM) tips, epitaxial strain, and so on. QD devices have now been demonstrated in many research laboratories, and commercial products are available on the market.

Zero Point Energy We noted before that at the lowest electron energy (n = 1), the *ground state energy,* remains finite despite the fact that V = 0 inside the solid (see Equation 3.156 and Figure 3.58c). In classical physics, both the kinetic energy (KE) and potential energy (PE) can have a zero value, but not in quantum physics! According to quantum mechanics, an electron in a box of length L must necessarily have energy. It cannot be inside the box and have zero energy. The minimum possible energy is E_1 and is called the zero point energy (ZPE). Zero point energy hypothesized by Max Planck in 1911 and developed by him and Walter Nernst between 1911 and 1916. It was first measured by Dr. Willis Lamb in 1947 as a slight upward shift of electrons in their atomic orbitals. The concept of a zero point energy has very far-reaching consequences, manifesting

themselves from the quantum world to the cosmos. The existence of a zero point energy for a quantum particle is a general phenomenon and is consistent with the Heisenberg uncertainty principle; i.e., particles with zero momentum cannot be localized. From the Heisenberg uncertainty principle (Equation 3.106), it follows that for a 1D box:

$$E = \frac{p^2}{2m} \text{ and also}$$

$$\frac{(\Delta p)^2}{2m} \geq \frac{\hbar^2}{2m(\Delta x)^2} \geq \frac{\hbar^2}{2m\left(\frac{L}{2}\right)^2} = \frac{2\hbar^2}{mL^2} > 0 \quad (3.162)$$

where we used $\Delta x \leq \frac{L}{2}$, and also for E_1 (see Equation 3.156) we obtain:

$$E_1 = \frac{k_x^2 \hbar^2}{2m} = \frac{4\pi^2\hbar^2}{2mL^2} > \frac{2\hbar^2}{mL^2} > 0 \quad (3.163)$$

because $k_x = 2\pi/L$ (see Equation 3.161).

That the zero point energy is a direct consequence of the Heisenberg uncertainty principle can simply be recognized by considering an electronic vibration ceasing at T = 0. If such a thing could happen, then position and momentum both would be zero, which would violate the Heisenberg uncertainty principle.

The zero point energy can be measured through the Casimir force, a small attractive force between two close parallel, uncharged plates in a vacuum (see Volume III, Chapter 8 on actuators). This force, inversely proportional to the fourth power of the distance between the plates, comes about because, according to quantum physics, even a perfect vacuum contains a zero point energy keeping virtual particles in a continuous state of fluctuation (see Equation 3.107). Given the equivalence of mass and energy expressed by Einstein's $E = mc^2$, the vacuum energy must be able to create particles, known as virtual particles. They flash briefly into existence and expire within an interval dictated by the uncertainty principle. Casimir realized that only the virtual photons with wavelengths that fit a whole number of times into the gap between the two plates should be part of the energy calculation (i.e., no half-quantum lengths are allowed). Moving the plates together, the energy density should thus decrease as there are more allowed virtual photon states pushing against the plates from the outside than from between the two plates (see Figure 3.69).

An example of a practical application of the existence of the zero point energy for confined electrons is the quantum well laser, which is more efficient than a diode laser. In a quantum well, as we saw above, there are no allowed electron states at the very lowest energies, but there are more available states in the lowest conduction band state and at the top of the valence band so that many more holes and electrons can combine and produce photons with identical energy for enhanced probability of stimulated emission (lasing; see also Chapter 5).

Potential Wells of Finite Depth and Size, Finite Barriers, Tunneling, and Interfaces

Finite-Depth Potential Wells A potential well is a potential energy function with a minimum. An infinitely deep potential well, as considered above, is an idealization. On the atomic scale there are no infinitely high and sharp barriers, and both $\psi(x)$ and $d\psi(x)/dx$ go to zero smoothly near a boundary. If the walls of an electron *prison* are not infinitely high, as sketched in Figure 3.56, but can be scaled by the particles inside as sketched in Figure 3.70, Equation 3.137 must be rewritten as:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} = (E - V_0)\psi(x) \quad (3.164)$$

with V_0 the potential barrier the particle must jump to get out of the well. We will now solve this equation in the three regions marked in Figure 3.70.

A quantum well with finite potential walls for trapped particles is a more realistic picture than the infinite potential wells assumed earlier and applies

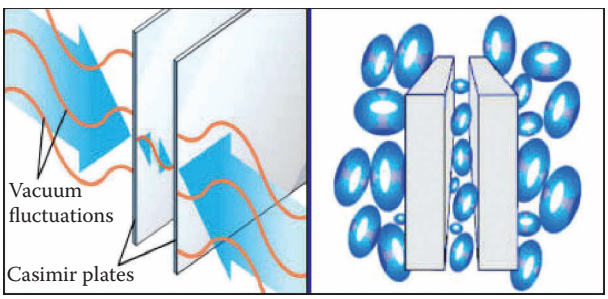


FIGURE 3.69 Vacuum fluctuations. The Casimir effect.

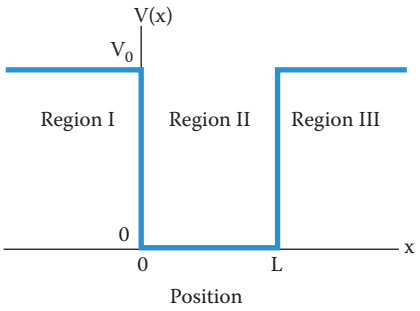


FIGURE 3.70 Quantum well with finite walls.

to most practical 3D (quantum dots), 2D (quantum wires), and 1D (quantum wells) confined structures. The representation also applies to nucleons inside the nucleus of an atom. Even the Krönig-Penney model, used to calculate the potential energy of an electron in a row of atoms in a linear solid, uses an array of periodic square wells of the type shown in Figure 3.70. In this model, each cell with V(x) = 0 represents an atom.

Outside the box sketched in Figure 3.70, in regions I and III, the boundary condition is that $V(x) = V_0$. These are regions that are "forbidden" to classical particles with $E < V_0$. With $E < V_0$ a classical particle cannot penetrate a barrier region: think about a particle hitting a metal foil and only penetrating the foil if its initial energy is greater than the potential energy it would possess when embedded in the foil and where otherwise it will be reflected. Defining α as:

$$\alpha^2 = \frac{2m(V_0 - E)}{\hbar^2} \tag{3.165}$$

yields:

$$\frac{d^2\psi}{dx^2} = \alpha^2\psi \tag{3.166}$$

In a region with $E < V_0$ there is an immediate effect on the waveform for the particle because $\mathbf{k}_x$ is real under these conditions, and we can write:

$$\mathbf{k}_x = \alpha = \frac{\sqrt{2m(V_0 - E)}}{\hbar} \tag{3.167}$$

We find for the general solution of Equation 3.164 in regions I and III a wave function of the form $\psi(x) = Ae^{+\alpha x} + Be^{-\alpha x}$, i.e., a mixture of an increasing and a decreasing exponential function. With a barrier that is infinitely thick we can see that the increasing exponential must be ruled out as it conflicts with the Born interpretation because it would imply an infinite amplitude. Therefore, in a barrier region the wave function must simply be the decaying exponential $Be^{-\alpha x}$, the important point being that a particle may be found inside a classically forbidden region (regions I and III).

If the barrier thickness is not infinite, then the increasing exponential component in the wave function cannot be ruled out because the wave function amplitude may not necessarily rise to infinity before the potential drops back to zero. In this case, the solutions for regions I and III are the following damped exponentials:

$$\psi_I(x) = Ae^{+\alpha x} \text{ (region I, } x < 0) \tag{3.168}$$

$$\psi_{III}(x) = Be^{-\alpha x} \text{ (region III, } x > L) \tag{3.169}$$

The values for *A* and *B* of the damped exponentials in Equations 3.168 and 3.169 are determined from the boundary conditions. These boundary conditions include the fact that wave function ψ and its derivative $d\psi/dx$ must be continuous at boundaries between regions I, II, and III (the boundary conditions require that $\psi_I = \psi_{II}$ at $x = 0$ and $\psi_{II} = \psi_{III}$ at $x = L$).

In Figure 3.71, the "leaky" waves in the forbidden regions are shown as exponential tails. The penetration depth of these waves is the distance outside the potential well over which the probability significantly decreases and is given by:

$$\delta x \approx \frac{1}{\alpha} = \frac{\hbar}{\sqrt{2m(V_0 - E)}} \tag{3.170}$$

Thus, the penetration distance that violates classical physics is proportional to Planck's constant and also depends on the value of $V_0 - E$ and on the mass of the particle. Because of this "barrier penetration," the electron density of a material extends outside the surface of the material.

In region II ($0 < x < L$) where V(x) = 0, the wave equation becomes:

$$\frac{d^2\psi}{dx^2} = -\mathbf{k}_x^2\psi \tag{3.171}$$

where:

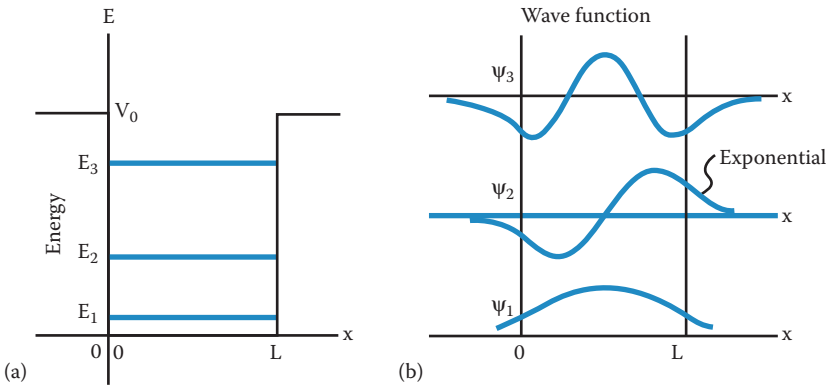


FIGURE 3.71 (a) Energies of a wave function inside the quantum well (region II). (b) The wave functions leak out of the well: see exponential tails.

$$\mathbf{k}_x = i\alpha \text{ and } \alpha = \frac{\sqrt{2mE}}{\hbar} \tag{3.172}$$

The solution here is an oscillating wave just as in the case of the well with infinite walls (Figure 3.56):

$$\psi_{II} = Ce^{i\mathbf{k}_x x} + De^{-i\mathbf{k}_x x}$$

or (3.173)

$$\psi_{II}(x) = \psi_0 \sin \mathbf{k}_x x + \psi_0' \cos \mathbf{k}_x x$$

Thus, the wave functions for a particle in a well with finite walls look very similar to the ones for the infinite square well … except that the particle now has a finite probability of "leaking out" of the well!

Finite Height Barriers (Step Functions)

(1) $E > V_0$ A potential barrier is the opposite of a potential well. It is a potential energy function with a maximum. For a barrier of finite height and thickness (Figure 3.72) and with $E > V_0$, we use again the TISE (Equation 3.137). For particles outside and above the barrier (regions I and III), V = 0 and we obtain:

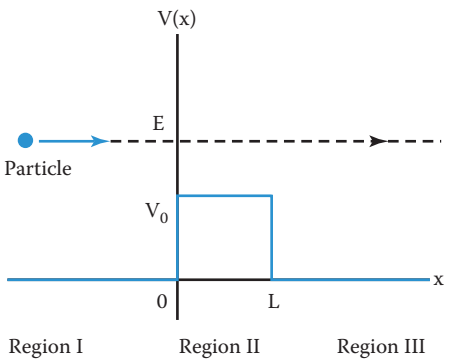


FIGURE 3.72 Finite barrier with boundary conditions.

$$\begin{aligned}\psi_I(x) &= Ae^{+ik_I x} + Be^{-ik_I x} (x < 0 \text{ and } V = 0)\\ \psi_{III}(x) &= Ee^{+ik_{III} x} + Fe^{-ik_{III} x} (x > L \text{ and } V = 0)\end{aligned} \tag{3.174}$$

These are oscillations with the wave vector:

$$\mathbf{k}_I = \mathbf{k}_{III} = \sqrt{\frac{2mE}{\hbar^2}} \tag{3.175}$$

We have replaced $\mathbf{k}_x$ here with the symbol $\mathbf{k}_I = \mathbf{k}_{III}$ or $\mathbf{k}_{II}$ to be more specific about the regime under consideration.

It is important to recognize here that with $E > V_0$ a particle would, in a classical picture, easily overcome the barrier, and one would expect a 100 % transmission. We will see that this is not the case when the particle is "wavy." In the barrier region (region II), where $0 < x < L$, we calculate:

$$\psi_{II}(x) = Ce^{+ik_{II} x} + De^{-ik_{II} x} \tag{3.176}$$

with the wave vector:

$$k_{II} = \frac{\sqrt{2m(E - V_0)}}{\hbar} \tag{3.177}$$

whereas in the case of $E < V_0$, $\mathbf{k}_x$ is real in a barrier region (see Equation 3.167), in the case of $E > V_0$, $\mathbf{k}_x$ (=$\mathbf{k}_{II}$) remains imaginary. If we are only considering waves moving from left to right (Figure 3.73), we can simplify the above wave functions to:

$$\text{Incident wave: } \psi_I(x) = Ae^{+ik_I x} \tag{3.178}$$

$$\text{Reflected wave: } \psi_I(x) = Be^{-ik_I x} \tag{3.179}$$

$$\text{Transmitted wave: } \psi_{III}(x) = Ee^{+ik_I x} \tag{3.180}$$

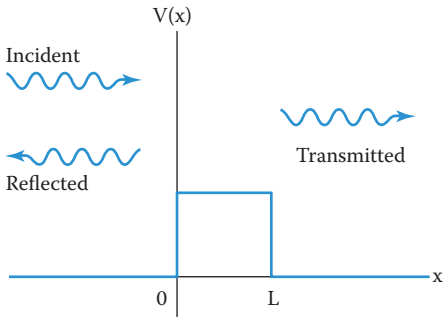


FIGURE 3.73 Wave with $E > V_0$ moving from left to right.

We define a "reflection coefficient" R as:

$$R = \frac{|B|^2}{|A|^2} \tag{3.181}$$

and likewise we can define a transmission coefficient T as:

$$T = \frac{|E|^2}{|A|^2} \tag{3.182}$$

The probability of the particles being reflected R or transmitted T is then:

$$R = \frac{|\psi_I(\text{reflected})|^2}{|\psi_I(\text{incident})|^2} = \frac{|B|^2}{|A|^2}$$

and:

$$T = \frac{|\psi_{III}(\text{transmitted})|^2}{|\psi_I(\text{incident})|^2} = \frac{|E|^2}{|A|^2} \tag{3.183}$$

The transmission probability is the probability that a particle incident on the left of the barrier emerges on the right of it.

(2) $E < V_0$ The situation where classically the particle does not have enough energy to surmount the potential barrier, $E < V_0$, is sketched in Figure 3.74.

With $E < V_0$, the particle will be reflected at x = 0 with the same kinetic energy and Equation 3.176 again applies, but the transmitted wave is now a damped exponential as we saw above. The wave function in region II becomes:

$$\psi_{II}(x) = \underset{\text{Unphysical} = 0}{Ce^{+\alpha x}} + \underset{\text{Damped}}{De^{-\alpha x}} \tag{3.184}$$

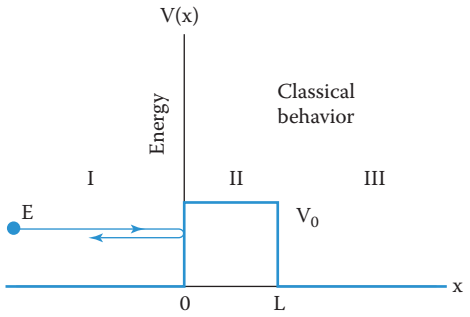


FIGURE 3.74 Wave with $E < V_0$ moving from left to right.

with $k_{II} = \alpha = \frac{\sqrt{2m(V_0 - E)}}{\hbar}$ (see Equation 3.167).

Tunneling We briefly return here to the tunneling phenomenon as described in Equation 3.167. The violation of classical physics in tunneling as described by this equation is allowed by the uncertainty principle. A particle can violate classical physics by ΔE for a short time, $\Delta t \sim \hbar/\Delta E$ (see Equation 3.107). The tunneling wave function is shown in Figure 3.75. The exponential decay of the wave function inside the barrier is given as:

$$\psi(x) = Ae^{-\alpha x} \tag{3.185}$$

with $\alpha^2 = \frac{2m(V_0 - E)}{\hbar}$ (Equation 3.165). If the barrier is narrow enough (L in Figure 3.75 is small), there will be a finite probability P of finding the particle on the other side of the barrier. The probability of an electron reaching across barrier L is:

$$p = |\psi(x)|^2 = A^2e^{-2\alpha L} \tag{3.186}$$

where A is a function of energy E and barrier height V_0. The probability of finding an electron

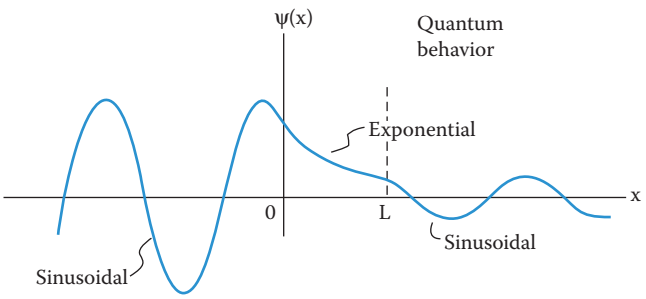


FIGURE 3.75 Tunneling wave function. The rectangular barrier stretches from x = 0 to x = L, the height of the barrier is V_0.

on the other side of a barrier of width L can be probed with a fine needle tip from a scanning tunneling microscope (STM) (see also Volume III, Chapter 6 on metrology). The tunneling current, based on Equation 3.186, picked up by the sharp needle point is given by:

$$I = f_w(E)A^2e^{-2\alpha L} \tag{3.187}$$

where $f_w(E)$ is the Fermi-Dirac function, which contains a weighted local density of electronic states in the solid surface that is being probed and states in the needle point (see also Fermi's golden rule further below). The weighted local density of electronic states is a material property of both probed surface and probe and may be obtained by measurements of the current I as a function of bias voltage V (dI/dV), which gives spatial and spectroscopic information about the quantum states of a nanostructure. From Equation 3.187, when L changes by 1 Å, the current changes by a factor of about 10! Obviously, the current is very sensitive to the gap distance. The size of the gap in practice is on the order of a couple of Angstroms! If the tip has two atoms vying for sitting at the very apex of the tungsten tip, the atom recessed by two atoms lower than the winning atom carries about 1 million times less current. That is why one wants such a fine tip.

Because particles must be either reflected (R) or transmitted (T) we have $R + T = 1$. By applying the boundary conditions $x \to \pm\infty$, $x = 0$, and $x = L$, we may also calculate the theoretical transmission or tunneling probability P from:

$$P = \frac{1}{1+G} \text{ with } G = \frac{(e^{\alpha L} - e^{-\alpha L})^2}{4\left(\frac{E}{V_0}\right)\left(1 - \frac{E}{V_0}\right)} \tag{3.188}$$

We note that P can be nonzero; i.e., particles may tunnel through a barrier even when $E < V_0$. The wave function does not fall abruptly to zero inside a region where its potential energy exceeds its total energy. This quantum mechanical result is a most remarkable feature of modern physics: there is a finite probability that the particle can penetrate the barrier and even emerge on the other side! We also notice here that even when $E > V_0$, $P < 1$. So a particle that has enough energy to overcome a barrier has a high probability to be reflected instead. This is counterintuitive as a classical particle with that energy would have a $P = 1$. We can summarize this situation as follows: quantum mechanics predicts an enhanced tunneling when $E < V_0$ and an enhanced reflection when $E > V_0$. This is equivalent to light reflecting from an interface with an abrupt change of refractive index.

Example 3.3: Assume that the work function (i.e., the energy difference between the most energetic conduction electrons and the potential barrier at the surface) of a certain metal is $\Phi = 5$ eV. Estimate the distance x outside the surface of the metal at which the electron probability density decreases to 1/1000 of that just inside the metal.

$$\frac{|\Psi(x)|^2}{|\Psi(0)|^2} = e^{-2kx} \approx \frac{1}{1000}$$

This leads to:

$$x = -\frac{1}{2k}\ln\left(\frac{1}{1000}\right) \approx 0.3 \text{ nm}$$

$$\text{Using } k = \sqrt{\frac{2m_e}{\hbar^2}(V_0 - E)} = 2\pi\sqrt{\frac{2m_e}{h^2}\Phi}$$

$$= 2\pi\sqrt{\frac{5\text{ eV}}{1.505\text{ eV}\cdot\text{nm}^2}} = 11.5\,\text{nm}^{-1}$$

with $m = m_e$.

In this section we considered the simplified problem of tunneling through a square barrier, but in most cases the barriers are not simply square shaped. One then needs to obtain a more general expression for the tunneling probability. These calculations are fairly involved, and we refer the reader to the specialized literature (e.g., Wolf's *Principles of Electron Tunneling Spectroscopy*[20]).

Some Tunneling History The concept of tunneling has no analogy in classical mechanics. The experimental manifestations of this effect are one of the many triumphs of the quantum theory. Based on electron tunneling, Fowler and Nordheim, in 1928, explained the main features of electron emission from cold metals by high external electric fields,

which had been unexplained since its first observation by Lilienfeld in 1922.

The discovery of the Esaki tunnel diode (see Figure 4.44) was very significant in the history of tunneling as it was the first electronic device where electron tunneling was clearly manifested in a semiconductor. Esaki described the tunnel diode in his 1957 thesis and received the 1973 Nobel Prize for his invention (http://nobelprize.org/nobel_prizes/physics/laureates/1973/esaki-bio.html). The next significant event was early in the 1970s, when the first quantum wells (QWs), which were also the first low-dimensional heterostructures, were demonstrated (see the AlGaAs/GaAs/AlGaAs structure in Figure 3.59).

Esaki was not only the originator of the Esaki tunnel diode; he also invented the double-barrier resonant tunneling diodes (abbreviated DBRT diode; see Figure 5.149) in 1974. Furthermore, in 1969, Esaki* and Tsu initiated research on semiconductor superlattices based on a periodic structure of alternating layers of semiconductor materials with wide and narrow bandgaps, in other words, a series of quantum wells or multiquantum well devices (MQWs).[21] The first superlattices were fabricated using an AlGaAs/GaAs material system (Figure 5.148).

In 1981, Gerd Binnig and Heinrich Rohrer invented the scanning tunneling microscope (STM), enabling the visualization and moving of individual atoms for the first time.

Tunneling is also of great importance in the history of nuclear physics. The decay of a nucleus is the escape of particles bound inside a barrier. The phenomenon of tunneling explains α-particle decay of heavy, radioactive nuclei. Inside the nucleus, an α-particle feels the strong, short-range attractive nuclear force as well as the repulsive Coulomb force. The nuclear force dominates inside the nuclear radius where the potential can be approximated by a square well. The rate for escape can be very small; particles in the nucleus "attempt to escape" 10^{20} times per second but may succeed in escaping only once in many years! Even if the quantum state (wave function) of the nucleus is completely defined with no uncertainty, one cannot predict when a nucleus will decay (Figure 3.76). Quantum mechanics tells us only the probability per unit time that any nucleus will decay (Figure 3.76a). Electrons, we saw above, are bound with negative total energy and can never escape the nucleus (Figure 3.76b). But a nucleon, such as an α-particle, is held inside the nucleus by the strong nuclear force and can escape by tunneling through the positive Coulomb barrier, which leads to the nuclear decay processes (G. Gamov, 1928). The potential barrier at the nuclear radius is several times greater than the energy of an α-particle. Based on Equation 3.107, for a short time, a particle can "borrow" energy from the uncertainty relation, gaining enough energy to jump over the potential barrier before giving it back. When it returns to its "proper" energy state, it is just outside the barrier instead of just inside, and rushes away. The process is as if the particle tunneled through the barrier, and it is purely a quantum effect.

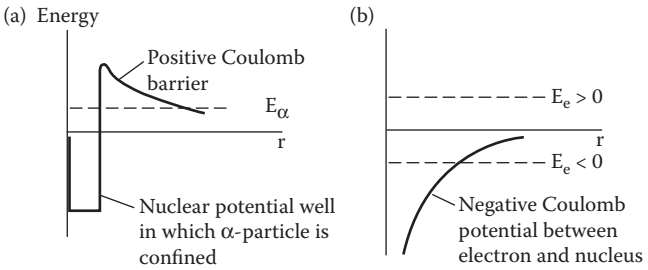


FIGURE 3.76 Nucleons escape the nucleus in radioactive decay (a); electrons cannot escape the nucleus (b).

Interfaces The theoretical treatment of particles hitting an interface is the same as the treatment of particles hitting one side of a square barrier (half square or half step function) and finds all types of applications. The results allow one to handle important problems involving the transmission and reflection of particle waves, such as encountered in vacuum/metal interfaces (work function), vacuum/semiconductor interfaces, semiconductor/metal interfaces, metal/metal interfaces (e.g., in thermocouples), and semiconductor/semiconductor interfaces (e.g., in diodes). For this analysis no new equations are needed as we can retrieve all the necessary expressions from our treatment of the full square barrier as summarized in Table 3.7.

* Esaki worked for IBM at the Thomas J. Watson Research Center, Yorktown Heights, New York, until 1992, when he returned to Japan to become president of Tsukuba University, Ibaraki.

TABLE 3.7 Transmitted and Reflected Particle Waves on an Interface

Energy	Equations	Wave Vectors
$E = \frac{\langle p^2 \rangle}{2m} > V_0$	Incident $\psi_I(x) = Ae^{+ik_I x} + Be^{-ik_I x}$ Reflected Transmitted $\psi_{II}(x) = Ce^{+ik_{II} x} + De^{-ik_{II} x}$ Unphysical = 0	$k_I = \frac{\sqrt{2mE}}{\hbar}$ $k_{II} = \frac{\sqrt{2m(E - V_0)}}{\hbar}$
$E = \frac{\langle p^2 \rangle}{2m} < V_0$	Incident $\psi_I(x) = Ae^{+ik_I x} + Be^{-ik_I x}$ Reflected Unphysical = 0 $\psi_{II}(x) = Ce^{+\alpha x} + De^{-\alpha x}$ Damped	$k_I = \frac{\sqrt{2mE}}{\hbar}$ $\alpha = \frac{\sqrt{2m(V_0 - E)}}{\hbar}$

Harmonic Potential Wells

We now revisit the harmonic oscillator from Figure 3.51 and solve it in the Schrödinger way so we may learn how quantization comes about in the case of an oscillating system. Simple harmonic oscillators describe many physical situations from springs to diatomic molecules and atomic lattices. In all cases, harmonic oscillations occur because the system contains a part that experiences a restoring force (spring) proportional to the displacement from equilibrium. In Figure 3.77, depicting a harmonic potential well, we discern two regions: inside the well (region I), with $E > V(x)$, and outside of the well (region II), with $E < V(x)$.

The time-independent Schrödinger equation (TISE) for a harmonic oscillator is given as:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} = \left(E - \frac{1}{2}kx^2\right)\psi(x) \qquad (3.189)$$

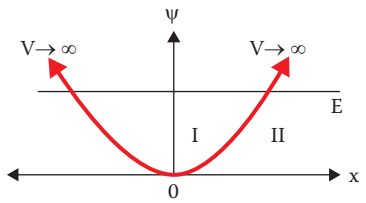


FIGURE 3.77 Harmonic potential well.

where $V(x) = \frac{1}{2}kx^2$ (in this equation k is the force constant, not the wave number **k**!) and $\omega = \left(\frac{k}{m}\right)^{\frac{1}{2}}$. The general solutions for Equation 3.189 are bell-shaped Gaussian functions multiplied by a Hermite polynomial $H_n(y)$:

$$\psi_n(x) = N_n H_n(\alpha^{\frac{1}{2}}x)e^{-\frac{\alpha x^2}{2}}$$

$$\text{with } n = 0,1,2,3\ldots \text{ and } \alpha = \left(\frac{K_m}{\hbar^2}\right)^{\frac{1}{2}}$$

$$N_n(\text{a normalization constant}) = \frac{1}{(2^n n!)^{\frac{1}{2}}}\left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} \qquad (3.190)$$

Some values for $H_n(y)$ are listed in Table 3.8.

The energy is given by:

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega, \; n = 0,1,2,3\ldots \qquad (3.191)$$

with n the vibrational quantum number. As shown in Figure 3.78, energies are evenly spaced and again cannot be zero in the ground state because of quantum confinement. Here we start the quantum numbers from n = 0 rather than n = 1.

The ground state (n = 0) of the wave inside the box is a simple Gaussian with no zero crossings (Figure 3.79a):

$$\psi_{n=0}(x) = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} e^{-\frac{\alpha x^2}{2}} \qquad (3.192)$$

with zero point energy:

$$E_0 = \frac{1}{2}\hbar\omega \qquad (3.193)$$

TABLE 3.8 Hermite Polynomials, $H_n(y)$

n	$H_n(y)$
0	1
1	2y
2	$4y^2 - 2$
3	$8y^3 - 12y$
4	$16y^4 - 48y^2 + 12$
5	$32y^5 - 160y^3 + 120y$

n = vibrational quantum number.

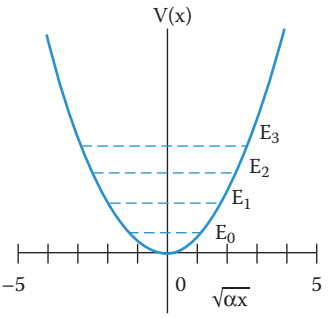


FIGURE 3.78 Energies of the quantum oscillator according to Equation 3.191, $\alpha = \left(\frac{K_m}{\hbar^2}\right)^{\frac{1}{2}}$.

The first excited state (n = 1): $\psi_{n=1}(x) = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} \sqrt{2\alpha}xe^{\frac{-\alpha x^2}{2}}$ has one zero crossing; the second excited state (n = 2): $\psi_{n=2}(x) = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} \frac{1}{\sqrt{2}}(2\alpha x^2 - 1)e^{\frac{-\alpha x^2}{2}}$ has two zero crossings; the third excited state (n = 3): $\psi_{n=3}(x) = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} \frac{1}{\sqrt{3}} \sqrt{\alpha}\, x(2\alpha x^2 - 3)e^{\frac{-\alpha x^2}{2}}$ has three crossings; and so on. From Figure 3.79b the lowest energy state (n = 0) of the QM oscillator has a maximum probability at the equilibrium position, whereas the classical oscillator always has its maximum probability at the extremes.

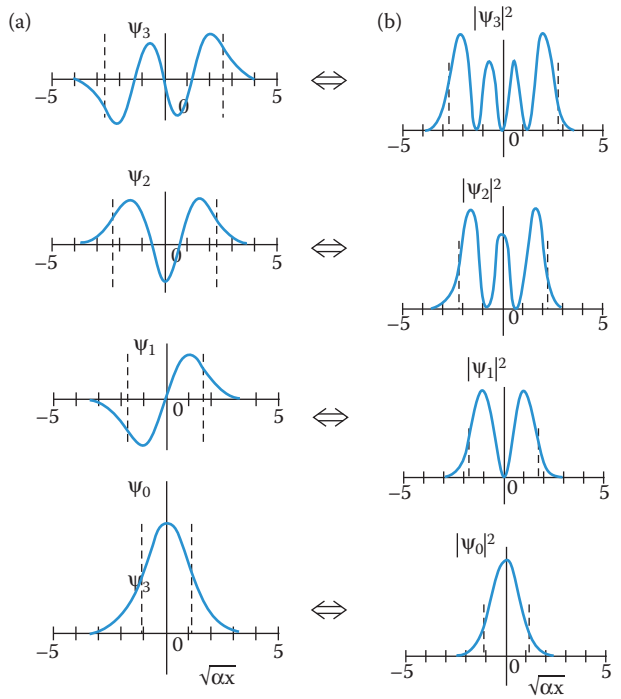


FIGURE 3.79 (a) The waves for n = 0 to n = 3. (b) Probabilities for the same waves.

As expected from Bohr's correspondence principle, the higher the quantum numbers, the better the quantized oscillator resembles the classical nonquantized oscillator from Figure 3.51. This is illustrated in Figure 3.80.

Quantized harmonic oscillators are all around us: diatomic molecules, vibrations within molecules, vibrations of atoms about equilibrium positions, oscillations of atoms or ions in crystal lattices (phonons), normal modes of electromagnetic fields in a cavity (blackbody radiation), Landau levels in the quantum Hall effect, and to a first approximation, any oscillatory behavior. We will encounter another example of the harmonic quantum oscillator solution presented here in the solution Einstein proposed in 1903 for the energies of the "atomic oscillators" in solids (see below under "Classical and Quantum Oscillators").

Example 3.4: The spacings between vibrational levels of molecules in the atmosphere, CO_2 and H_2O, are in the infrared frequency range: $\Delta E = h\nu = \hbar\omega \sim 0.01$ eV. As a consequence, Earth has an atmosphere acting as a greenhouse.

Particles in an Atom: Central Force

An electron bound to the hydrogen nucleus is an example of a central force system: the force depends on the radial distance between the electron and the nucleus only. The potential energy associated with a central or inverse square law force is very important

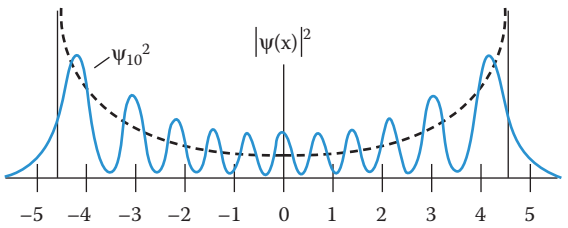


FIGURE 3.80 Quantized oscillator. Probability densities *P* for n = 10 states of a quantum mechanical harmonic oscillator. The probability densities for classical harmonic oscillators with the same energies are shown in black. In the n = 10 state, the wavelength is shortest at x = 0 and longest at x = |L|. The higher the quantum number *n*, the closer the solution resembles the classical one represented in Figure 3.51.

and was illustrated earlier for the case of a hydrogen atom in Figure 3.53. The potential energy of the electron-proton system is:

$$V(r) = -\frac{e^2}{4\pi\varepsilon_0 r} \tag{3.194}$$

The solutions of Schrödinger's equation with this potential are spherical Bessel functions. In Figure 3.53 the probability of finding the ground state for a hydrogen electron (n = 1) as a function of the radial distance from the proton was shown. The value of $|\Psi(x,t)|^2$ at some location at a given time is proportional to the probability of finding the particle at that location at that time. The wave functions for hydrogen are like vibrating strings or membranes, but the vibrations are in three dimensions and they are described by spherical Bessel functions. In Figure 3.81 we show solutions for higher quantum number cases (up to n = 4). In Figure 3.81A we provide a summary, and Figure 3.81B details some of these solutions for hydrogen. The lowest shells are spherical, 1s (n = 1, l = 0, m = 0), 2s (n = 2, l = 0, m = 0), 3s (n = 3, l = 0, m = 0), and 4s (n = 4, l = 0, m = 0), and hold two electrons each. The next shells are three 2p orbitals (n = 2, l = 1 and m = 1, 0, −1), which are dumbbell shaped and fit into the atom along perpendicular axes. A total of six electrons can fit. Level three (n = 3) has one s orbital and three p's—just like the ones at level two—and five three d's (with m = −2, −1, 0, +1, +2). The d orbitals are shaped for the most part like four-leaf clovers. At level four (n = 4) we encounter for the first time the 4f orbitals. They hold 14 electrons and look somewhat like onion blossoms. The electrons in the 4f orbitals are the furthest removed from the nucleus, are easily excited, and exhibit thousands of distinct energy states. They are associated with the many unusual optical, magnetic, and catalytic properties of the rare earth elements.

We come back to the mathematical formulation of the solutions of Schrödinger's equation for a central force when analyzing energy levels for quantum dots. These "artificial atoms" have energy quantization just as like atoms and molecules.

Summary: Most Important Periodic Potential Profiles and the Sommerfeld Model

Summarizing, the quantization for three of the most important potential profiles leads to the following mathematical solutions of Schrödinger's equation: for the central force, we obtain spherical Bessel functions; for an infinite square well potential, sines, cosines, and exponentials; and for an oscillator, Hermite polynomials. The quantized energy levels for each case are summarized in Figure 3.82 (only the 2D case is shown). Notice that the separation between consecutive energy levels increases with increasing *n* for an infinite square well potential (*b*) are evenly spaced for a harmonic well potential (*c*), and in the case of an inverse square law potential (*a*) the separation becomes closer with the larger *n*.

Sommerfeld's model assumes that the electrons in a metal experience a constant zero potential so

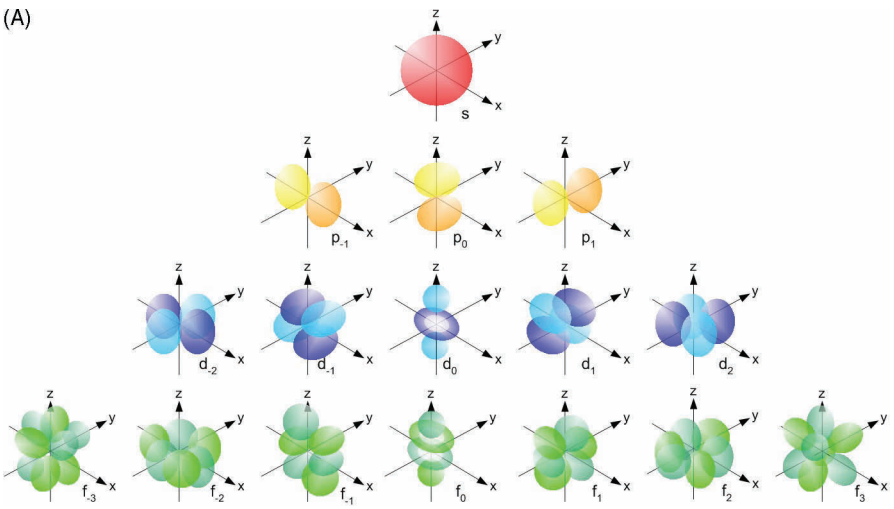


FIGURE 3.81 (A) Summary or overview of atomic orbitals up to n = 4.

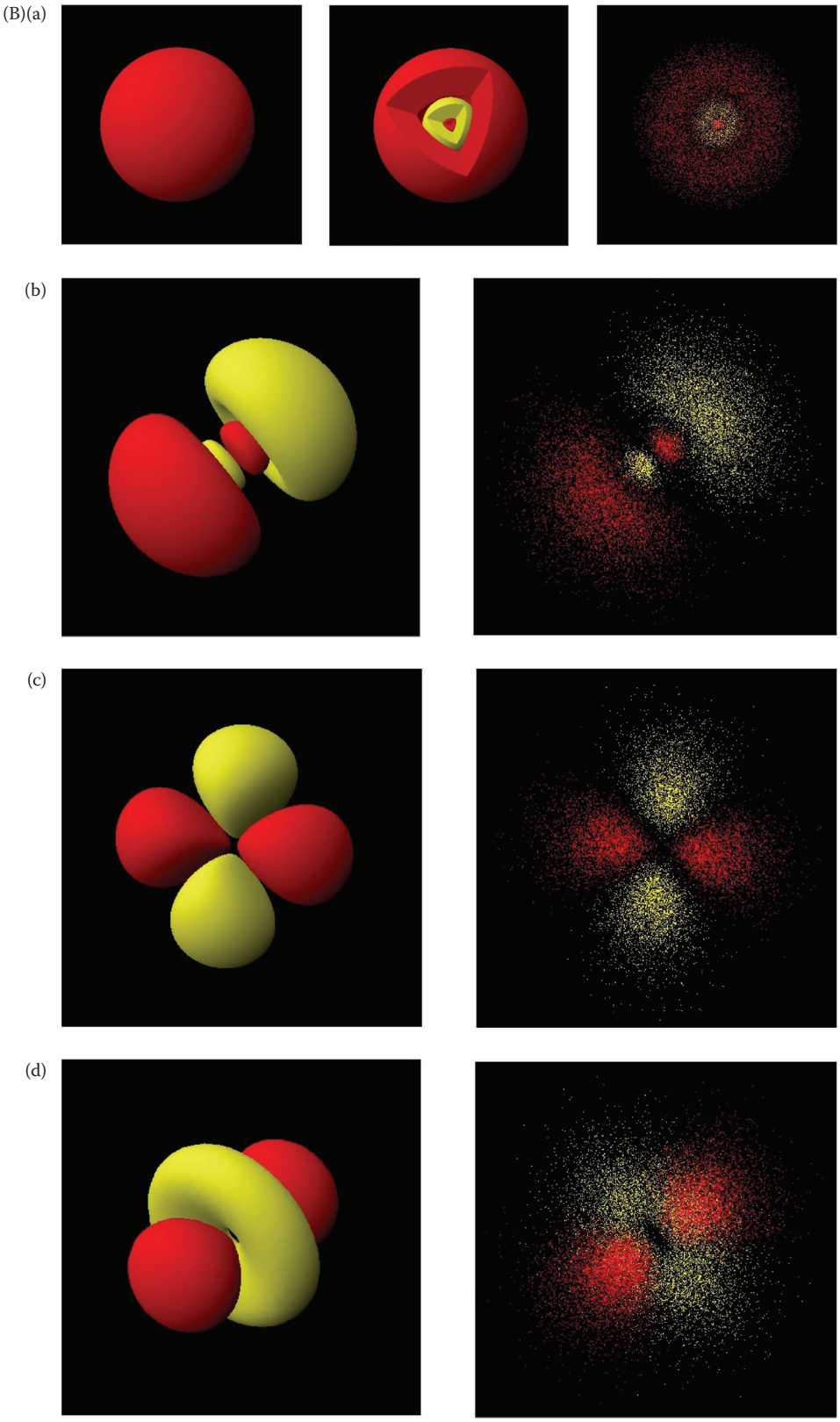


FIGURE 3.81 (*Continued*) (B) Where red = negative phase of Ψ and yellow = positive phase of Ψ. The density of dots reflects the magnitude of Ψ. (a) n = 3, l = 0, m = 0 for hydrogen. (b) n = 3, l = 1, m = –1 for hydrogen. (c) n = 3, l = 2, m = –1 for hydrogen. (d) 4 = 3, l = 3, m = 0 for hydrogen.

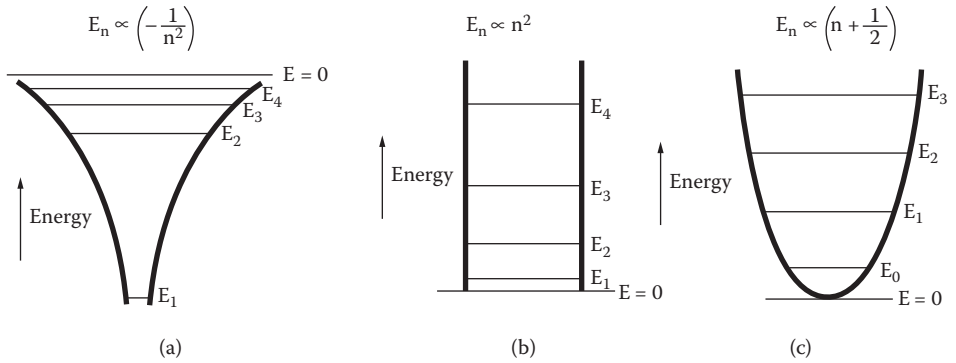


FIGURE 3.82 Quantized energy levels for a particle in an inverse square law potential (a) an infinite square well potential (b) and a harmonic well potential (c) 2D cases.

that the electrons are completely free to move about in the crystal. The important thing Sommerfeld's model does is introduce a finite surface for the metal at L and adopting a Fermi distribution for the charge carriers. What is most different from the classical theory as a result is the notion that only a few electrons, those with energies close to the Fermi level, contribute to the conduction mechanism. The Sommerfeld theory successfully explains specific heat, electrical conductivity, thermionic emission, thermal conductivity, and paramagnetism. However, the model fails to explain why some solids are good conductors, others are semiconductors, and yet others are insulators. The model also cannot account for the fact that some metals such as Be, Zn, and Cd exhibit a positive Hall constant; the free electron model always predicts a negative Hall coefficient! This model further predicts that the electrical conductivity is proportional to the electron concentration, but in reality divalent metals (Be, Cd, and Zn) and even trivalent metals (Al, In) have consistently lower conductivities than monovalent metals (Cu, Ag, and Au). Also, measurements of the Fermi surface, a concept introduced further below, indicate that it is often not spherical, contradicting Sommerfeld's model, which predicts a perfect sphere.

We prepare ourselves now to launch the more realistic Krönig-Penney model, where the potential energy of an electron in a row of atoms in a linear solid is modeled as an array of periodic square wells. We start by learning to combine atoms in simple molecules in the valence bond and molecular orbital theory; we then learn how to work with Bloch functions; and at last we are ready to introduce the band theory of solids.

Bringing Atoms Together

Molecules: Valence Bond and Molecular Orbital Theory

Molecules are formed from atoms by quantum mechanical forces. The so-called covalent bonds have no classical counterparts. They exist only because of the fermionic nature of the valence electrons. In the valence bond (VB) theory, atoms form electron-pair covalent bonds through the overlap of atomic orbitals of adjacent atoms. In the molecular orbital (MO) theory, shared, delocalized, valence electrons are viewed as occupying regions of space extending over all the binding atoms. When two atoms come together to form a molecule a valence bond is formed, and the valence electrons are in orbitals (called molecular orbitals) spread over the entire molecule, i.e., the electrons are delocalized. Half-filled atomic orbitals from the binding partners overlap and form bonds with two electrons of opposite spin occupying the molecular orbital. The latter corresponds to Pauli's exclusion principle, i.e., there are a maximum of two electrons with opposite spin per orbital. A diagram of the potential energy versus the internuclear distance for two approaching hydrogen atoms is shown in Figure 3.83a. The example involves the formation of a sigma bond (σ) from the overlap of s orbitals from the binding hydrogen atoms. At an internuclear distance, r_0, of 74 picometers, equilibrium is reached, and the attraction between the binding partners is maximized with an H-H bond strength of –436 kJ/mol. Hydrogen atoms react to form a molecule because the energy of the system is less than the sum of the individual constituents. From Figure 3.83b and c, we learn that compared

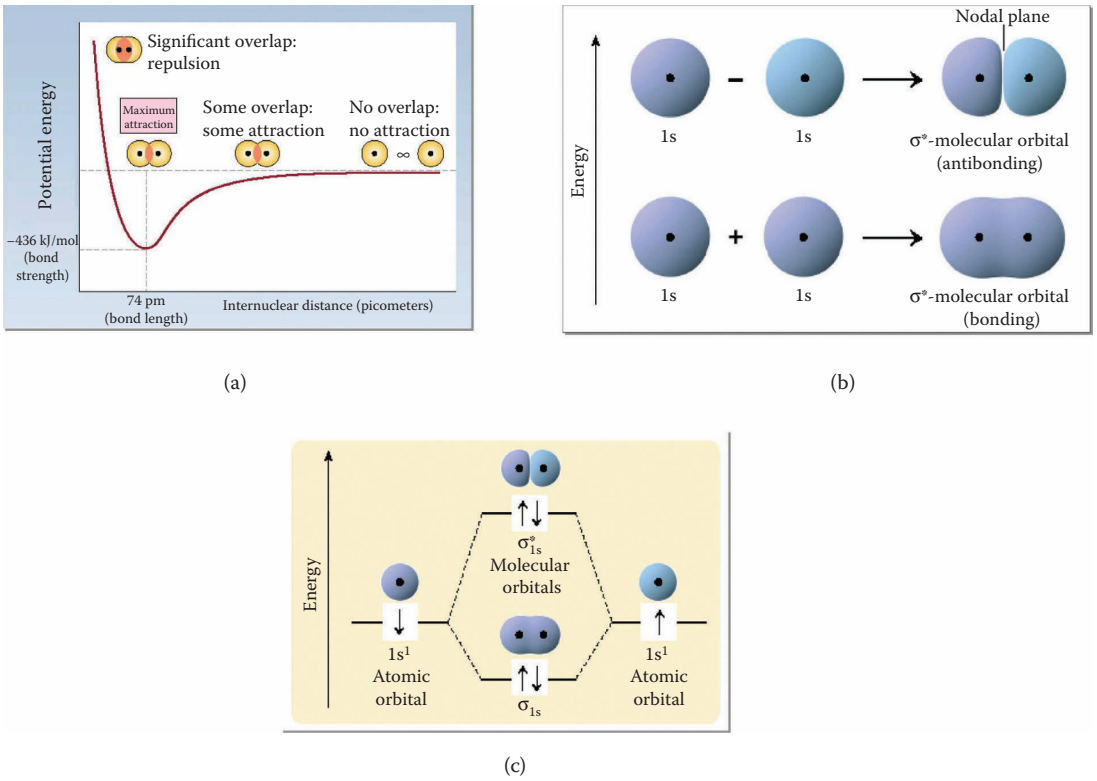


FIGURE 3.83 (a) A diagram of the potential energy versus the internuclear distance for two approaching hydrogen atoms. A sigma bond (σ) is formed from the overlap of the *s* orbitals from the binding hydrogen atoms. At an internuclear distance, r_0, of 74 picometers, equilibrium is reached, and the attraction between the binding partners is maximized with a H-H bond strength of –436 kJ/mol. (b) A bonding and antibonding molecular orbital is formed. (c) The bonding molecular orbital (σ_{1s}) is lower in energy than the parents' atomic orbitals, and the antibonding (σ^*_{1s}) is higher in energy.

with a single hydrogen atom there are twice as many theoretically permitted electron shells; the bonding molecular orbital (σ_{1s}) is lower in energy than the parents' atomic orbitals, and the antibonding (σ^*_{1s}) is higher in energy. The available electrons of the just created molecule are assigned, just as in the case of an atom, orbitals of successively higher energy and in the case of orbitals of equal energy—*degenerate orbitals*—these are filled one electron at a time before paring begins (Hund's rule), and only two electrons of opposite spin can occupy the same energy level (Pauli's exclusion principle). The total number of molecular orbitals equals the number of atomic orbitals contributed by the atoms constituting the molecule. When we examine a very large *N*-atom molecule, such as a long carbon chain hydrocarbon, we find a splitting of each one-atom energy level into *N* energy levels, each one corresponding to a somewhat different electron shell form.

Molecules also have excited states originating from a variety of sources. Because they are not usually spherically symmetric, molecules can rotate around several of their axes. The rotational energy is again quantized and depends on the *l* quantum number. The atoms in the molecule also vibrate compared to each other around an equilibrium distance. This can be described well by a harmonic oscillator spectrum (Equation 3.191).

To form a solid, one keeps adding atoms in three dimensions, making a very large molecule, and that is exactly what we will do next. We will see that in solids, just like in molecules, higher energy states occur where they interact to form a higher band of allowed energies.

Solids: A First Look at The Band Model of Solids

For the total number *N* of atoms in a solid (~10^{23} cm^{-3}), *N* energy levels split apart within a width ΔE at r_0 as shown in Figure 3.84a, where first two, then six, and then *N*, 3s atomic levels combine. This leads to a band of energies for each initial atomic energy level, and 2*N* electrons may occupy an energy band

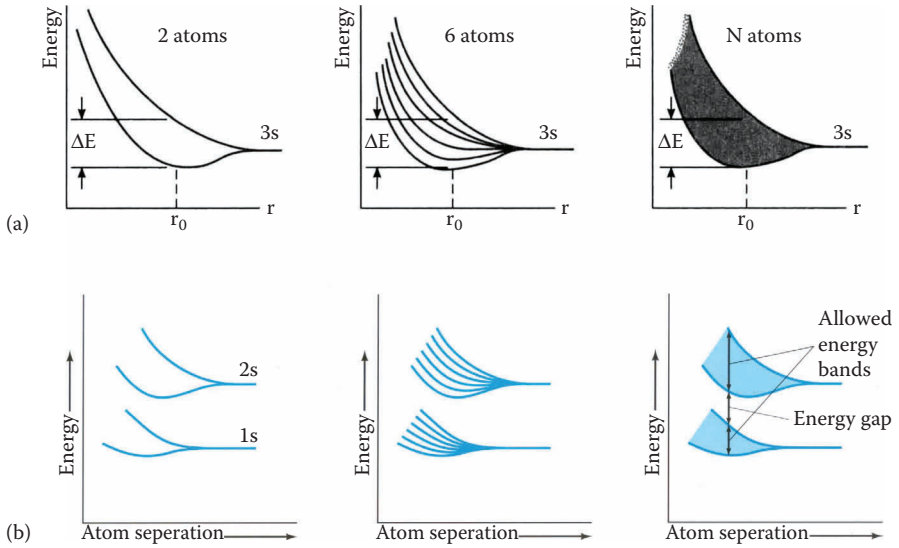


FIGURE 3.84 (a) Two, six, and then *N*, 3s atomic levels combine. This leads to a band of energies for each initial atomic energy level, and 2*N* electrons may occupy an energy band containing *N* energy levels. (b) Sets of 1s and 2s levels are combined into allowed energy bands separated by a forbidden energy zone, the bandgap, E_g.

containing *N* energy levels. In Figure 3.84b, sets of 1s and 2s levels are combined into allowed energy bands separated by a forbidden energy zone, the bandgap, E_g. A group of energy states incompletely filled at room temperature and empty at 0 K is called a conduction band. A conduction band is able to support the movement of electrons. A band of energy levels, missing some electrons at room temperature and completely filled at 0 K, is called a valence band. The valence band supports the movement of missing electrons, which are called holes. As illustrated in Figure 3.85, the relative position of conduction band, valence band, and forbidden bandgap are instructive for the classification of materials into conductors, semiconductors, and insulators. In conductors, the conduction band is partially filled, allowing electrons to move freely, or the valence band overlaps with the conduction band, again enabling free electron movement. In insulators there is a substantial forbidden energy gap between completely filled valence band and empty conduction band (~9 eV in Figure 3.85). Semiconductors are similar to insulators, but the bandgap is narrower, and electrons and/or holes are available at room temperature. Because in semiconductors the energy gap is small, thermal energy might suffice for some electrons to jump from the valence band to the conduction band; however, more often doping with impurities is needed to generate enough charge carriers at room temperature.

Altering the band structure, confining the geometry or the potential of electrons, leads to engineered states with very interesting unseen properties.

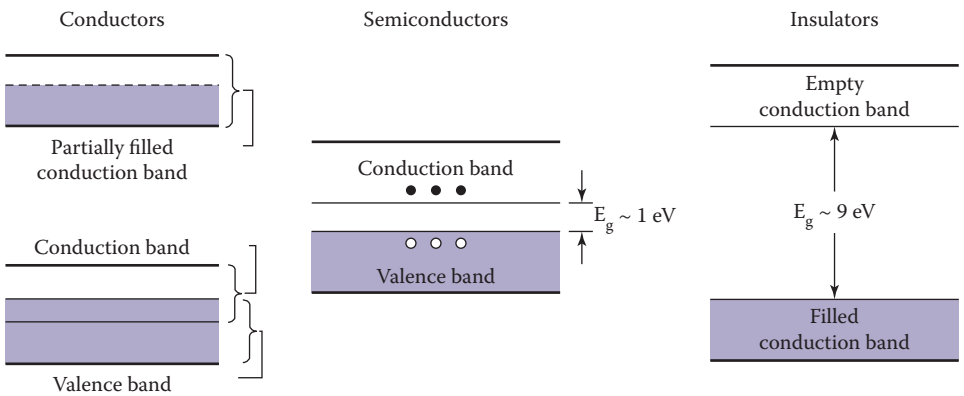


FIGURE 3.85 Classification of materials into conductors, semiconductors, and insulators based on the band model.

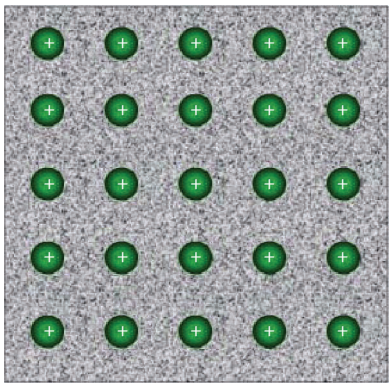


FIGURE 3.86 Positive ion cores in a metal.

Bloch Functions

Introduction Armed with some understanding of quantum mechanics and an introduction to the valence bond (VB) and molecular orbital (MO) theory, we are in a much better position to understand the band model of solids. The Drude free electron gas model surveyed above explained a number of important metallic properties, but it did not explain the most intriguing problems: what is it that distinguishes a metal from an insulator, and how can the same material, say carbon, form a conductor, a superconductor, a semiconductor, or an insulator? These facts remain unanswerable until one also takes into account that an electron gas moves through space occupied by a periodic array of positive-charged atomic cores (Figure 3.86). When electrons are free to roam in an infinite solid, e.g., in a metal, the electronic "orbitals" are traveling wave solutions, but when the metal is limited to a cube of size L^3 or a periodic array of positive ions is assumed, this will act to restrict the allowed energies and hence other quantities such as momentum, spin, and so on. Let us specify the force acting on the "free" electrons as $V(\mathbf{r})$.

Because the ion cores in a crystal are arranged periodically, the potential an electron feels is also periodic; with the periodicity of the underlying Bravais lattice in three dimensions, this yields the Born and von Karman's periodic boundary condition:

$$V(\mathbf{r}+\mathbf{R}) = V(\mathbf{r}) \tag{3.195}$$

for all Bravais lattice vectors $\mathbf{R}$ (see also Chapter 2). In Figure 3.87 we illustrate how the potential $V(\mathbf{r})$ for a single atom compares with that of two and with that of an array of atoms. Electrons in isolated atoms

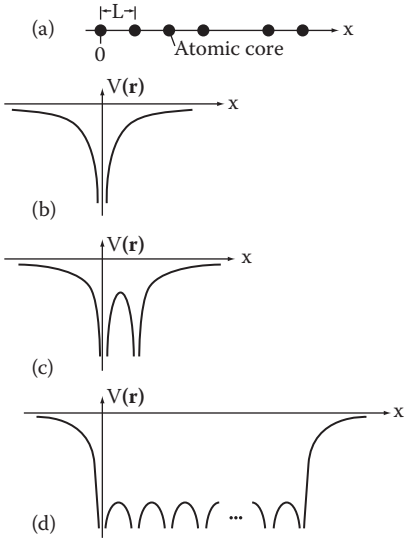


FIGURE 3.87 Bringing atoms together in a lattice. (a) Atomic cores; (b) isolated atom with binding energy as a function of $V(\mathbf{r})$; (c) two atoms brought together; and (d) array of atoms with binding energy as a function of $V(\mathbf{r})$.

occupy discrete allowed energy levels E_1, E_2, and so on, with a potential energy of the electron at a distance r from a positively charged nucleus q given as (see Figure 3.87a):

$$V(\mathbf{r}) = \frac{-qe}{4\pi\varepsilon_0 \mathbf{r}} \tag{3.196}$$

The 1D potential energy of an electron caused by an array of nuclei of charge q separated by a distance $\mathbf{R}$ is:

$$V(\mathbf{r}) = \sum_n \frac{-qe}{4\pi\varepsilon_0 |\mathbf{r} - n\mathbf{R}|} \tag{3.197}$$

where n = 0, ±1, ±2, and so on. This is shown in Figure 3.87d. The periodic potential in one dimension, shown here, is a Krönig and Penney-like potential (see Figure 3.98). From this figure, $V(\mathbf{r})$ is lower in the solid than in the isolated atoms. In the lowest binding energy states, conduction electrons move in a nearly constant potential as shown in Figure 3.88.

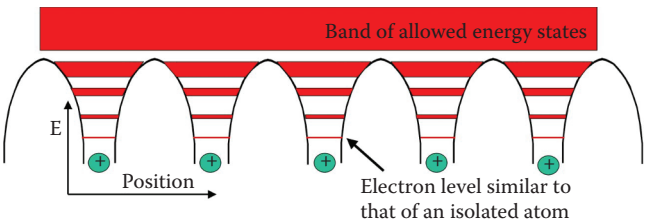


FIGURE 3.88 Band formation.

These electrons are trapped within the metal; a work function of several electron volts prevents them from escaping from the surface.

The higher energy states of tightly bound electrons are very similar to those of the isolated atoms. Lower binding electron states from the atoms become bands of allowed states in the crystal, and we will learn shortly that with partially filled bands the solid becomes a conductor.

Consider now what potentials an electron would see as it moves through a crystal lattice (limited to 1D for now). The electrostatic potential, $V(x)$, is periodic such that $V(x + L) = V(x)$. Bloch's theorem states that because the potential repeats every L length, the magnitude of the wave function (but not necessarily the phase) must also repeat every L length. This is the case because the probability of finding an electron at a given point in the crystal must be the same as found in the same location in any other unit cell. Generalizing, according to Bloch's theorem (1928), if $\psi_0(\mathbf{r})$ is a solution of Schrödinger's equation in free space, then a solution in a potential field that is periodic with period $\mathbf{R}$ is a product of $\psi_0(\mathbf{r})$ and another function $V(\mathbf{r})$, which is itself periodic with period $\mathbf{R}$. This theorem is one of the most important formal results in all of solid-state physics because it tells us the mathematical form of an electron wave function in the presence of a periodic potential energy. Thus, the wave function of the electron in a periodic potential $V(\mathbf{r})$ has the form:

$$\psi(\mathbf{r}) = \psi_0(\mathbf{r}) \cdot V(\mathbf{r}) \tag{3.198}$$

These are the so-called Bloch functions. In the Bloch approach an electron is considered to belong to the crystal as a whole rather than a particular atom. We have encountered the solution for $\psi_0(\mathbf{r})$ already. It is a running wave that can be written as a sine or cosine function or, more generally, in the form $e^{i\mathbf{k}\cdot\mathbf{r}}$, where $\mathbf{k}$ is equal to $2\pi/\lambda$, λ being the wavelength of the electron in this case. The running wave represents the behavior of the free electron. In the presence of the periodic potential, the running wave is modulated to give:

$$\psi(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} \cdot V(\mathbf{r}) \tag{3.199}$$

where $V(\mathbf{r}+\mathbf{R}) = V(\mathbf{r})$ for all R in the Bravais lattice. The Bloch function can be interpreted (approximately) as describing the electron distribution within a single unit cell, and an overall phase variation term $e^{i\mathbf{k}\cdot\mathbf{r}}$ represents a phase difference of the wave function in adjacent unit cells. The latter can take on several values depending on the wave vector $\mathbf{k}$. For core electrons (those tightly bound to the nucleus) Equation 3.199 represents a strongly localized wave function similar to the electron orbitals around a hydrogen atom. The less strongly bound valence electrons are described by more extended wave functions that have significant amplitude between neighboring atoms. Free electrons are described by wave functions with a high energy E, such that the wave vector remains real between atoms [where $V(\mathbf{r})$ is high]. This important result shows that the wave function for the electron itself has the periodicity of the lattice. As a consequence, just like with x-rays or vibrations of atoms in crystals (phonons), only certain wavelengths (that is, energies) are permissible for electrons in crystals. Importantly, we shall see that this leads, among other things, to a natural distinction among metals, insulators, and semiconductors. From Equation 3.199 we must have standing waves in the crystal that have a period equal to a multiple of the period of the crystal's electrostatic potential (similar to a multilayer antireflection coating in optics). Indeed, when an electron travels in a solid and enters a region with a lower potential energy (e.g., closer to a positively charged atom), the kinetic energy goes up, and the wave function acquires a shorter wavelength. This behavior of electron waves entering a low potential region resembles that of light entering a high refractive index region. Just as for light, changes in wavelength give rise to reflections, in this case electron wave reflection. Inside a crystal, electrons experience a periodic potential caused by the regularly spaced atomic cores in the crystal lattice, leading to multiple electron wave reflections and electron wave interference. The multiple reflections result in eigenmodes that are affected by the exact shape of the periodic potential $V(x)$. For an infinite crystal, the eigenfunctions describe a state with a well-defined energy and a corresponding spatial distribution of the electron throughout the entire crystal. In a simple linear lattice with lattice spacing a, we have $V(\mathbf{r} + \mathbf{R}) = V(\mathbf{r})$, as shown above.

It is important to note that because the wave function repeats each unit cell, we only have to consider what happens in one unit cell to describe the entire crystal. Thus, we can restrict ourselves to values of **k** such that $-\pi/a$ to $+\pi/a$ [implying $\mathbf{ka} \leq 1$ or $(2\pi/\lambda)\, a \leq 1$], or we can describe the electron behavior in a solid with wave vectors that lie in the first Brillouin zone.

Bloch Function Applied to a Six-Atom Linear Lattice To get a better appreciation about Bloch functions, let us consider an array of six atoms (N = 6), as illustrated in Figure 3.89. The Bloch function for an electron in this linear solid is $\psi_n(x) = V(x)e^{ik_n x}$, with a function $V(x)$ that depends on the electronic states involved (see Figure 3.90). The Born and von Karman's periodic boundary condition (Equation 3.195) can only be satisfied if the wave vector **k** has N possible values from $\mathbf{k} = \pi/L$ to $\mathbf{k} = \pi/a$, with L the total length of the six-atom crystal and a the lattice spacing. In other words, $\mathbf{k}_n = n\pi/L$, with $n = 1, 2, 3\ldots L/a$, or there are $N = L/a$ states. This is really nothing else than the electron in a box problem, except that the box is now divided in six compartments!

For N = 6, there are six different superpositions of the atomic states that form the crystal states as shown in Figure 3.91 (where we only consider the 1s combinations).

Bloch Function for Metals Let us now take an example metal and determine which states its electrons

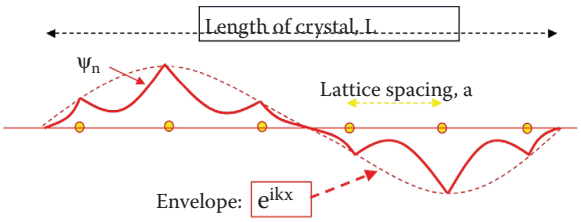


FIGURE 3.89 The states in a six-atom linear crystal (N = 6).

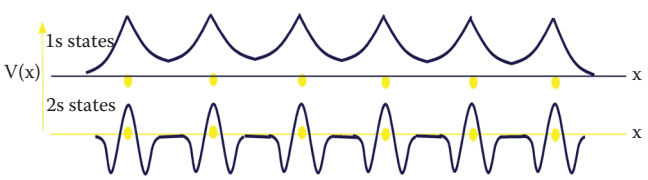


FIGURE 3.90 The periodic potential $V(x)$ depends on the electronic states of the atoms involved.

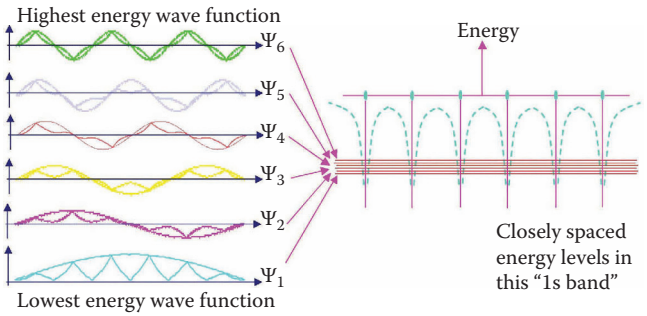


FIGURE 3.91 Six different super positions of the 1s atomic states form the crystal states.

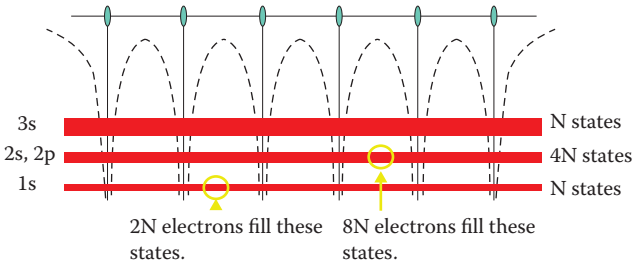


FIGURE 3.92 Filling the Bloch functions of sodium with electrons.

are occupying. In Figure 3.92 we have filled the Bloch functions of sodium with electrons according to Pauli's principle. We consider N sodium atoms, and because for sodium Z = 11 ($1s^2 2s^2 2p^6 3s^1$), this means we have $11N$ electrons to distribute. Notice that the 3s band is only half filled (N orbital states and N electrons). Electrons in this 3s band are easily promoted to higher states in the band (Figure 3.93), and this is what makes sodium a good conductor. In other words, to get a good conductor one needs a band partially filled with electrons.

The Bloch wave functions in metals are stationary waves, and in a perfectly periodic metal lattice, an electron would freely move without scattering from the atomic cores. This corresponds to a metal without any resistance at all! An electron in a periodic potential has a well-defined wave vector and momentum, and it is only when there are defects in the crystal,

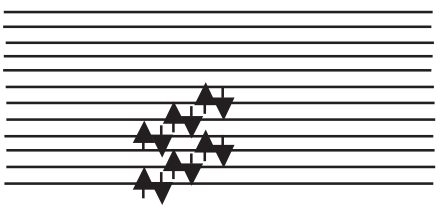


FIGURE 3.93 Half-filled 3*s* band of sodium.

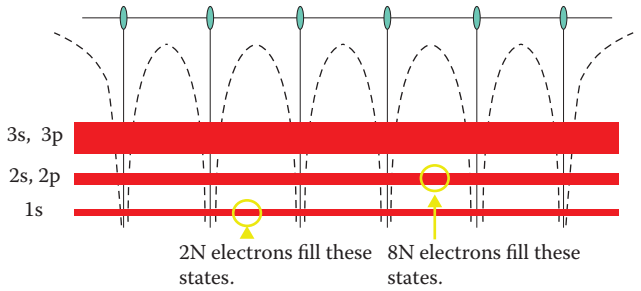


FIGURE 3.94 Filling the Bloch functions of silicon with electrons.

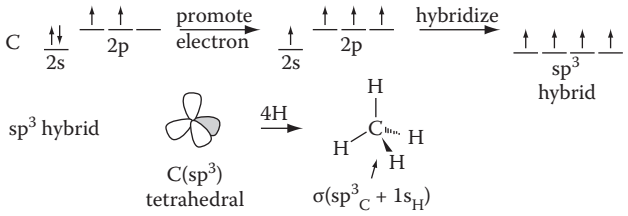


FIGURE 3.95 Hybridization of one 2s and three 2p orbitals in carbon to form four sp^3 hybrid orbitals. In methane carbon, four equivalent bonds are formed with hydrogen.

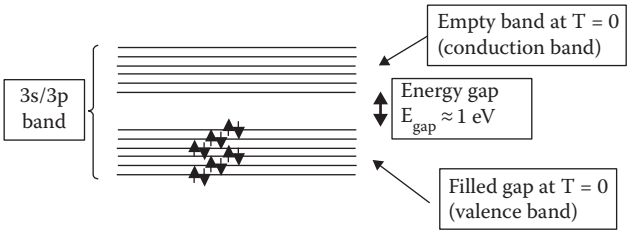


FIGURE 3.96 Two hybrid sp^3 bands in Si split in an empty conduction band and a filled valence band.

breaking up the lattice periodicity, that an electron may scatter to other Bloch states. Lattice vibrations also may break the perfect lattice periodicity, and as a consequence, electrons in metals scatter more at higher temperatures. This is all very different from Drude's hypothesis: it is not collisions with lattice atoms that determine the resistance of a metal but rather defects and lattice vibrations.

Bloch Functions for Semiconductors and Insulators

Next we will fill the Bloch function of a semiconductor such as Si with electrons. This is illustrated in Figure 3.94. In the case of Si, Z = 14, and the atom orbitals are $1s^2 2s^2 2p^6 3s^2 3p^2$. With a total number of N atoms, we will have $14N$ electrons to accommodate. At first blush it appears that, like Na, Si will also have a half-filled band because the $3s3p$ band has $4N$ orbital states and $4N$ electrons. By this analysis, Si should be a good metal, just like Na. But something unique happens for C and Si and other group IV elements.

In group IV elements, bonding orbitals are hybrid combinations of s and p states, so-called sp^3 hybrids. Hybrid atomic orbitals were introduced to reconcile the discrepancy between what atomic orbital theory predicts and what is seen experimentally. Take carbon, for example; the electron configuration is $1s^2$, $2s^2$, $2p^2$, so one expects carbon, in say methane, to form two bonds via the two unpaired p electrons; in practice we know that carbon forms four equivalent bonds. The hybridization process, in the case of C in methane, is explained in Figure 3.95. A $2s$ electron is promoted, and one $2s$ orbital and three $2p$ orbitals form four equivalent sp^3 hybrid orbitals used in binding four hydrogens to make a tetrahedral-shaped methane molecule. Similarly, in a Si lattice, four sp^3 orbitals link all the Si atoms tetrahedrally together. In the Si case, the hybridizing orbitals are one $3s$ orbital and three $3p$ orbitals, and these four equivalent bonding orbitals are completely filled in the single-crystal Si with two electrons each.

The superposition of sp^3 bands between neighboring Si atoms results in a filled bonding band, the valence band, and an empty antibonding band, the conduction band, as illustrated in Figure 3.96.

The electrons in a filled band cannot contribute to conduction because within reasonable E fields they cannot be promoted to a higher kinetic energy. Therefore, at T = 0, Si is an insulator. At higher temperatures, however, electrons are thermally promoted into the conduction band. For a Si crystal at room temperature, the amount of energy an electron must gain to overcome the bandgap is about 40 times more than the average amount of thermal energy. As a consequence, in a semiconductor the number n of free electrons increases rapidly with T (much faster than the scattering time τ decreases), whereas for a metal scattering time, τ gets shorter with increasing T (Figure 3.97).

Thus, energy bands and the gaps between them determine the conductivity and other properties of solids. Insulators have a valence band that is full and a large energy gap (a few eV). Consequently, no states of higher energy are available for electrons to go to.

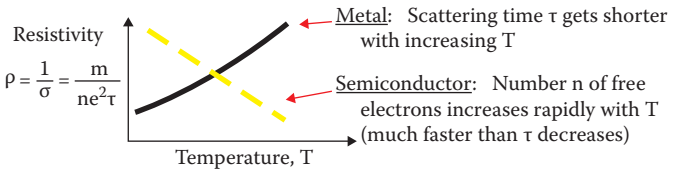


FIGURE 3.97 Resistivity as a function of temperature for a metal and a semiconductor.

Semiconductors are insulators at T = 0, but they have a small energy gap (~1 eV) between valence and conduction bands, so they conduct only at higher *T*. Metals have an upper band that is only partly full; in other words, the Fermi level lies within the valence band, and when applying an electric field, lots of states of higher energy are available for electrons to go to.

The Krönig-Penney Model

Solution of Schrödinger's Equation for a Periodic Potential

Bloch's theorem, along with the use of periodic boundary conditions, enables the calculation of the energy bands of electrons in a crystal if the potential energy function experienced by the electron is known. This was demonstrated for a simple finite square well potential model by Krönig and Penney in 1931, representing the first solution of Schrödinger's equation for a periodic potential. The Krönig-Penney model uses a simple 1D model of a crystalline solid as shown in Figure 3.98. The period of the potential is (a + b = L). The potential (V) is assumed equal to zero in the region of an atom, e.g., for $0 < x < a$, and to equal V_0 in the region between atoms, e.g., $-b < x < 0$. The calculations are a repeat of the calculations for a square barrier carried out above. For the zero regions, where electrons essentially act as free particles, the Schrödinger equation that applies is (Equation 3.171):

$$\frac{d^2\psi}{dx^2} = -k_x^2\psi$$

with (Equation 3.172):

$$k_x = i\frac{\sqrt{2mE}}{\hbar}$$

and the solution is an oscillating wave like in Equation 3.173:

$$\psi = Ae^{ik_x x} + Be^{-ik_x x}$$

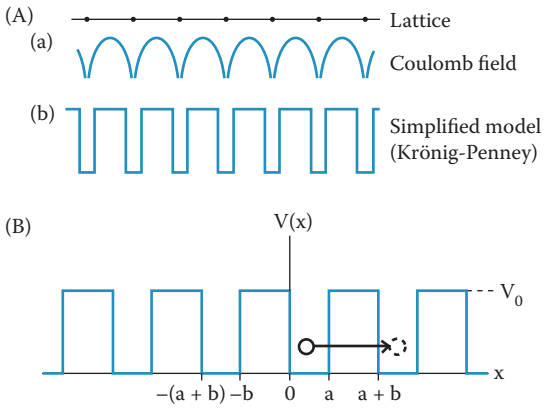


FIGURE 3.98 (A) (a) The periodic lattice potential in a real crystal (see also Figure 3.94). The bullets represent the positions of the nuclei. (b) One-dimensional periodic potential used in the Krönig-Penney model. A central question is whether an electron with energy *E* will be able to propagate from one lattice cell to another. (B) Electrons are essentially free between $0 < x < a$ (and in any similar region along the lattice) and have to tunnel through the barrier regions.

In the nonzero regions, electrons must tunnel through the rectangular barriers for which the following Schrödinger equation applies (Equation 3.166):

$$\frac{d^2\psi}{dx^2} = \alpha^2\psi$$

where α is (Equation 3.167):

$$\alpha = \frac{\sqrt{2m(V_0 - E)}}{\hbar}$$

and the solution is the summation of the damped exponentials of Equations 3.168 and 3.169:

$$\psi(x) = Ce^{+\alpha x} + De^{-\alpha x}$$

It is assumed that the energy E of the electron is always smaller than V_0; in other words, electrons stay in the solid.

We will not detail solving Schrödinger's equations for a linear array of atoms here. A detailed treatment can be found, for example, in A.J. Dekker[22] and better yet in S.O. Pillai.[23] Suffice it to say that one finds two solutions: one for the regions where V(x) = 0, and one for the regions where V(x) = V_0, and that these solutions come with constants *A*, *B*, *C*, and *D*. The boundary conditions are that the wave functions and their derivatives are continuous across the potential boundaries and that

the solutions must be Bloch functions of the form $\psi_n(x) = V(x)e^{ik_x x}$, with $V(x + nL) = V(x)$, $n = 0, +1, +2, +3, \ldots$, representing the symmetry of the assemblage of atoms. With application of these boundary conditions including the use of Bloch's theorem, the constants A, B, C, and D can de determined and the wave functions calculated. One can then show that, under the simplifying conditions, V_0 tends to infinity and b approaches zero, whereas the product $V_0 b$ remains finite, and solutions to the wave equations exist only if:

$$P\frac{\sin\beta a}{\beta a} + \cos\beta a = \cos ka \qquad (3.200)$$

where P is $\frac{mV_0 ba}{\hbar^2}$, a measure for the potential barrier height and width, and with β equal to $\left(\frac{2mE}{\hbar^2}\right)^{\frac{1}{2}}$. In Figure 3.99 we plot $P\frac{\sin\beta a}{\beta a} + \cos\beta a$, the left side of Equation 3.200, versus βa, fixing P at a value of $3\pi/2$, as a typical example. Because β^2 is proportional to the energy E, the x-axis in this figure is a measure of energy. Importantly, the right side in Equation 3.200, the term $\cos ka$, can only have values between –1 and +1, as marked by the two dashed horizontal lines in the figure. Obviously this condition can only be satisfied with values of βa for which the left side lies between +1 and –1.

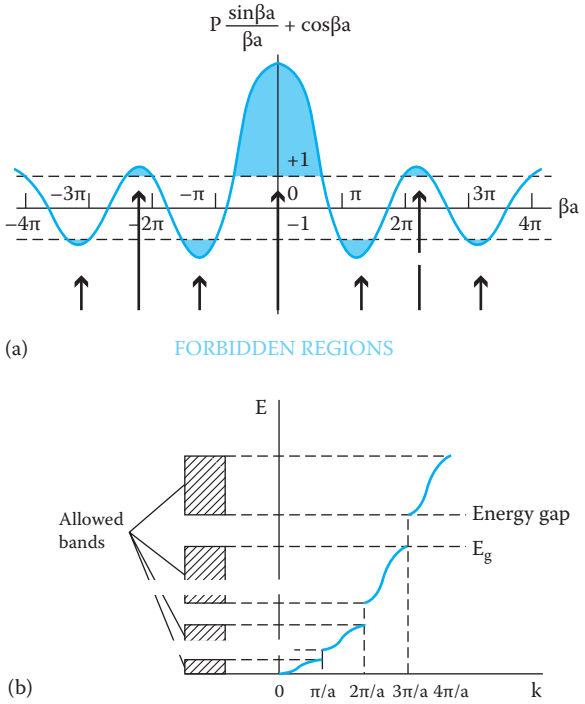


FIGURE 3.99 (a) Plot of $P\frac{\sin\beta a}{\beta a} + \cos\beta a$ versus βa. Arrows point to forbidden regions. (b) The energy as a function of wave number **k**. The allowed values of energy are given by those ranges of $\beta = \left(\frac{2mE}{\hbar^2}\right)^{\frac{1}{2}}$ for which the function lies between +1 and –1.

Analysis of the Solution

Some very important insights can be gained from an analysis of Equation 3.200 and an inspection of Figure 3.99:

1. Because β is related to E, electrons possess energies within certain bands but not outside those bands: there are allowed bands of energy and forbidden bands of energy (see arrows).
2. If V_0 increases, then P increases, the binding energy goes up, and the width of a particular allowed band decreases. The left side of Equation 3.200 becomes steeper, and in the limit, with P tending to infinity, the allowed energy bands reduce to the single energy levels we encounter in isolated atoms. In the latter situation, the equation only has solutions if $\sin\beta a = 0$, in other words, if $\beta a = \pm n\pi$ with $n = 1, 2, 3, \ldots$ and the energy spectrum becomes that of an electron in a constant potential box of atomic dimensions. In such cases, the energy levels are the energy levels of a potential well with $L = a$: $E_n = \frac{\pi^2\hbar^2}{2ma^2}n^2$ (see Equation 3.156). Each electron is confined to one atom by an infinite potential well, so electrons are completely bound to atoms.
3. In the case we decrease V_0 and reduce P to zero—in other words, when the binding energy goes to zero—Equation 3.200 is reduced to $\cos(\beta a) = \cos ka$ ($\beta = \mathbf{k}$), and the energy E is now given by $E = \frac{\hbar^2 k^2}{2m}$, i.e., the parabolic E-**k** relation for free electrons (see Figure 3.55). By varying V_0, the model thus covers the whole range from completely free electron to completely bound electron. In Figure 3.100 we illustrate the energy level structure as a function of varying degrees of binding strength.

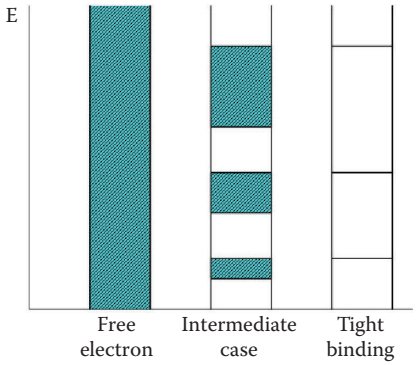


FIGURE 3.100 Energy level structure as a function of binding strength.

4. The width of allowed energy bands increases with increasing values of βa, i.e., with increasing energy, because the first term in the equation decreases on average with increasing βa.
5. From Figure 3.99a we see that at the boundary of an allowed energy band the cos (**k**a) = ±1 with **k** = nπ/a (dashed horizontal lines in the figure). Based on Equation 3.200 we can represent the energy also as a function of wave number **k**. The result is shown in Figure 3.99b and Figure 3.101 further below. This particular way of displaying the electronic levels in a periodic potential is known as the extended-zone scheme. Discontinuities in the *E*-**k** curve occur for **k** = nπ/a with n = 1, 2, 3,…. The zones in **k**-space that correspond to allowed energies for motion of an electron are the Brillouin zones (BZ), which we first encountered in Chapter 2.* The periodic potential $V(x)$ splits the free electron *E*-**k** curve into "energy bands" separated by gaps at each BZ boundary. Electrons can never have an energy within this energy gap. In a metal, the periodic potential $V(x)$ is very small or 0; in a semiconductor, the potential is large and consequently the energy splitting is large. The first Brillouin zone extends between **k** = –π/a and **k** = +π/a; the second is in the range of **k** from $-2\pi/a$ to $-\pi/a$ plus the range from π/a to $2\pi/a$. There can be no energy value for an electron between the bottom of the conduction band, E_c, and the top of the valence band, E_v. Therefore, the value $E_c - E_v = E_g$ is an energy gap at **k** = ±π/a. The existence of forbidden energies has very fundamental consequences. It happens for electrons in crystals with a periodically varying potential and for photons in systems with a periodically varying refractive index, in which case we call the gap a photonic bandgap (see Chapter 5). Brillouin zones are further detailed in the section below.

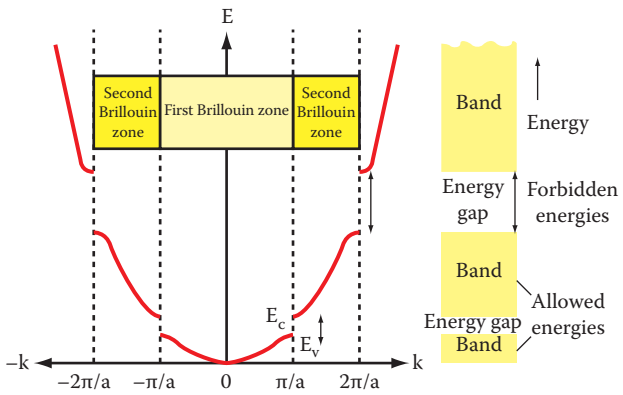


FIGURE 3.101 Energy versus wave number for motion of an electron in a one-dimensional periodic potential. The range of allowed **k** values goes from $-\pi/a$ to $+\pi/a$ corresponding to the first Brillouin zone for this system. Similarly, the second Brillouin zone consists of two parts: one extending from π/a to $2\pi/a$, and another part extending between $-\pi/a$ and $-2\pi/a$. This representation is called the extended zone scheme. Deviations from free electrons parabola are easily identified.

6. Finally, inspection of Equation 3.200 reveals that when **k** is substituted by $\mathbf{k} + 2\pi n/a$, where n is an integer, the right side of the equation remains the same, or **k** is not uniquely determined. In other words, in a given energy band, the energy is a periodic function of **k**. Given this periodicity, it is often convenient to introduce a "reduced wave vector" as $-\pi/a \le \mathbf{k} \le +\pi/a$. By shifting the second Brillouin zone $2\pi/a$ left or right, one can obtain the reduced zone scheme. A representation of energy versus reduced wave vector is marked by a double pointed arrow in Figure 3.102a. Obviously one does not need all *E*-**k** curves in all BZs. All information is already contained in the first Brillouin zone in a reduced zone scheme because of the $2\pi/a$ periodicity. In Figure 3.102a, we also display the electronic

* From Chapter 2 we remember that the x-ray diffraction pattern of a crystal is a map of the reciprocal lattice. It is a Fourier transform of the lattice in real space or also the representation of the lattice in **k**-space. This is true for the x-rays considered in Chapter 2 but also for the matter waves associated with electrons, neutrons, and so on. We are presenting BZs for all dimensionalities of the electronic structure further below.

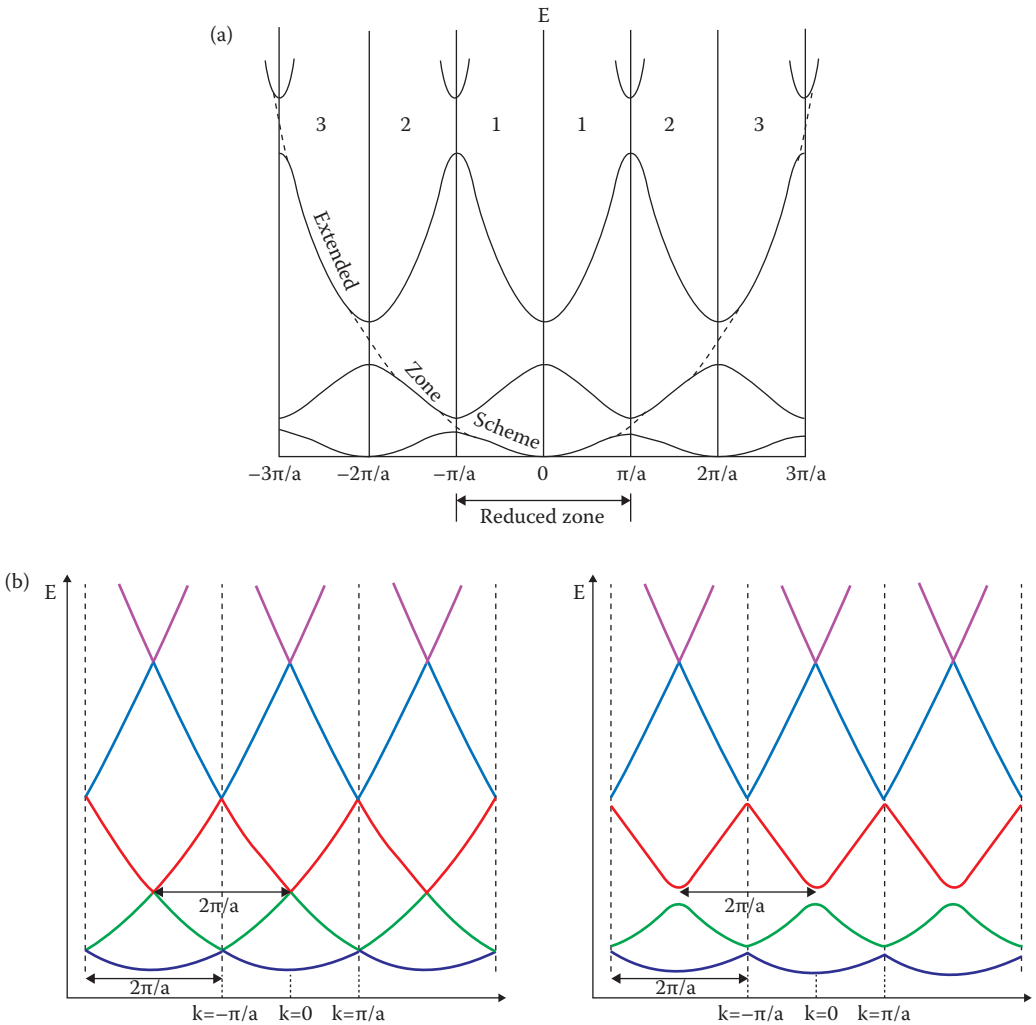


FIGURE 3.102 (a) Different representations of *E*-**k** in the presence of a periodic potential, band splitting. The extended zone scheme: plot *E*-**k** from **k** = 0 through all possible Brillouin zones (**bold curve**); the periodic or repeated zone scheme: redraw *E*-**k** in each zone and superimpose; the reduced zone scheme: all states with $|\mathbf{k}| > \pi/a$ are translated back into the first BZ. (b) Different *E*-**k** plots in the presence of a vanishing periodic potential: zone folding (left) and a nonvanishing periodic potential (right).

levels in a repeated or periodic zone scheme and repeat the extended zone scheme from Figure 3.101. In Figure 3.102b, we show what happens when the periodic potential becomes vanishing small; the bandgaps disappear and we get zone folding.

It is easy to show that the number of **k** values in each BZ is just *N*, the number of primitive unit cells in the sample. For this consider the finite, linear crystal with a total length of L = Na. The allowed values of the electron wave vector **k** in the first Brillouin zone for this arrangement are exactly N ($k = 0, \pm 2\pi/L, \pm 4\pi/L, \ldots, \pm N\pi/L$). Each primitive cell contributes exactly one independent value of **k** to each energy band. Thus, 2*N* electrons resulting from spin degeneracy can occupy each band. A monovalent element with one atom per primitive cell has only one valence electron per primitive cell and thus *N* electrons in the lowest energy band. This band will only be half-filled, and the material will be a conductor. The Fermi energy, the energy dividing the occupied and unoccupied states, for such a monovalent element will be in the middle of the valence band. If each atom contributes two valence electrons to the band (divalent element), the band, at T = 0 K, will be filled to the top. Thus, the simple rule whether an element is an insulator or a metal is given by whether the number of electrons is odd or even. This rule works surprisingly well;

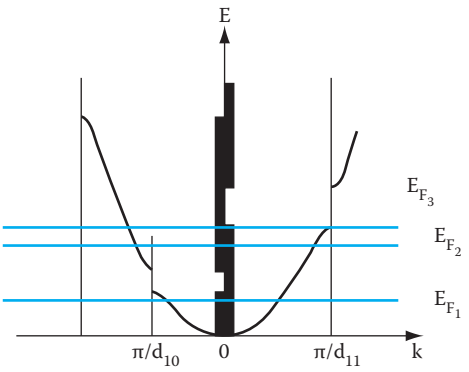


FIGURE 3.103 The presence of more than one periodicity in a crystal may cause the overlap in the integrated density of states as shown here as black rectangles along the *E*-axis. The material depicted in the *E*-**k** diagram is a semimetal conductor; see also Figure 3.104b.

the exceptions are caused by bands with more complicated structures in three dimensions than those discussed here. Indeed, the fact that a monovalent element is a conductor does not mean that a divalent element will always be an insulator. Although true in the 1D, it is not necessarily so in 2D or 3D! Bands along different directions in **k**-space can overlap, so that electrons can partially occupy both of the overlapping bands and thus form a semimetal. A semimetal is a metal with a carrier concentration several orders of magnitude smaller than the 10^{22} cm^{-3} typical for ordinary metals. Graphite and the pentavalent conducting elements, for example, are semimetals. A 2D band diagram for a semimetal is shown in Figure 3.103. In this example we show a material with three valence electrons and Fermi levels E_{F_1} to E_{F_3}. This case illustrates how bands in different directions overlap, resulting in a semimetal (this corresponds to the situation depicted in Figure 3.104b).

However, it remains true that only crystals with an even number of valence electrons in a primitive cell can be insulators. Depending on the energy band structure, for a given number of electrons, you can get a different filling, and it is the band structure that determines electronic and optical behavior. Some example band configurations leading to insulating and metallic behavior are shown in Figure 3.104.

The Effective Mass, Velocity of Charge Carriers, and Crystal Momentum

Up to this point we have implied that the mass of an electron in a solid is the same as the mass of a free electron (m_e). In reality, for some solids the measured electron mass is larger than that of the free electron, and in other cases it is smaller. The cause for this deviation is found in the interactions between the drifting electrons and the atoms in the crystal. The mass of an electron in a crystal is called the *effective mass*, m_e^*, and in general it is different from the mass of a free electron. The effective mass, m_e^*, of an electron is the mass of the free electron modified by the presence of the periodic potential of an array of positive lattice ions. Rather than moving undisturbed through the lattice, electrons are constantly jostled by atom movements (phonons). By lowering the temperature, atoms move more sluggishly, and this reduces the lattice resistance for electrons. However, temperature reduction does not reduce the

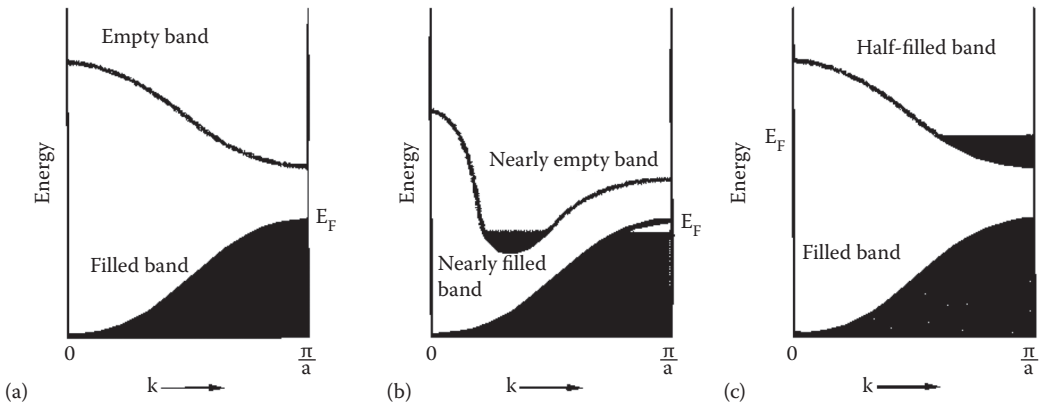


FIGURE 3.104 Occupied states and band structures leading to (a) semiconductor or insulator: two electrons per atom, *N* states per band, and two electrons/state; (b) semimetal: two electrons per atom, *N* states per band, and two electrons/state, a semimetal because bands overlap; (c) and metal: three electrons per atom (Li), *N* states per band, and two electrons/state, metal because of electron population.

influence of lattice defects; the only avenue here is to purify the crystal further, but even the purest crystal features some remaining defects. A third phenomenon controlling the speed of an electron through a crystal is the electron mobility, $\mu_e = \frac{e\tau}{m_e}$ (Equation 3.9). Applying the same voltage to equally pure samples of Si and GaAs, one finds that electrons in GaAs are accelerated much more. The electrons in Si and GaAs are of course the same, but their host lattices influence them differently. We refer to this property by saying that an electron has an effective mass in a given material; therefore, the effective mass m_e^* of an electron in gallium arsenide is less than that of an electron in silicon (making μ_e in GaAs larger; see Equation 3.9).

For a free particle such as an electron, we derived a wave solution $\psi(x) = Ae^{ik_x x}$ with energy $E = \frac{\hbar^2 k_x^2}{2m_e}\left(= \frac{p^2}{2m_e}\right)$ (Equation 3.144). For a free electron, the quantity $\hbar\mathbf{k}$ represents the true momentum $\mathbf{p}$ of the electron. For electrons in a crystal, we need to define a *crystal momentum* that is different from $\mathbf{p}$ because the energy for electrons in a crystal does not vary in the same fashion with $\mathbf{k}$ as it does for free electrons—in other words, the dispersion curve $E(\mathbf{k})$ is different for an electron in lattice than that of a free electron. When dealing with interactions of the electron with the lattice, we must use the conservation of crystal momentum $\hbar\mathbf{k}$ and not that of a free electron.

We will now show how the effective mass and the velocity of an electron in a lattice can both be derived from a knowledge of the energy dispersion curves $E(\mathbf{k})$. Remember that an electron state is a wave packet with a group velocity v_g given as the derivative of ω, the angular frequency of the de Broglie waves with respect to the wave number $\mathbf{k}$ or:

$$v_g = \frac{d\omega}{d\mathbf{k}} \tag{3.101}$$

From quantum theory the frequency of a wave function with energy E is $\omega = \frac{E}{\hbar}$ (because $E = \hbar\omega$), so that we may write:

$$\frac{dE}{d\mathbf{k}} = \hbar\frac{d\omega}{d\mathbf{k}} \tag{3.201}$$

Substituting this expression in Equation 3.101 results in:

$$v_g = \frac{1}{\hbar}\left[\frac{dE}{d\mathbf{k}}\right] \text{ or more generally } \mathbf{v}_g = \frac{1}{\hbar}\nabla_k E(\mathbf{k}) \tag{3.202}$$

The group velocity of a wave packet is basically the velocity at which the wave packet transports energy through the system. It also gives the average velocity of the Bloch electron. The influence the crystal is exerting on the electron motion is contained in the dispersion relation $E(\mathbf{k})$, and the velocity of an electron in a crystal then depends on the dispersion curve $E(\mathbf{k})$. Let us illustrate this point with a simple example. For a free electron, with the parabolic dispersion function $E(\mathbf{k}) = \frac{\hbar^2 k_x^2}{2m}$ (Equation 3.144), we derive $\frac{dE}{d\mathbf{k}} = \frac{\hbar^2\mathbf{k}}{m_e}$, and using Equation 3.202 we calculate for the free electron velocity:

$$v_g = \frac{\hbar\mathbf{k}}{m_e} = \left(\frac{\hbar\mathbf{k}}{m_e}\right)\left(\frac{2\pi\mathbf{p}}{h}\right) = \frac{\mathbf{p}}{m_e} \tag{3.203}$$

In this equation we have a linear relation of v_g with $\mathbf{k}$, which is of course expected because E is proportional to $\mathbf{k}^2$. In general, E is not proportional to $\mathbf{k}^2$, or at least only in small regions as schematically reproduced here in Figure 3.105d.

In Figure 3.105c, we show the velocity v as a function of $\mathbf{k}$. We observe that at the bottom of the energy band ($\mathbf{k} = 0$), the velocity is zero and then increases with $\mathbf{k}$ until it reaches a maximum value at $\mathbf{k} = \mathbf{k}_0$, corresponding to the inflection point in the E-$\mathbf{k}$ curve. Beyond the inflection point, the velocity starts to decrease and finally assumes the zero value at $\mathbf{k} = \pi/a$, which is at the top of the band.

An electron has a well-defined mass, and when accelerated it obeys Newtonian mechanics. To calculate the acceleration ($a = dv_g/dt$) of an electron in a crystal we derive from Equation 3.202 that:

$$a = \left(\frac{2\pi}{h}\right)\frac{d}{dt}\left(\frac{dE}{d\mathbf{k}}\right) \text{ or also } a = \left(\frac{2\pi}{h}\right)\frac{d^2E}{d\mathbf{k}^2}\frac{d\mathbf{k}}{dt} \tag{3.204}$$

The term $\frac{d^2E}{d\mathbf{k}^2}$ we can get from the E-$\mathbf{k}$ relationship, but we still need to derive $d\mathbf{k}/dt$ tocalculate a. For an

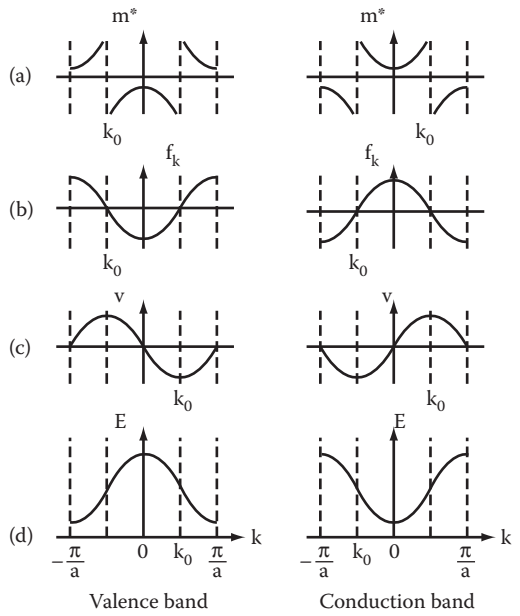


FIGURE 3.105 Effective mass m^*, f_k, velocity v, and energy E as a function of $\mathbf{k}_0$ for electrons in the conduction band (right) and holes in the valence band (left). See text for details.

electron with velocity $\mathbf{v}_g$, subjected to the influence of an external field E applied for a time dt, the distance traveled is $\mathbf{v}_g dt$, so that the work dE done by the electrical field on the electron is:

$$dE = -eE\mathbf{v}_g dt \tag{3.205}$$

Substituting the value for $\mathbf{v}_g$ from Equation 3.202 in the above equation, we get:

$$dE = -eE\left(\frac{2\pi}{h}\right)\left(\frac{dE}{d\mathbf{k}}\right)dt \tag{3.206}$$

or:

$$\frac{d\mathbf{k}}{dt} = -\frac{2\pi eE}{h} \tag{3.207}$$

or since $\mathbf{F} = -e\mathbf{E}$:

$$\hbar\frac{d\mathbf{k}}{dt} = \mathbf{F} \tag{3.208}$$

The last expression shows us that in a crystal $\hbar\dfrac{d\mathbf{k}}{dt}$ is equal to the external force on the electron, whereas in free space the force is equal to $d(mv)/dt$. The electron in a crystal is subject to both forces from the lattice and from external fields. In case there is also a magnetic field present, the force term $\mathbf{F}$ in Equation 3.208 must include the Lorentz force, $\mathbf{F}_L = -e(\mathbf{v}_g \times \mathbf{B})$ (see Equation 3.64), so that the equation of motion of an electron with group velocity $\mathbf{v}_g$ in a constant magnetic field $\mathbf{B}$ is given by:

$$\hbar\frac{d\mathbf{k}}{dt} = -e(\mathbf{v}_g \times \mathbf{B}) \tag{3.209}$$

Combining Equation 3.204 with Equation 3.207 (no magnetic field) we finally derive for a:

$$a = \left(\frac{4\pi^2}{h^2}\right)eE\left(\frac{d^2E}{d\mathbf{k}^2}\right) \tag{3.210}$$

Comparing this equation with that for a free, classical particle where we have $m_e(dv/dt) = eE$ and $a = (dv/dt) = (eE/m_e)$, we then define the effective electron mass m_e^* as:

$$m_e^* = \left(\frac{h^2}{4\pi^2}\right)\left(\frac{d^2E}{d\mathbf{k}^2}\right)^{-1} \tag{3.211}$$

so that:

$$a = \left(\frac{eE}{m_e^*}\right) \tag{3.212}$$

From Equation 3.211, the effective mass is determined by $\left(\dfrac{d^2E}{d\mathbf{k}^2}\right)^{-1}$ (see Figure 3.105a). For a free electron $m_e^* = m_e$ because $E = \dfrac{\hbar^2\mathbf{k}^2}{2m_e}$ and $\dfrac{d^2E}{d\mathbf{k}^2} = \dfrac{\hbar^2}{m_e}$. All the equations for a free electron may be used for an electron in a crystal, provided that m_e in each case is replaced by the suitable m_e^*. For example, we may write:

$$E = \frac{\hbar^2\mathbf{k}^2}{2m_e^*} \tag{3.213}$$

Typically one considers states near the top of the valence band to be holes (particles of charge $+e$) with free electron-like dynamics but with effective mass m_h^*, and states near the bottom of the conduction band to be electrons with free electron-like dynamics but effective mass m_e^*. The effective mass is inversely proportional to the curvature of the band, and in general m^* is different in each direction of the crystal and is a tensor.

From experimental values for m_e^*, it is apparent that the effective mass need not always be larger than m_e. It can be smaller, and it may even be negative. For most metals, it is from one-half to twice m_e. For some transition metals, it is much higher than m_e, and for semiconductors it is lower. From Figure 3.105a, near $\mathbf{k} = 0$,

the effective mass approaches m_e. With **k** increasing, m_e^* also increases, and it reaches a maximum value at the inflection point ($\mathbf{k}_0$) of the E-**k** curve. Above the inflection point ($\mathbf{k} > \mathbf{k}_0$), m_e^* becomes negative, and as **k** approaches π/a it decreases to a small negative value. Near the bottom of the band, the effective mass m_e^* has a constant positive value because the quadratic equation $E \propto k^2$ holds over a small region here (second derivative is a constant). As **k** increases, the quadratic relation no longer holds, and m_e^* changes with **k**. Beyond $\mathbf{k}_0$, the mass m^* becomes negative because the region is close to the top of the band, and a negative mass is to be expected. The way to interpret this is to consider that for $\mathbf{k} > \mathbf{k}_0$ the velocity decreases, and therefore the acceleration is negative, implying a negative mass. In this region of **k**-space, the lattice exerts such a large retarding force on the electron that it overcomes the applied force and produces a negative acceleration. In other words, in the upper half of the band the electron behaves as a positively charged particle, referred to as a hole. Because we can describe the holes in terms of an E-**k** diagram, we can again define an effective mass simply by putting a minus sign in front of Equation 3.210. This turns out to be positive and given by the curvature of the E-**k** diagram at the top of the valence band. Therefore, here we finally have our answer why even metals might produce a positive Hall effect and why the conductivity is not simply proportional to the electron density; it matters where in the band these electrons find themselves!

Neither the electron nor the hole as described by its effective mass exists outside of the material; they are more properly referred to as quasiparticles. The fictitious positively charged particle is even stranger than a conduction band electron with an effective mass. If we add a hole to a completely filled valence band, we must end up with an empty electron state. Therefore, the hole must have a charge of $+e$ (compared with an electron's $-e$). If an empty state exists at low energy, then it is energetically favorable for an electron from a higher energy state to fall into it. In terms of holes, this means it is energetically favorable for the hole to rise to the top of the valence band, i.e., the energy scales for holes are reversed from those for electrons.

In Figure 3.105b we plot the degree of freedom of an electron, f_k, as a function of **k**. Here f_k, a measure of the extent to which an electron in state **k** is free, is defined as:

$$f_k = \frac{m}{m^*} = \frac{m}{\hbar^2}\left(\frac{d^2E}{d\mathbf{k}^2}\right) \tag{3.214}$$

If the effective mass is large, f_k is small, or the particle behaves as a heavy particle. When $f_k = 1$, the electron behaves as a free electron. In the lower half of the band, f_k is positive, and in the upper half f_k is negative.

The behavior of electrons and holes near the band edges determines most of the optical and electronic properties of a solid-state device. From the preceding, near the band edges, the electrons and the holes can be described by a simple effective mass picture, i.e., the electrons behave as if they are in free space except their masses are m_e^* and m_h^*, respectively. In Figure 3.106, we show a four-band model for a generic semiconductor with a parabolic approximation of the bands. This simplified band structure was first proposed by Kane (1957) and is valid near $\mathbf{k} = 0$. The valence band features a heavy holes (hh) band (I), a light holes (lh) band (II), and a split-off (so) band (III).

Electron transitions 1 and 2 are from the heavy holes band to the conduction band and from the light holes band to the conduction band, respectively. Split-off hole transitions are also possible.

As shown in Figure 3.106, along a given direction, the top two degenerate valence bands can be approximately fitted by two parabolic bands with different curvatures: the heavy holes band (the wider band,

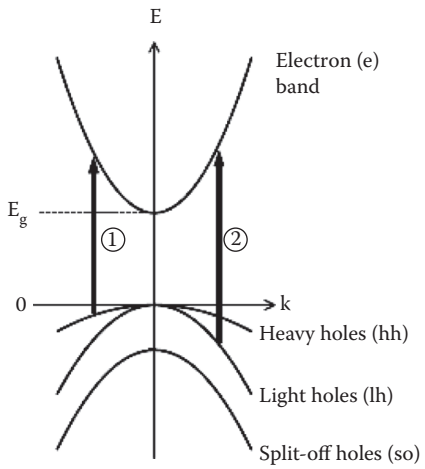


FIGURE 3.106 Four-band model. In the valence band, we have a heavy holes (hh) band (I), a light holes (lh) band (II), and a split-off (so) band (III).

with smaller $\frac{\partial^2 E(k)}{\partial k_i \partial k_j}$) and the light holes band (the narrower band, with larger $\frac{\partial^2 E(k)}{\partial k_i \partial k_j}$). Thus, the effective mass, in general, is tensorial with components:

$$\frac{1}{m^*_{ij}} \equiv \frac{1}{\hbar^2} \frac{\partial^2 E(k)}{\partial k_i \partial k_j} \tag{3.215}$$

which represents a generalization of Equation 3.211.

Example 3.5: In Si, $m_{hh} = 0.53\ m_0$ and $m_{lh} = 0.16\ m_0$; and in GaAs, $m_{hh} = 0.51\ m_0$ and $m_{lh} = 0.074\ m_0$.

Because there are multiple valence bands available for holes but there is only one conduction band for electrons, electromagnetic radiation absorption by holes is different from that of electrons. Free electrons located at wave vector $\mathbf{k} = 0$, at the bottom of a single parabolic conduction band, may reach higher energy states only if their momentum is increased. As a consequence, in n-type semiconductors, conduction band electrons may only be excited by the simultaneous absorption of a photon and the absorption or the emission of a phonon to conserve momentum (see also direct and indirect bandgap transitions discussed further below). This three-particle process (photon-electron-phonon) is less probable than a two-particle process (electron-photon). With p-type semiconductors, one can have such direct (i.e., no phonons needed) transitions between the different degenerate valence bands (heavy and light hole bands) because they occur at the same **k**. Because no momentum or **k** change is required, holes produce stronger free carrier absorption than conduction band electrons.

Below we take a closer look at the density of states (DOS) function, which describes the number of allowed energy states that are available in a system per unit energy and per unit volume (i.e., in units of number of states/eV/cm^3). From the preceding we can appreciate that the density of states, the effective mass, and the electron mobility are all correlated. When traveling through a crystal and bumped off course by a phonon or an impurity, the new energies an electron can adopt depend on the number of available states at the bottom of the conduction band. The density of states at the bottom of the conduction band is larger for Si than it is for GaAs, and as a consequence an electron traveling through silicon has a greater chance of being knocked off course into an allowed energy state. In general, then, a small effective mass is indicative of a relatively low number of energy levels at the bottom of the conduction band. The effective mass of an electron in silicon is six times heavier than that in GaAs, and as a consequence, based on Equation 3.9, the electron mobility in GaAs is six times larger or an electron can race six times faster through a GaAs lattice compared with a Si lattice (see Table 3.9). GaAs transistors, although much more difficult to fabricate than Si transistors, are used in cases where speed is of utmost importance; this includes military applications and the latest generations of supercomputers (e.g., the Cray 3).

Table 3.9 Electron and Hole Mobilities

Material	Mobility (cm^2/V-s)
Si	$\mu_n = 1500$, $\mu_p = 460$
Ge	$\mu_n = 3900$, $\mu_p = 1900$
GaAs	$\mu_n = 8000$, $\mu_p = 380$

Particle Distributions Functions

Introduction

Particle distributions functions *f(T, E)* represent the probability that a particular state with energy *E* is occupied by a particle (say an electron) in equilibrium at a given temperature *T*. Earlier in this chapter we introduced the Maxwell-Boltzmann distribution, a classical distribution of particles, illustrated in Figure 3.6. In this distribution function, it is assumed that all the particles are distinguishable. This kind of consideration comes from the fact that all particles have characteristic wave properties according to the de Broglie hypothesis. Two particles can be considered to be distinguishable if their separation *d* is large compared with their de Broglie wavelength and are considered to be indistinguishable if their wave packets overlap significantly. In such case, a Fermi-Dirac particle distribution and (f_{FD}) must be applied. The thermal de Broglie wavelength is roughly the average de Broglie wavelength

of the gas particles in an ideal gas at the specified temperature and is given by:

$$\lambda_{DB} = \sqrt{\frac{2\pi\hbar^2}{mk_B T}} \sim \frac{h}{p} \tag{3.216}$$

We can take the average interparticle spacing in an ideal gas to be approximately $(n = V/N)^{1/3}$, where V is the volume, N is the number of particles, and n is the density of particles. When the thermal de Broglie wavelength is much smaller than the interparticle distance, the gas can be considered to be a classical or a Maxwell-Boltzmann gas, or:

$$\lambda_{DB} <<< d = n^{-\frac{1}{3}} \tag{3.217}$$

On the other hand, when the thermal de Broglie wavelength is on the order of, or larger than, the interparticle distance, quantum effects will dominate, and the gas must be treated as a Fermi gas or a Bose gas, depending on the nature of the gas particles. Particles become indistinguishable when $d = n^{-1/3} \sim \lambda_{DB}$. The de Broglie temperature is given by:

$$T_{DB} = \frac{2\pi\hbar^2}{mk_B} n^{\frac{2}{3}} \tag{3.218}$$

Example 3.6: For an electron gas in metals, $n = 10^{22}$ cm^{-3}, $m = m_e$, and $T_{DB} \sim 3 \times 10^4$ K, but for a gas of Rb atoms, $n = 10^{15}$ cm^{-3} and $m_{atom} = 10^5 m_e$, so we obtain a $T_{DB} \sim 5 \times 10^{-6}$ K. In other words, at room temperature electrons are indistinguishable with overlapping wave functions, whereas Rb atoms only become so at very low temperatures.

Besides the Fermi-Dirac distribution, we will also introduce the Bose-Einstein distribution function (f_{BE}) and explain under what conditions they each apply. We will see that at $T < T_{DB}$, f_{BE} and f_{FD} are strongly different from f_{MB} and at $T >> T_{DB}$: $f_{BE} \approx f_{FD} \approx f_{MB}$.

Fermi-Dirac

When assigning electrons to the energy levels one must require, as Sommerfeld did, that the allowed wave functions obey Pauli's principle (introduced in 1925), i.e., one can only put two electrons with opposite spin in each level of quantum number n. When that is done this level is filled. One then proceeds to the next higher level for the next pair of electrons. The result is obvious—all the lowest levels are filled with pairs of electrons until they reach some maximum value of energy E_F, the so-called Fermi energy. The exclusion principle applies to all "spin one-half" particles, which include electrons, protons, and neutrons. If we draw the distribution function, i.e., the probability of filling a level, $f(E,T)$, as a function of E, for two different temperatures, we find the result shown in Figure 3.107.

The probability of an energy state being occupied is called the Fermi factor, $f(E,T)$. As T approaches zero, the Fermi-Dirac distribution becomes a step function. At 0 K, f(E, T) = 1 until we reach the maximum level E_F, after which it falls to zero. If we increase the temperature, thermal energy can excite electrons to energy levels higher than E_F. Because the kinetic energy of the lattice ions is of the order of kT (~0.025 eV), electrons cannot gain much more than kT in collisions with lattice ions. This is determined by the Fermi-Dirac distribution, which applies for any particle that follows Pauli's exclusion principle—no more than two particles of opposite spin being allowed in a given energy level. Half-integer spin particles are called fermions, and Fermi-Dirac statistics and Pauli's exclusion principle hold. The "material" particles, such as electrons, protons, and neutrons, are all fermions, and without Pauli's exclusion principle, the plethora of chemical elements and the variety of our physical world would simply not exist. Other particles such as α-particles, deuterons, photons, and mesons do not obey Pauli's exclusion principle. Such particles are called bosons

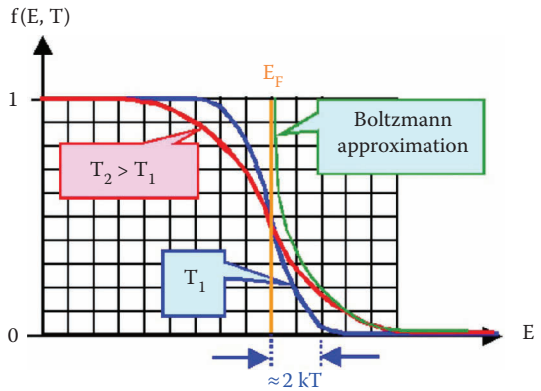


FIGURE 3.107 The Fermi-Dirac distribution function for two different temperatures. The Boltzmann approximation is indicated as well (green line).

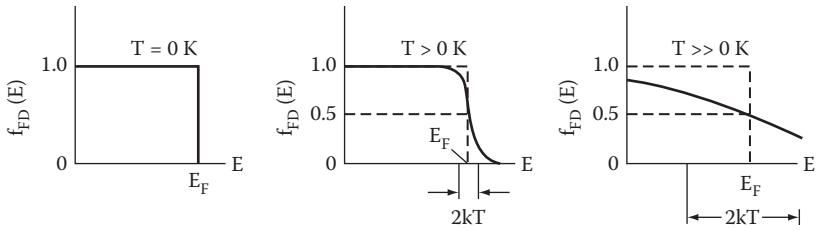


FIGURE 3.108 Evolution of Fermi distribution function as a function of temperature. At the highest temperatures the Fermi-Dirac function smears out and starts resembling a Boltzmann distribution.

and have either zero intrinsic spin or integral spin quantum numbers. Fermion and boson particle distributions are compared further below in Figure 3.108. The wave function that describes a collection of fermions must be antisymmetric with respect to the exchange of identical particles. One fermion of a system in a certain state prevents all other fermions from being in that state. The mathematical expression for the Fermi-Dirac distribution as a function of energy and temperature is:

$$f(E, T)_{FD} = \frac{1}{e^{\left[\frac{(E-\mu)}{kT}\right]}+1} \text{ or } \frac{1}{Ae^{\frac{E}{kT}}+1} \qquad (3.219)$$

where μ is the chemical potential and $A = e^{-\mu}$. At T = 0, one can see that $\mu(T = 0) = E_F$, the Fermi energy. Above T = 0 there is no abrupt energy that separates filled from unfilled levels, so the definition of the Fermi energy must be slightly modified. Note that when f(E, T) = 1/2, there is a 50-50 probability of finding the level occupied and $E = E_F$. Thus, at temperature *T*, the Fermi energy is defined as that energy for which the probability of being occupied is 0.5. The Fermi energy plays a very important role in the band theory of semiconductors and metals. In metals the Fermi energy is an effective cutoff level for the allowed energies of the electrons. By analogy, imagine a sea of electrons with a "depth" of E_F. At room temperature only a small fraction of the electrons will ever have an energy much above that sea level ($\approx 2kT_1$ at T_1). The energy E_F corresponds to the chemical potential, μ, i.e., the amount of energy needed to add an electron to the system ($E_F = \mu$). For fermions, the chemical potential may be either positive or negative. From thermodynamics, the chemical potential, and thus the Fermi energy, is related to the Helmholtz free energy: $\mu = F(n + 1) - F(n)|_{T,V}$, where F = U − TS. For comparison, we have indicated the Maxwell-Boltzmann distribution in Figure 3.107 as well (green line).

With $T > 0$ K, the $f(E, T)$ function has the general shape as sketched in Figure 3.108, but as the temperature keeps on increasing it gradually smears out, and finally at very high temperature ($T \gg 0$ K) it begins to look like an ordinary Boltzmann distribution as shown in Figure 3.108 (see also Figure 3.6). At ordinary temperatures, say 1000 K, kT = 0.088 eV, whereas for gold E_F = 5.51 eV and for sodium it is 3.12 eV-typical values for metals. In other words, $E_F \gg kT$, and an electron gas in a metal cannot usually be treated as an ordinary Maxwell-Boltzmann gas. With $E_F \gg kT$, the electron distribution is quantum-mechanical, and the electron gas is said to be degenerate. A Fermi gas well described by a Fermi-Dirac distribution that is approximately step-like is termed degenerate.

Remember that Drude used a Maxwell-Boltzmann distribution for the free electrons in a metal. From the above we now recognize that this was the wrong particle distribution to use. The Maxwell-Boltzmann distribution, shown in Figure 3.6, is a classical distribution of particles, but with particles satisfying quantum statistics, a Fermi distribution must be used. The electronic velocity vector component distribution for a single direction v_i according to Drude (based on Equation 3.17, which gives the distribution of speeds of molecules rather than their component velocities), is given as:

$$f_{MB}(v_i) = n\left(\frac{m}{2\pi k_B T}\right)^{\frac{3}{2}} e^{-\frac{\frac{1}{2}mv_i^2}{kT}} \qquad (3.220)$$

where n = N/V with *n* the electron density, *V* the volume, and *N* the sum of all electrons. The Maxwell-Boltzmann distribution assumes that the described particles are distinguishable. In other words, we can make a distinction as to which particle is in which energy state. For an ideal gas, that is certainly the case,

but this is not true for electrons, and this is the reason why Drude's model ultimately failed. Sommerfeld used the same classical gas as Drude but used the quantum Fermi-Dirac distribution for the electrons' velocity vector component in one direction (v_i):

$$f_{FD}(v_i) = \frac{(m/\hbar)^3}{4\pi^3} \frac{1}{\exp\left[\left(\frac{1}{2}mv_i^2 - kT_0\right)/kT\right] + 1} \tag{3.221}$$

The temperature T_0 is determined by the normalization $n = \int f(v_i)dv_i$.

We can understand now why the electron gas does not contribute to the heat capacity of metals. As a metal is heated, the only way an electron can take up energy is by moving into a somewhat higher allowed energy level. But a typical electron is buried deep inside the Fermi sea, and there are no empty levels to move to because each level above the electron is already occupied by a pair of electrons of opposing spin. Only the relatively few electrons at the top of the distribution can find empty levels to move into. These electrons find themselves in the so-called Maxwellian tail of the Fermi-Dirac function. They are the only electrons that contribute to the heat capacity, as we will calculate below. Thus, at ordinary temperatures the electronic heat capacity is almost negligible, as we can see from Figure 3.16. But why then can all the electrons in a free electron gas contribute to the electrical conductivity? We would expect that also in this case, electrons take energy from the electrical field and move into higher energy levels. What happens in this case, however, is a shift of the entire set of electron energy levels from lower to higher values. The Fermi distribution, in this case, as we will show below in Figure 3.121, is shifted bodily by the applied field. Thus, electrons can acquire an average drift velocity in the field without violating Pauli's exclusion principle.

Bose-Einstein Distribution

Bosons are a third type of particles that come with their own kind of distribution law, i.e., the Bose-Einstein distribution. Violations of the Maxwell-Boltzmann statistics are observed if the density of particles is very large (neutron stars), the particles are very light (electrons in metals, photons), or they are at very low temperatures (liquid helium). Classical and quantum mechanics particle distributions are compared in Table 3.10. From this table we recognize that with a very large E or small T all three distributions reduce to the classical Maxwell-Boltzmann form.

Whole-integer spin particles are called bosons, for which Bose-Einstein statistics are applicable. The wave function that describes a collection of bosons must be symmetric with respect to the exchange of identical particles. One boson of a system in a certain state increases the probability of finding another boson in this state. Satyendra Bose and Albert Einstein developed these statistics in the

TABLE 3.10 Classical and Quantum Distributions of Particles, with $\mu = E_F$ the Chemical Potential of the Particle

Distribution Name	Properties of the Particles	Examples	Distribution Function, F
Maxwell-Boltzmann	Spin does not matter. Unlimited number of particles per state. Particles are identical but distinguishable. Wave functions do not overlap.	Classical gas. Fermions and bosons at high T ($\mu - E >> kT$).	$F_{MB} = e^{\left(\frac{\mu - E}{kT}\right)}$ or $Ae^{\left(\frac{-E}{kT}\right)}$
Bose-Einstein	Boson particles are indistinguishable with integer spin (0, 1, 2…). Unlimited number of particles per state. Wave functions overlap.	Liquid ^{4}He, photons, Cooper pairs, excitons, Rb.	$F_{BE} = \frac{1}{e^{\left(\frac{E - \mu}{kT}\right)} - 1}$ or $\frac{1}{Ae^{\frac{E}{kT}} + 1}$
Fermi-Dirac	Fermion particles are identical with half-integer spins (1/2, 3/2, 5/2…). Wave functions overlap. Never more than one particle per state.	Free electrons in metals, protons, and neutrons; electrons in white dwarfs.	$F_{FD} = \frac{1}{e^{\left(\frac{E - \mu}{kT}\right)} + 1}$ or $\frac{1}{Ae^{\frac{E}{kT}} + 1}$

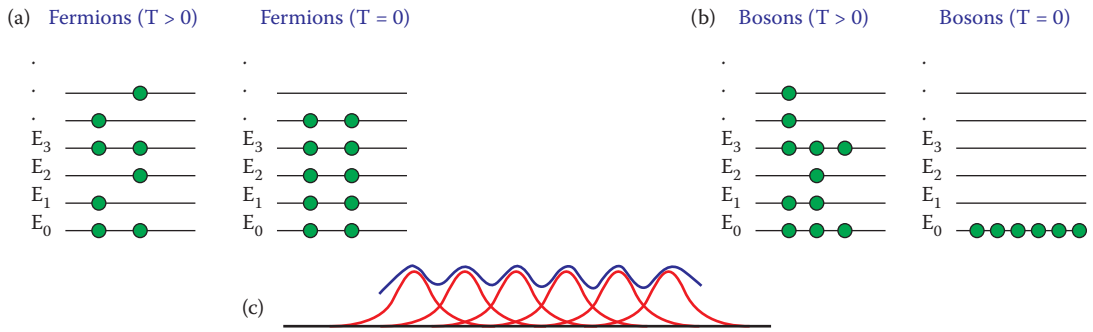


FIGURE 3.109 (a) Fermions at $T > 0$ and at T = 0 filling up a set of energy levels. (b) Bosons at $T > 0$ and T = 0. The lowest state is macroscopically populated. (c) The particles get so close together that their wave functions overlap, and the condensate is described by a single, macroscopic ψ.

1924–1925 time frame. Bose, a reader in physics at Dacca University, first derived Planck's blackbody law anew from a combination of light quanta of zero mass and statistics. By treating radiation as a quantum gas and counting particles instead of wave frequencies, Bose cut quantum theory loose from its classical antecedents. Then Einstein developed these statistics further and applied them to collections of atoms—gas or liquid—obeying the same rules, and he predicted the phenomenon of what later would be called the Bose-Einstein condensation (BEC) or *superatoms*. If one concentrates a large number of identical bosons in a small region at low temperatures, their wave functions start to overlap, and the bosons lose their individual identities and become one object. The BEC is a macroscopic matter wave, and it is formed of atoms that are delocalized and indistinguishable. In other words, order appears in momentum space (atoms adopt a collective behavior). In Figure 3.109a and b we compare how fermions and bosons fill up a set of energy levels and in c we illustrate how bosons, when they get close together, can be described by a single macroscopic wave function ψ. Einstein was himself not convinced that BEC could happen: "*Die Theorie ist schön, aber ist sie auch war*?" The Bose-Einstein distribution is given as:

$$f_{BE}(E) = \frac{1}{e^{\left(\frac{(E-\mu)}{kT}\right)} - 1} \text{ or } \frac{1}{Ae^{\frac{E}{kT}} - 1} \qquad (3.222)$$

where A = exp(−μ). The boson statistics would mark Einstein's last great contribution to quantum theory.

Bose-Einstein, Maxwell-Boltzmann, and Fermi-Dirac distributions are compared in Figure 3.110 for the same value of A = 1. It is clear that these

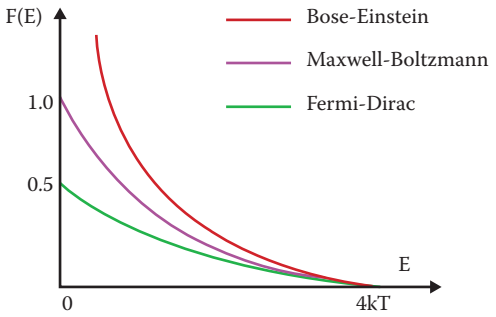


FIGURE 3.110 Bose-Einstein, Maxwell-Boltzmann, and Fermi-Dirac distributions compared.

distributions deviate the most at low temperatures, so one expects to see quantum phenomena emerge at low temperatures.

With bosons there is no exclusion principle, and a macroscopic number of bosons occupy the lowest energy quantum state. The blackbody law, for example, is a direct result of photons, which are bosons, all trying to get into the same energy state. Among other things, bosons also explain superconductivity, lasers (detailed in Chapter 5), and superfluidity in ^{4}He. A superfluid has no viscosity; in effect it experiences no friction and is able to flow forever while at the same time the thermal conductivity becomes very large!

Fermi Surfaces, Brillouin Zones, Density of State Functions, and Conductivity as a Function of Quantum Confinement

Introduction

In this section we consider how electronic processes, like charge transport, are affected by the dimensions of the solid. For this, we need to introduce the density of state functions for electronic systems with different levels of quantum confinement. The

density of state function, $G(E)$, is also called the *degeneracy* of a system and is often abbreviated as DOS. The DOS function describes the number of allowed energy states that are available in a system per unit energy and per unit volume (i.e., in units of number of states/eV/cm^3). Earlier, we showed that for a gas of molecules, the number of molecules per unit volume, i.e., density $n(E)$, that have energies between E and $E + dE$, is $n(E)dE$, where $n(E) = G(E)f_{MB}(E)$ with $f_{MB}(E)$, the Maxwell-Boltzmann distribution, representing the probability function of occupancy of a state with energy E. In the case of an electron gas, the number of electrons, $n(E)$, with energies between E and $E + dE$, is again given by the product of $G(E)f_{FD}(E)$ and a distribution function $f_{FD}(E)$, which in this case is the Fermi-Dirac distribution. Thus, combining the density of available states (DOS) with the Fermi function, one can calculate the density of filled states $n(E)$. In other words, we obtain a number for those charge carriers that can contribute to conductivity and charge transport.

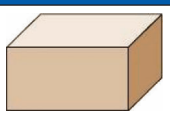
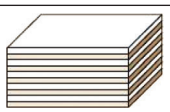
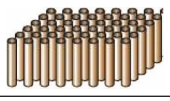
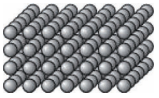
Systems with 3, 2, 1, and 0 degrees of freedom for electrons to move in are referred to as bulk material, quantum wells, quantum wires, and quantum dots, respectively, as summarized in Table 3.11. In this table, λ_{DB} represents the de Broglie wavelength (see Equation 3.216). It is the size of L_1, L_2, L_3 compared with λ_{DB} that determines whether we are dealing with a bulk semiconductor (L_1, L_2, $L_3 > \lambda_{DB}$), a quantum well (L_1, $L_2 > \lambda_{DB} > L_3$), a quantum wire ($L_1 > \lambda_{DB} > L_2$, L_3), or a quantum dot ($\lambda_{DB} > L_1$, L_2, L_3). We will see that the DOS function, $G(E)$, strongly depends on the degrees of freedom of the electrons in the system.

To calculate the charge carrier transport or current density ($\mathbf{J} = -ne\mathbf{v}$), using solids of different dimensionalities, we need to know the charge carrier density n as a function of dimensionality, but also the concept of Fermi surfaces needs to be introduced, and we need to get reacquainted with the Brillouin zones of Chapter 2. A Fermi surface is a surface of constant energy E_F in **k**-space, and at absolute zero temperature, it separates occupied from unoccupied quantum states. The way Fermi surfaces fill Brillouin zones determines whether a material will be a metal, semimetal, semiconductor, or an insulator. For example, if the first Brillouin zone is completely filled and there is a large gap between it and the next Brillouin zone, we have an insulator; a smaller gap makes for a semiconductor. If the first band is not completely filled or overlaps with an empty second Brillouin zone, a conductor or a semimetal results, respectively.

Bulk Materials: L_1, L_2, $L_3 > \lambda_{DB}$

Fermi Surface for Bulk Materials A Fermi surface is a surface of constant energy E_F in **k**-space, and at absolute zero temperature, it separates occupied from unoccupied quantum states. With a large number of electrons N at T = 0, the electrons will occupy the lowest energy states, consistent with the exclusion

Table 3.11 Examples of Reduced-Dimensional Material Geometries and Definitions of Their Dimensionality and of the Associated Type of Confinement*

$L_{1,2,3} > \lambda_{DB}$	No nanostructures	No confinement	Bulk material	
$L_{1,2} > \lambda_{DB} > L_3$	2D nanostructures	1D confinement	Wells	
$L_1 > \lambda_{DB} > L_{2,3}$	1D nanonstructures	2D confinement	Wires	
$\lambda_{DB} > L_{1,2,3}$	0D nanostructures	3D confinement	Dots	

*Here λ_{DB} represents the de Broglie wavelength.

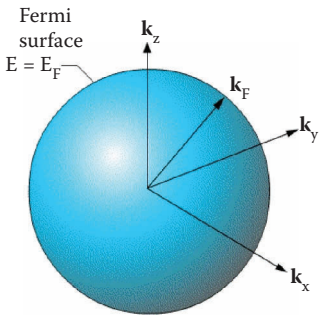


FIGURE 3.111 Free electron Fermi surface.

principle. Thus, the N electrons will fill up the lowest $N/2$ energy levels (two electrons per level). The energy of the last filled (or half-filled) level at T = 0 is called the Fermi energy, E_F. An unrestricted 3D Fermi electron gas is isotropic, so that a surface of constant energy E in **k**-space is a sphere with a surface called the Fermi surface (Figure 3.111). This surface separates the occupied from the unoccupied states at 0 K. At T = 0 K, all the free electron states are occupied up to an energy E_F, and all fall within a Fermi sphere with a Fermi wave vector, $\mathbf{k}_F$. When electrons are not free, a Fermi surface still exists, but it is distorted by the interaction with lattice ions (see, for example, Figure 3.117). The Fermi energy E_F is an effective cutoff level for the allowed energies of the electrons and corresponds to the chemical potential, μ, i.e., the amount of energy needed to add an electron to the system ($E_F = \mu$). The quantities $\mathbf{k}_F$ and E_F are called the Fermi wave vector and Fermi energy, respectively, in honor of Enrico Fermi. To calculate their exact value, we need to introduce the density of states function (see below, Equations 3.238 and 3.232, respectively).

Brillouin Zones for Bulk Materials Brillouin zones for crystals were first introduced in Chapter 2 in the context of x-ray diffraction. An x-ray diffraction pattern of a crystal represents a map of the reciprocal lattice, a Fourier transform of the lattice in real space, or also a representation of the lattice in **k**-space. Each Brillouin zone is a primitive cell of the reciprocal lattice, which corresponds to the Wigner-Seitz primitive cell of the real lattice (Figure 2.34). The boundaries of Brillouin zones satisfy the von Laue conditions for diffraction, and the Brillouin zone surface describes all **k** vectors that are constructively diffracted by the crystal. In other words, Bragg planes bound the Brillouin zones. This is true for the x-rays considered in Chapter 2 but also for the matter waves associated with electrons, neutrons, etc., discussed in this chapter. In discussing the Ewald sphere in Chapter 2 we also pointed out that the Bragg planes bisect the shortest vectors of the reciprocal lattice and that one can define the first Brillouin zone as the set of points in **k**-space that can be reached from the origin without crossing any Bragg planes. The second Brillouin zone is the set of points that can be reached from the first zone by crossing only one Bragg plane. The nth Brillouin zone can be defined as the set of points that can be reached from the origin by crossing $n-1$ Bragg planes, but no fewer (see the Brillouin zone for a square and a triangular 2D lattice in Figure 2.33). To get reacquainted with the procedure of constructing Brillouin zones, we start off here with a simple square lattice of atoms with interatomic distance, a. Its reciprocal lattice is also a square, with reciprocal lattice base vector of length $2\pi/a$. That reciprocal lattice is shown in Figure 3.112a, where we also have drawn the three shortest vectors marked $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$. Three lines, Bragg planes, are drawn as perpendicular bisectors of these three vectors. By drawing all the lines equivalent by symmetry with those bisector lines, we obtain the regions in **k**-space that constitute the first three Brillouin zones as marked (Figure 3.112b).

We turn now to a 3D crystal where we find again that a zone structure exists that may be different for different directions in **k**-space. The coordinates $\mathbf{k}_x$, $\mathbf{k}_y$, and $\mathbf{k}_z$ specify a point in **k**-space: a value of the vector from the origin that specifies the momentum of the electron. The zones in **k**-space corresponding to allowed energies for motion of an electron in a solid are 3D Brillouin zones forming nested sets of polyhedra. Different potentials exist in different directions, so the electron wavelength and crystal momentum, $\mathbf{k} = 2\pi/\lambda$, differ with direction, and many different parabolic E-**k** relationships exist, depending on the crystalline momentum. For a simple example, consider a cubic crystal with a Bragg reflection (Equation 2.20 with d = a) occurring when:

$$\mathbf{k}_x = \frac{2\pi n_x}{a},\ \mathbf{k}_y = \frac{2\pi n_y}{a},\mathbf{k}_z = \frac{2\pi n_z}{a},$$
$$n_x, n_y, n_z, = \pm 1, \pm 2, \pm 3, \ldots \qquad (3.223)$$

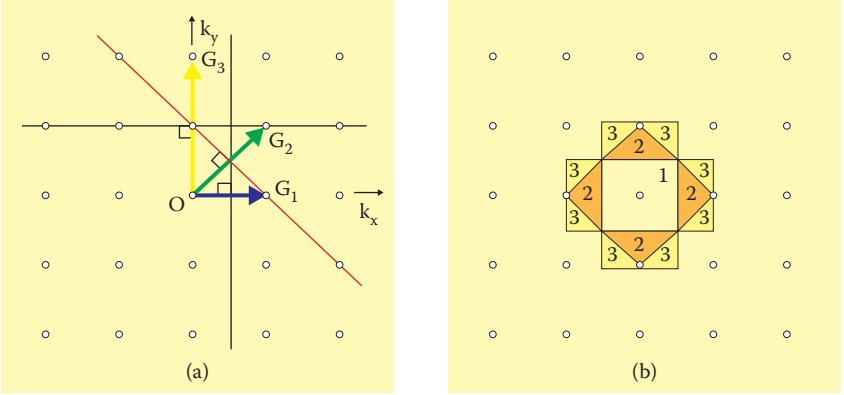


FIGURE 3.112 Illustration of the definition of the Brillouin zones for a 2D square Bravais lattice. To go from (a) to (b): the numbers in (b) denote the zone to which the region belongs. The numbers are ordered according to the length of the vector **G** used in the construction of the outer boundary of the zone (a).

These are the zone boundary values corresponding to discontinuities in the allowed energy levels. From Bragg's law, reflecting planes with the largest interplanar spacings (d = a) have the smallest values for **k**. The planes with the largest spacing are the {100} planes. Thus, the first Brillouin zone in **k**-space is bounded by the {100}* planes at $\mathbf{k}_x = \frac{\pm\pi}{a}$, $\mathbf{k}_y = \frac{\pm\pi}{a}$, $\mathbf{k}_z = \frac{\pm\pi}{a}$, which therefore is a cube with faces π/a from the origin as shown in Figure 3.113. The next set of reflecting planes are the {110} planes with spacing $d = a/2^{1/2}$. Corresponding planes in **k**-space are:

$$\pm\mathbf{k}_x \pm \mathbf{k}_y = \frac{2\pi}{a}, \quad \pm\mathbf{k}_x \pm \mathbf{k}_z = \frac{2\pi}{a}, \quad \pm\mathbf{k}_x \pm \mathbf{k}_z = \frac{2\pi}{a} \tag{3.224}$$

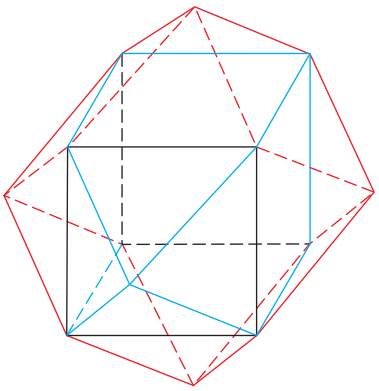


FIGURE 3.113 First two Brillouin zones for a simple cubic lattice. The first zone is the cubic volume $8\pi^3/a^3$ in **k**-space. The second zone is the **k**-space between the cube and the circumscribed dodecahedron.

* See Chapter 2 on crystallography on Miller indices.

These 12 planes outline the dodecahedron shown in the same figure. The second Brillouin zone is the **k**-space between the cube and the dodecahedron.

In Figure 3.114 we show what happens when one fills up the first Brillouin zone (BZ) of a face-centered cubic structure with electrons. In this case the largest spacing that satisfies the Bragg condition involves both {111} and {200} planes. These are the set of planes that would be reached first (largest a is smallest **k**). The zone boundaries then derive partly from planes derived from {200} and partly from planes derived from {111}. Envision a growing Fermi sphere inside this first Brillouin zone. Near the center of the zone the electrons are virtually free and behave like an electron gas with an energy given by Equation 3.144. At these energies, well below the zone boundaries, the surfaces of constant energies are represented as spheres within the Brillouin zone as drawn in Figure 3.114a. But when adding more and more electrons, we approach the diffraction

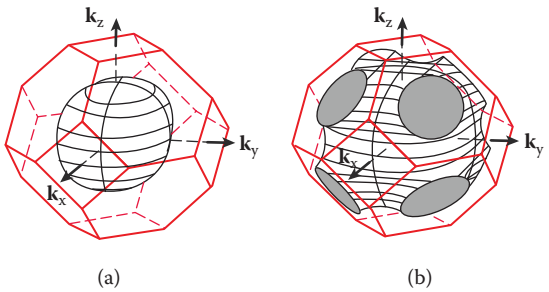


FIGURE 3.114 (a) The first Brillouin zone of a face-centered cubic (FCC) structure with surfaces of constant energy of electrons shown for nearly free electrons near the bottom of the zone and (b) electrons at the zone boundary.

boundaries, and the wave vectors near or at the BZ feel the periodic potential of the crystal, whereas the others do not. All wave vectors that end on a BZ will fulfill the Bragg condition and thus are diffracted. Because electrons cannot transgress this boundary, the spherical surface becomes distorted by bulging out toward the zonal plane (Figure 3.114b). These electrons no longer behave as a free electron gas. Wave vectors completely in the interior of the first BZ, or between any two BZs, will never get diffracted; they move pretty much as if the potential would be constant; i.e., they behave very close to the solutions of the free electron gas.

If the first Brillouin zone is completely filled and there is a considerable gap between it and the next Brillouin zone, we have an insulator at hand. With a smaller gap, we are dealing with a semiconductor. If the first band is not completely filled or overlaps with an empty second Brillouin zone, we are dealing with a conductor or a semimetal, respectively.

Some other jargon often associated with the difference between a metal, a semiconductor, and an insulator is the characterization of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). If the difference in the energy between the HOMO and the LUMO is zero, then just a little energy can promote an electron in an unoccupied level. Therefore, with an electrical potential difference, some electrons are very mobile and give rise to electrical conductivity in a metal. When the temperature increases, there are more electrons excited toward empty orbitals. However, the conductivity decreases because the vibration of the nuclei increases the collisions between the transported electrons and the nuclei, so there is a less-efficient transport or the resistance increases. When the HOMO-LUMO energy difference is nonzero, there is an electronic gap. If the bandgap is small, thermal excitations can promote electrons to unoccupied levels; consequently, those electrons can contribute to the electrical conductivity. This is the case of semiconductors, and that is why the conductivity of semiconductor increases with temperature. Insulators are characterized by a huge HOMO-LUMO energy difference, and the electrons cannot reach the unoccupied levels; there is no measurable conductivity.

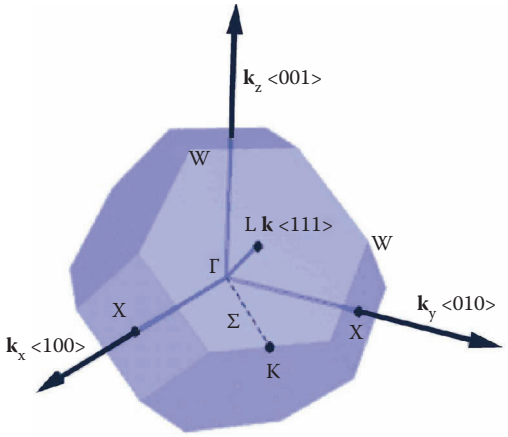


FIGURE 3.115 First Brillouin zone for Si. Points of high symmetry on the Brillouin zone have specific importance. The most important point for optoelectronic devices is the center at $\mathbf{k} = 0$, known as the gamma point Γ (000). One finds *X* at the surface boundaries (100), (001), and (010); *K* at (110); and *L* at (111). *K* is in the middle of an edge joining two hexagonal faces, *L* is at the center of a hexagonal face, and *W* is a corner point. Σ means directed from Γ to *K*.

In Figure 3.115 we show the first Brillouin zone for Si, a semiconductor. Letters are used to mark many of the high symmetry points on this first BZ boundary. The gamma point (Γ) is always the zone center, where $\mathbf{k} = 0$ ($\mathbf{k}_x = \mathbf{k}_y = \mathbf{k}_z = 0$), the *X* point is at $\mathbf{k}_x = 2\pi/a$ and $\mathbf{k}_y$ and $\mathbf{k}_z = 0$ (center of a square face), and the *L* point at $\mathbf{k}_x = \mathbf{k}_y = \mathbf{k}_z = \pi/a$ (center of a hexagonal face) (*a* is the lattice constant, i.e., cube edge). *K* is in the middle of an edge joining two hexagonal faces, *L* is at the center of a hexagonal face, and *W* is a corner point. Σ means directed from Γ to K. Most semiconductors have band edges of allowed bands at one of these points. Zone edges or surfaces are marked with symbols from the Roman alphabet, whereas the interior is marked with symbols from the Greek alphabet. Two example Greek letters not yet introduced are: Δ which means directed from Γ to *X* and Λ which means directed from Γ to L. These symbols are not marked in Figure 3.115 where we only introduced Γ and Σ but they are used in Figure 3.116. In Figure 3.116 the energy bands in 2D are plotted along the major symmetry directions in the first BZ for two of the industrially most important semiconductors, namely, Si and GaAs.

Group IV and III–V semiconductors have a band structure that appears somewhat similar because their basic character is controlled by sp^3 hybridization and tetrahedral bonding. From Figure 3.116a, in the

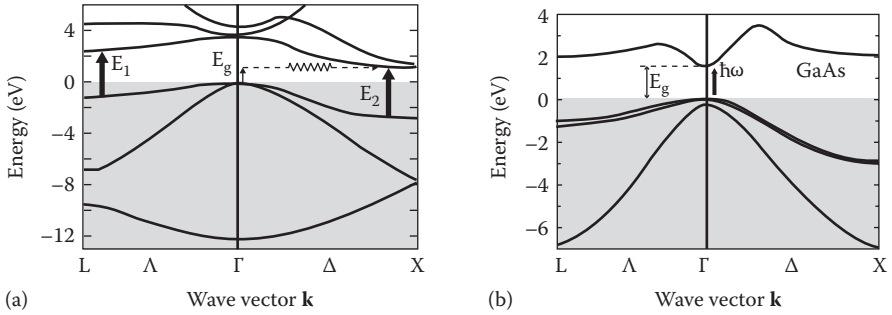


FIGURE 3.116 (a) Si is an indirect semiconductor because the maximum of the valence band (E_v at Γ) does not coincide with the minimum of the conduction band (E_c at X). For silicon, the valence band maximum (VBM) occurs at $\mathbf{k} = 0$ (Γ point). However, the conduction band minimum (CBM) occurs to the left of *X*. This makes Si an indirect bandgap material. The shortest distance between E_c and E_v is the bandgap E_g of Si: 1.1 eV (a). GaAs is a direct bandgap semiconductor because the VBM and CBM occur at the same point in the Brillouin zone, and it features a strong absorption. For GaAs, $E_g = 1.5$ eV (b).

case of silicon, the valence band maximum (VBM) occurs at $\mathbf{k} = 0$ (Γ point). However, the conduction band minimum (CBM) occurs to the left of *X*. Thus, the VBM and the CBM occur at different points in the Brillouin zone, and this makes Si an indirect bandgap material (with a bandgap $E_g = 1.1$ eV). For an electron to move from the VBM to the CBM, its **k** vector needs to be changed, making the process less likely because a phonon is required for the electron to change its momentum (three-particle process: electron, photon, and phonon). Its indirect bandgap makes Si a weak absorber of light and thus a poor optoelectronic material. The Si band diagram also has two critical points at E_1 (3.2 eV) and at E_2 (4.3 eV). Critical points (van Hove singularities) occur whenever $dE/dk = 0$, in other words, when the bands are parallel to each other. At those critical points, band transitions are more likely because no momentum change is required (parallel band effect). The Si band structure also features six equivalent conduction band minima at *X* along six equivalent <100> directions, and the valence band maxima for the heavy-hole, light-hole, and split-off bands are located exactly at the Γ point. Germanium and diamond also are indirect bandgaps. GaAs is a direct bandgap semiconductor (Figure 3.116b) because the VBM and CBM occur at the same point in the BZ, and it features a strong absorption at $\hbar\omega > E_g$ (for GaAs, $E_g = 1.5$ eV), making it an excellent optoelectronic material.

The Fermi surface for Cu, shown in Figure 3.117, is a sphere entirely contained within the first Brillouin zone. Even if the free electron Fermi sphere does not intersect a BZ boundary, its shape can still be affected at points close to the boundary where the energy bands begin to deviate from the free electron parabolic shape. This is the case with Cu, where in the <111> directions, contact is made with the hexagonal Brillouin zone faces. Eight short "necks" reach out to touch the eight hexagonal zone faces. The Fermi surface shown in Figure 3.117 can extend throughout many unit cells if the necks in the <111> directions are joined together with similar surfaces in adjacent cells. A section of such a continuous zone structure is shown in Figure 3.118. A magnetic field may close off these necks, a phenomenon that may be studied using cyclotron resonance. These necks are clearly evident from de Haas-van Alphen oscillations (http://www.lanl.gov/orgs/mpa/nhmfl/users/pages/deHaas.htm and http://physics.binghamton.edu/Sei_Suzuki/pdffiles/Note_dHvA.pdf) for magnetic fields in the <111> directions, which contain two periods, determined by the extremal "belly" (maximum) and "neck" (minimum) orbits. The de Haas-van Alphen

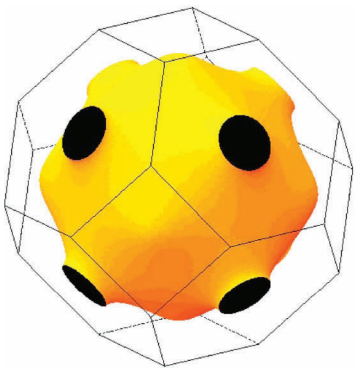


FIGURE 3.117 Brillouin zone for Cu with free electron sphere bulging out in the <111> directions to make contact with the hexagonal zone faces.

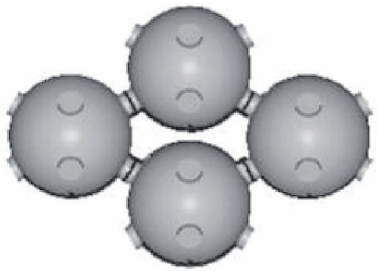


FIGURE 3.118 Continuous Fermi surfaces extending through adjacent unit cells in gold structures.

effect has its origin in the Bohr-Sommerfeld quantization of the orbits of conduction electrons under the influence of a magnetic field. A measurement of the temperature dependence of the oscillation amplitude permits a determination of the cyclotron frequency or, equivalently, the electron mass.

Density of States for Bulk Materials The DOS function $G(E)$ is a property that quantifies how closely packed energy levels are in some physical system. It is usually expressed as a function of internal energy E, $[G(E)]$, or as a function of the wave vector $\mathbf{k}$, $[G(\mathbf{k})]$. The density of $\mathbf{k}$-states, $G(\mathbf{k})$, in the Fermi sphere drawn in Figure 3.111 is the number of allowed states between $\mathbf{k}$ and $\mathbf{k} + d\mathbf{k}$, i.e., the number of states between a sphere of radius $\mathbf{k}$ and a sphere of radius $\mathbf{k} + d\mathbf{k}$ (analogous to the calculation of the density of states for gas molecules illustrated in Figure 3.7). In three dimensions the volume between the two shells, $\mathbf{V}$ is given by:

$$\mathbf{V}d\mathbf{k} = 4\pi\mathbf{k}^2 d\mathbf{k} \tag{3.225}$$

The density of states $G(\mathbf{k})_{3D}$ is then derived by dividing this volume by the volume of a single energy state. From Figure 3.66b we recognize that each value of $\mathbf{k}$ occupies a volume $V = (2\pi/L)^3$, so that the number of states per unit volume of $\mathbf{k}$-space is $1/V$ or $(L/2\pi)^3$. We also introduce a factor of two here from Pauli's exclusion principle: each energy state can accommodate two electrons, or:

$$G(\mathbf{k})_{3D}\,d\mathbf{k} = 2\frac{4\pi\mathbf{k}^2 d\mathbf{k}}{\left(\frac{2\pi}{L}\right)^3} = \frac{L^3\mathbf{k}^2}{\pi^2}d\mathbf{k} \tag{3.226}$$

To obtain the density of states in terms of the energy $[G(E)_{3D}]$, we use the relation between E and $\mathbf{k}$, derived earlier as $\mathbf{k} = \left(\frac{2m_e^* E}{\hbar^2}\right)^{1/2}$ (Equation 3.150). For every $\mathbf{k}$-state there is a corresponding energy state. Differentiating the latter expression with respect to energy leads to:

$$d\mathbf{k} = \left(\frac{2m_e^* E}{\hbar^2}\right)^{-\frac{1}{2}} \frac{m_e^*}{\hbar^2} dE \tag{3.227}$$

The density of energy states $G(E)_{3D}$, the number of allowed states between E and $E + dE$, is then obtained using the chain rule and dividing by L^3 (=V) because we want an expression per unit energy and unit volume:

$$G(E)_{3D} = G(\mathbf{k})_{3D}\frac{d\mathbf{k}}{dE} = \frac{k^2}{\pi^2}\frac{d\mathbf{k}}{dE} \tag{3.228}$$

Thus, we calculate the density of states for a parabolic band in a bulk material (three degrees of freedom) as:

$$G(E)_{3D}\,dE = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}}dE \tag{3.229}$$

for $E \geq 0$. There are no available states at $E = 0$, and the effective mass of the electron takes into account the effect of the periodic potential on the electron. The minimum energy of the electron is the energy at the bottom of the conduction band (CB), E_c. For the valence band (VB), similar results hold. In the case that the origin of the energies is not chosen to be the bottom of the band (i.e., $E_c \neq 0$), for the conduction band Equation 3.229 comes with an energy term $(E - E_c)^{1/2}$, and in the case of the valence band the energy term is given as $(E_v - E)^{1/2}$. Here the contributions from the light- and heavy-hole bands add to give the total DOS. Note that, just like the DOS in a 3D ideal gas (Equation 3.19), we obtain a square root dependence in energy E. This function is illustrated in Figure 3.119 (full line). Thus, in the case of free 3D motion, the electronic DOS has a smooth square root dependence on energy E with three continuously varying wave vectors $\mathbf{k}_x$, $\mathbf{k}_y$, and $\mathbf{k}_z$. This holds for materials that are large compared with the de Broglie wavelength (L_1, L_2, $L_3 > \lambda_{DB}$). An important feature of nanostructure devices is that their DOS is very different from this expression, and

G(E)dE =

$\frac{1}{V}$ × [the number of states in the energy range from E to E + dE]

G(E)f(E)dE =

$\frac{1}{V}$ × [the number of filled states in the energy range from E to E + dE]

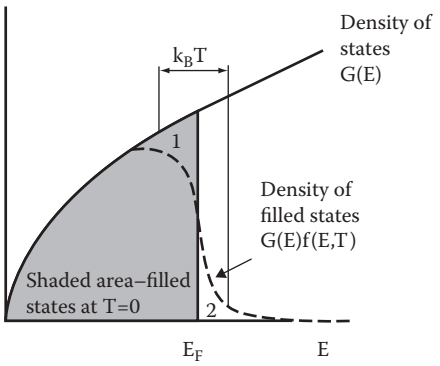


FIGURE 3.119 Density of states (DOS) (full line) and occupied states (shaded area filled at T = 0 K) around the Fermi energy [$G(E)_{3D}$]. The density of filled states is marked with a broken line. As the temperature increases above T = 0 K, electrons from region 1 are excited into region 2.

this will turn out to have important consequences for their electrical and optical properties.

Equation 3.229 gives us the number of possible electronic states of a large 3D device (L_1, L_2, $L_3 > \lambda_{DB}$). To deduce how these states are occupied with electrons, we need to multiply the density of state function [$G(E)_{3D}$] with the Fermi-Dirac distribution function, $f_{FD}(E, T)$, which gives the probability that a state of energy E is occupied by an electron. The number of electrons with energies between E and $E + dE$ is then written as:

$$n(E)_{3D}dE = G(E)f_{FD}(E)dE \tag{3.230}$$

where $G(E)_{3D}dE$ is the number of states between E and $E + dE$. This function is shown as a broken line in Figure 3.119.

The integral of $n(E)_{3D}dE$ over all energies gives the total number of electrons n_{3D} per unit volume (=n_{3D}/V). From Equation 3.230, with f(E) = 1 (T = 0 K), the number density of filled states n_{3D} is:

$$n_{3D} = \int_0^{\infty} n(E)_{3D}dE = \int_0^{E_F} G(E)_{3D}f(E)dE \text{ or}$$
$$n_{3D} = \int_0^{E_F} \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}}dE = \frac{\left(\frac{2m_e^*E_F}{\hbar^2}\right)^{\frac{3}{2}}}{3\pi^2} \tag{3.231}$$

Note that at T = 0, $n(E)_{3D}$ is zero for $E > E_F$, so we only have to integrate from E = 0 to E = E_F. In Figure 3.119 this integral corresponds to the shaded area.

For a metal with a total number n_{3D} of valence electrons per unit volume, we can now calculate the maximum energy (E_F) and the maximum **k** value ($\mathbf{k}_F$). The maximum energy (the energy at the surface of the sphere) at T = 0 K is obtained by solving for E_F:

$$E_F = \frac{\hbar^2}{2m_e^*}\left(3\pi^2 n_{3D}\right)^{\frac{2}{3}} = (0.365\ \text{eV nm}^2)(n_{3D})^{\frac{2}{3}} \tag{3.232}$$

This Fermi level is the top of the collection of electron energy levels at absolute zero temperature, and it depends on the number of electrons per unit volume (n_{3D}). Example numbers for Fermi energies for different metals can be found in Table 3.13.

Because fermions cannot exist in identical energy states (see the exclusion principle), at absolute zero, electrons pack into the lowest available energy states and build up a "Fermi sea" of electron energy states. In this state (0 K), the average energy per electron in a 3D electron gas is calculated as:

$$E_{av} = \frac{\text{Total Energy}}{\text{Number of Electrons}} = \frac{\int_0^{E_F} EG(E)_{3D}dE}{\int_0^{E_F} G(E)_{3D}dE}$$
$$= \frac{1}{n_{3D}}\int_0^{E_F} EG(E)_{3D}dE = \frac{3}{5}E_F \tag{3.233}$$

For Cu, for example, this average energy is 4 eV, huge compared with typical thermal energies of 0.025 eV (kT at 300 K). Contrast this result with that of a gas of molecules where a Boltzmann distribution leads to an energy E of zero at 0 K and of the order of kT at a temperature T!

The total number N of **k**-states within the Fermi sphere can be calculated as:

$$N = \frac{\frac{4}{3}\pi k_F^3}{\left(\frac{2\pi}{L}\right)^3} = \frac{k_F^3}{6\pi^2}v \tag{3.234}$$

because each **k**-state occupies $(2\pi/L)^3$ and the total volume of the Fermi sphere equals $\frac{4}{3}\pi k_F^3$. The ratio of

these two gives the total number of states N, each of which accommodates two electrons; in other words, the density of states $n_{3D}(=N/V)$ at the Fermi level is:

$$n_{3D} = \frac{2N}{V} = \frac{k_F^3}{3\pi^2} = \frac{\left(\frac{2m_e^* E_F}{\hbar^2}\right)^{\frac{3}{2}}}{3\pi^2} \qquad (3.235)$$

giving us back Equation 3.231. Thus, the density of single-particle states available per unit volume per unit energy of states at the Fermi level in an unrestricted (3D) electronic device is derived as:

$$G(E)_{3D}(E = E_F) = \left.\frac{dn_{3D}}{dE}\right|_{E=E_F} = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} E_F^{\frac{1}{2}} \qquad (3.236)$$

This is a very important quantity as the rates of many processes are proportional to $G(E)$. Increasing n_{3D}, the density of electrons in 3D increases both E_F (Equation 3.232) and $G(E)_{3D}$ (Equation 3.236). This is of course the same outcome as obtained in Equation 3.229, the only difference being that we replaced E with E_F here.

The result in Equation 3.236 can be simplified by comparing it with Equation 3.235 to obtain:

$$G(E)_{3D}(E = E_F) = \left.\frac{dn_{3D}}{dE}\right|_{E=E_F} = \frac{3n_{3D}}{2E_F} \qquad (3.237)$$

What this means is that the density of single-particle states available per unit volume and per unit energy of states at the Fermi level in 3D is 1.5 times the density of conduction electrons divided by the Fermi energy.

Example 3.7: Calculate the Fermi energy E_F of Na. Na has an atomic density of 2.53×10^{28} atoms/m^3. We are assuming $m_e = m_e^*$, i.e., $h^2/2m_e = 1.505$ eV.nm^2.

Answer: Sodium has one valence electron ($3s$) per atom, so the electron density is $n = 2.53 \times 10^{28}$/m^3. Using Equation 3.232 we obtain:

$$E_F = \frac{1}{4\pi^2}\frac{h^2}{2m_e}\left(3\pi^2 n_e\right)^{\frac{2}{3}}$$

$$= \frac{1}{4\pi^2}\left(1.505 \text{ eV} \cdot \text{nm}^2\right)\left(3\pi^2 \times 25.3 \text{ nm}^{-3}\right)^{\frac{2}{3}} = 3.31 \text{ eV}$$

Obviously electrons in a metal have a very large kinetic energy, even at T = 0!

From Equation 3.235, $\mathbf{k}_F$ is given by:

$$\mathbf{k}_F = (3n\pi^2)^{\frac{1}{3}} \qquad (3.238)$$

which depends only on particle concentration n.

We can now also ask, what is the speed in the highest occupied state E_F? From $\mathbf{k}_F$, the Fermi wave vector, we calculate the velocity $\mathbf{v}_F$ of the electrons on the Fermi surface, i.e., the Fermi velocity (see Equation 3.203, with group velocity $\mathbf{v}_g = \mathbf{v}_F$) as:

$$\mathbf{v}_F = \frac{\mathbf{p}_F}{m_e} = \frac{\hbar \mathbf{k}_F}{m_e} = \sqrt{\frac{2E_F}{m_e^*}} \qquad (3.239)$$

The Fermi velocity is the average velocity of an electron in an atom at absolute zero. This average velocity corresponds to the average energy given above (Equation 3.233). In Equation 3.239, $\mathbf{p}_F$ is the Fermi momentum, i.e., the momentum of fermions at the Fermi surface, and with the expression for $\mathbf{k}_F$ in Equation 3.238 we obtain:

$$\mathbf{v}_F = \frac{\hbar}{m}\left(3\pi^2 n_{3D}\right)^{\frac{1}{3}} \qquad (3.240)$$

We see that in 3D, the higher the electron density, the faster the electrons are moving. In the presence of an electrical field, all the electrons in a conductor move together, so the exclusion principle does not prevent the free electrons in filled states from contributing to the conduction. This is illustrated in Figure 3.120, where we show the Fermi function in one dimension versus velocity at an ordinary temperature. Over a wide range of velocities, the Fermi function equals 1, and speeds $\mathbf{v}_F$ marked in this figure are given by Equations 3.239 and 3.240.

The dashed curve represents the Fermi function after the electric field has been applied for a period t.

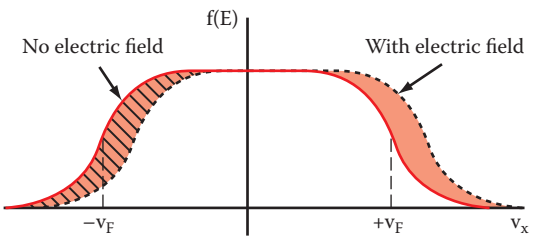


FIGURE 3.120 The Fermi function versus velocity in one dimension for a conductor with (solid) and without (broken line) electrical field in the +x-direction. The effect is greatly exaggerated (see text).

Table 3.12 Fermi Quantities*

Fermi Quantity	Equation 3D	Typical Value
Fermi wave vector	$\mathbf{k}_F = (3\pi^2 n)^{\frac{1}{3}}$	~10^8 cm^{-1}
Fermi energy	$E_F = \frac{\hbar^2}{2m_e^*}(3\pi^2 n)^{\frac{2}{3}}$	~1–10 eV
Fermi wavelength	$\lambda_F = 2\pi/\mathbf{k}_F$	Few nanometers for metals, several tens of nanometers for semiconductors
Fermi temperature	$T_F = E_F/k_B$	~10^4–10^5 K
Fermi momentum	$p = \hbar k_F$	
Fermi velocity	$v_F = \frac{\mathbf{p}}{m_e^*} = \frac{\hbar k_F}{m_e^*}$	~10^8 cm/s
*All values at T = 0 K.		

All the electrons have been shifted to higher velocities, but the net effect is equivalent to shifting electrons near the Fermi level only.

A typical value for the Fermi velocity at 0 K is ~10^6 m/s or 1000 km/s; as pointed out before, this is a surprising result because for a classical gas at 300 K the thermal velocity $\mathbf{v}_{rms} = \sqrt{\frac{8\mathbf{k}_B T}{\pi m}} = 10^5$ m/s (see Equation 3.21) with a velocity that goes to zero at T = 0 K!

The Fermi wavelength λ_F is given as $2\pi/\mathbf{k}_F$; it represents the de Broglie wavelength associated with the Fermi wave vector $\mathbf{k}_F$. The Fermi temperature is the temperature T_F at which $k_B T_F = E_F$. Thus, it is the energy of the Fermi level of an assembly of fermions divided by Boltzmann's constant. The quantity T_F is not to be confused with the temperature of the electron gas. Below the Fermi temperature, a substance gradually expresses more and more quantum effects of cooling. For temperatures much lower than the Fermi temperature, the average energy of the phonons of the lattice will be much less than the Fermi energy, and the electron energy distribution will not differ greatly from that at T = 0. We also recognize here that photons travel much faster than electrons! Photons travel at c = 3.0 × 10^8 m/s (in vacuum) versus electrons that travel at $\mathbf{v}_F$ (Fermi speed) = 1.57 × 10^6 m/s (copper wire).

The Fermi quantities we introduced in this section are summarized in Table 3.12, and in Table 3.13 we list calculated free electron Fermi surface parameters for some metals at room temperature.

Electronic Conductivity for Bulk Materials Drude, at the end of the nineteenth century, could not have known about the Fermi-Dirac distribution. Electrons, being fermions, follow this distribution, and at room temperature it is almost the same as at absolute zero temperature, resulting in a velocity distribution for fermions very different from the one predicted by Maxwell-Boltzmann statistics. Only electrons near the Fermi level contribute to electrical conductivity, and as a result of the wave nature of the electrons, they can pass through a perfect crystal without suffering any resistance at all. This means that the mean free path of an electron passing through a perfect crystal with all nuclei at rest is infinity rather than the interatomic spacing of the order of 1 nm assumed by Drude. In such an ideal crystal, a Bloch function ψ_k evolves into $\psi_{k+\Delta k_x}$, and

Table 3.13 Calculated Free Electron Fermi Surface Parameters for Metals at Room Temperature

Metal/Valency	Electron Concentration n (cm^{-3})	Fermi Wave Vector (cm^{-1})	Fermi Velocity (cm s^{-1})	Fermi Energy E_F (eV)	Fermi Temperature $T_F = E_F/k_B$ (in K)
Cu(1)	8.45×10^{22}	1.36×10^8	1.57×10^8	7.00	8.12×10^4
Ag(1)	5.85	1.20	1.39	5.48	6.36
Au(1)	5.90	1.20	1.39	5.51	6.39
Be(2)	24.2	1.93	2.23	14.14	16.41
Mg(2)	8.60	1.37	1.58	7.13	8.27
Zn(2)	13.10	1.57	1.82	9.39	10.90
Al(3)	18.06	1.75	2.02	11.63	13.49
Ga(3)	15.30	1.65	1.91	10.35	12.01
In(3)	11.49	1.50	1.74	8.60	9.98
Pb(4)	13.20	1.57	1.82	9.37	10.87
Sn(4)	14.48	1.62	1.88	10.03	11.64

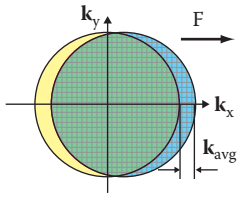


FIGURE 3.121 Displacement of the Fermi surface with an applied electrical field. Collisions with thermal vibrations and defects (not stationary ions or other electrons, as Drude envisaged) stop the Bloch oscillations and cause electrons to settle to a drift velocity. See also Figure 3.71, where the Fermi factor in one dimension for a conductor with (solid) and without (broken line) electrical field in the +*x*-direction is shown.

when this electron state reaches the Brillouin zone at $k = +\pi/a$, it re-enters the crystal at $-\pi/a$ or we get resistance-less Bloch oscillations.

In the absence of an electric field, the same number of electrons is moving in the ±*x*-, ±*y*-, and ±*z*-directions, so the net current is zero. But when a field E is applied, e.g., along the *x*-direction, the Fermi sphere in Figure 3.121, in the absence of collisions, is displaced at a uniform rate by an amount related to the net change in momentum, $\Delta\mathbf{p}_x$, of the free electron gas (FEG) as a whole. The equation of motion (Newton's law) describes this situation as (see Equation 3.208):

$$m_e^* \frac{dv_x}{dt} = \hbar\left(\frac{dk_x}{dt}\right) = -E_x e = F \qquad (3.241)$$

The quantity $\hbar\mathbf{k}$ is the crystal momentum, and thus one can say that the force caused by the electric field is equal to the time derivative of the crystal momentum. Integrating the previous expression we obtain:

$$\mathbf{k}_x(t) - \mathbf{k}_x(0) = -\frac{eE_x t}{\hbar} \qquad (3.242)$$

The shift in Fermi sphere creates a net current flow because more electrons move in the +*x*-direction than the –*x*-direction. According to this model, the Fermi sphere moves with constant velocity in **k**-space. This means that the electron velocity increases indefinitely. This is of course not possible, and it is evident that scattering processes must limit the electron velocity and hence the finite electrical conductivity of metals. A viscous term must be introduced in the equation of motion. In a real crystal, scattering events with a scattering time τ, involving collisions of electrons near the Fermi surface (these are the only ones that can move into empty states), prevent the observation of the ideal Bloch oscillations and oppose the electrical field effect so that the Fermi sphere reaches a steady state when the new center is displaced in the *x*-direction by an average wave vector:

$$\mathbf{k}_{avg} = -\frac{eE_x\tau}{\hbar} \qquad (3.243)$$

The wave vector changes calculated from Equation 3.243 are very small changes; for example, with an electric field *E* of 1 V/m and with a value for the scattering time τ of ~10^{-14} s we calculate a value for $|\mathbf{k}_{avg}|$ of ~15 m^{-1}, a quantity very small compared with a BZ dimension, which we calculate as:

$$\mathbf{k}_{BZ} \approx \frac{2\pi}{a} \approx \frac{2\pi}{3.00\times10^{-10}\,m} \approx 2\times10^{10}\ m^{-1}$$

The effect of the Fermi shift in Figure 3.121 are greatly exaggerated. If at t = 0 an electrical field is applied to a Fermi sphere centered at the origin of **k**-space, the sphere will move to a new position in a characteristic time between scattering events given by:

$$\tau = \frac{\lambda}{v_{avg}} \qquad (3.244)$$

From Equation 3.243 the average velocity of the Fermi sphere is given by:

$$\mathbf{v}_{avg} = \frac{\hbar\mathbf{k}_{avg}}{m_e^*} = -\frac{eE_x\tau}{m_e^*} \qquad (3.245)$$

At the steady state the current density is then given as:

$$\mathbf{J} = -ne\mathbf{v}_{avg} = \left(\frac{ne^2E_x\tau}{m_e^*}\right) = \sigma E \qquad [3.3][3.6]$$

where *n* is the electron density and:

$$\sigma = \frac{ne^2\tau}{m_e^*} \qquad (3.7)$$

These are the same expressions as derived in Drude's model, but with τ given now as $\tau = \frac{\lambda}{v_{avg}}$ (Equation 3.244) and m_e^* replacing m_e. The average velocity,

derived as $\bar{v} = \sqrt{\frac{8k_B T}{\pi m}}$ (Equation 3.21) from the Boltzmann distribution of speeds in Drude's model, is replaced here by $\mathbf{v}_{avg} = \frac{\hbar \mathbf{k}_{avg}}{m_e^*} = -\frac{eE_x\tau}{m_e^*}$ (Equation 3.245). In Drude's model the mean free time between collisions of electrons with lattice ions, τ, is related to the average velocity in $\bar{v} = \mathbf{v}_{dx} = -\frac{eE_x\tau}{m_e}$ (Equation 3.6) as:

$$\tau = \frac{\lambda}{v_x} \approx \frac{a}{v_x} \tag{3.246}$$

Drude assumed the electron mean free path, λ, to be equal to the lattice constant, *a*, which is of the order of 1 nm, in which case this equation yields a typical value for τ of about 10^{-14} s. Based on Equations 3.7, 3.21, and 3.246, Drude then derived the following relationship for the resistivity:

$$\rho = \frac{1}{\sigma} = \frac{m_e \mathbf{v}_d}{ne^2\lambda} = \frac{m_e}{ne^2 a}\sqrt{\frac{8kT}{\pi m_e}} \tag{3.22}$$

Using the lattice constant, *a*, for the mean free path, this equation leads to values for the conductivity of a metal that are six times too small. Moreover, from Equation 3.22 the temperature dependence of the resistivity is determined by v_{dx}, which in this model is proportional to $\sqrt{T}$, whereas in practice the temperature dependence of the resistivity is represented by the empirical relationship:

$$\rho = \rho_0 + \alpha T \tag{3.23}$$

where ρ_0 is the resistivity at a reference temperature, usually room temperature, and α is the temperature coefficient. If the Boltzmann distribution function is applied to the electron gas, one thus immediately finds the velocity of the electron to change as $\sqrt{T}$. According to Drude's model, the mean free path is obviously temperature-independent because it is calculated from the scattering cross-section of rigid ions (with lattice constant *a*). This results in a resistivity proportional to $\sqrt{T}$, provided that the number of electrons per unit volume *n* is temperature-independent (Equation 3.22). However, people at that time had been well aware that the resistivity of typical metals increases linearly with increasing temperature well above room temperature (Equation 3.23). To be consistent with the Maxwell–Boltzmann distribution law, one then had to assume *n* to change as $\frac{1}{\sqrt{T}}$ in metals. This was not physically accepted, and the application of the Maxwell–Boltzmann distribution to the electron system was apparently the source of the problem.

In the quantum mechanical model, only a *few electrons*, all moving at the temperature-independent very high Fermi velocity $\mathbf{v}_F \sim 10^8$ cm/s (Equation 3.239), carry the current, instead of *all electrons* moving at the average drift velocity $\mathbf{v}_d$ (~0.1 cm/s) in Drude's model (see p. 82 "Drude Fails"). Only electrons near the surface of the Fermi sphere find empty orbitals in which they can scatter. Electrons in the inner part of the Fermi sphere find no empty states with similar energy as those electrons near the Fermi sphere. Therefore, inner electrons cannot scatter, and those electrons do not contribute to the current transport process. Replacing v_{th} with v_F in Equation 3.8, one obtains a value for the resistivity that is 100 times larger than the experimental numbers—understandable because v_F is 16 times larger than v_{th}, and we know that the numbers obtained with v_{th} were already six times too large. The resolution lies in the calculation of the mean free path λ; in a perfectly ordered crystal $\lambda = \infty$; in a real crystal, λ is determined by scattering phenomena. We need to replace the inner atomic distance a in Equation 3.22 with the quantum mechanics value for λ, the mean free path of the conduction electrons. Experiments have shown that the electrons can move surprisingly far without any interaction; the mean free electron pass can be up to 10^8 atom distances at low temperatures. Electrons are not scattered by the regular building blocks of the lattice because of the wave character of the electrons. Scattering mechanisms instead are:

1. Lattice defects (foreign atoms, vacancies, interstitial positions, grain boundaries, dislocations, stacking disorders), and
2. Thermal vibration of the lattice (phonons).

Item 1 is more or less independent of temperature, whereas item 2 is independent of lattice defects but dependent on temperature. The mean free path

now does not depend on the radius of the ions but rather on deviations of the ions from a perfectly ordered array such as seen from lattice thermal vibrations and the presence of impurities. The ion vibrations lead to an effective area A that results in an electron mean free path given as $\lambda = 1/n_{ion}A$. Lattice ions are basically points only, but their thermal vibration has them occupy an electron scattering area $A = \pi r^2$, where r is the amplitude of the thermal vibrations. The energy of thermal vibration in a simple harmonic oscillator is proportional to the square of the amplitude of the vibration (r^2). In other words, the area A is proportional to the energy of the vibrating lattice ions. From the equipartition theory we know that the average vibration energy is proportional to kT, so it follows that A is proportional to T and λ is proportional to $1/T$. Because the mean free path is inversely proportional to temperature at high temperatures, it follows that $\sigma \propto \frac{1}{T}$, in agreement with the experimental evidence (Equation 3.23). This solves the issue of the wrong temperature dependence of the resistivity. We need to also calculate a correct absolute value for the resistance using the quantum mechanical mean free path.

The quantum mechanical mean free path of the conduction electrons, say in Cu, is defined as:

$$\lambda = \mathbf{v}_F \tau \tag{3.247}$$

where $\mathbf{v}_F$ is the velocity at the Fermi surface, because all collisions involve only electrons near the Fermi surface. With Fermi velocities $\mathbf{v}_F$ of typically 10^8 cm/s (see Table 3.13) and with room-temperature resistivities for many metals of $\rho \sim 1$–$10\ \mu\Omega\cdot$cm, the corresponding relaxation time is $\sim 10^{-14}$ s and the resulting averaged free electron path at room temperature is about 100 Å. So it is of the order of a few 10s to 100s of interatomic distances. At low temperatures for very pure metals the mean free path can actually be made as high as a few centimeters ($\tau \approx 2 \times 10^{-9}$ s at 4 K for very pure Cu). Compared with using the lattice constant, a, for the mean free path, this equation leads obviously to values for the conductivity that are 100 times higher—in agreement with experimental data.

For a current $\mathbf{J}$ of 1 A/mm^2 in a conductor with an electron density of $n = 10^{22}$ cm^{-3}, we calculate an average speed $\mathbf{v}_{avg}$ of the Fermi sphere of $J/ne \sim 0.1$ cm/s (see Equation 3.3), which is much less than the Fermi velocity $\mathbf{v}_F \sim 10^8$ cm/s. In a conductor, charges are always moving at the Fermi velocity, but the Fermi sphere moves much slower because electrons travel at fast Fermi velocities for a short average time, τ, and then "scatter" because of collisions with atom vibrations, grain boundaries, impurities, or material surfaces (especially in very thin films).

The value for $n(E)_{3D}$ in $\mathbf{J} = -n(E)_{3D}e\mathbf{v}_{avg}$ (Equation 3.3) at $T > 0$ is obtained from Equation 3.230:

$$n(E)_{3D}\,dE = G(E)_{3D} f_{FD}(E)\,dE$$

$$= \frac{8\pi\sqrt{2}m_e^{\frac{3}{2}}}{h^3} E^{\frac{1}{2}} \frac{1}{e^{\frac{E-E_F}{kT}}+1}\,dE \tag{3.248}$$

For a large semiconductor, Equation 3.248 was illustrated in Figure 3.119 with a broken line. We will see below that $G(E)$ depends strongly on dimensionality, so we also expect the current density to vary strongly with dimensionality.

Quantum Wells: $L_{1,2} > \lambda_{DB} > L_3$

Quantization As we saw earlier, devices that come with a length L that in one direction is comparable with the size of the electron de Broglie wavelength are known as quantum wells (1D confinement). A planar quantum well structure may be made from a thin region of a narrow gap semiconductor sandwiched between two layers of a wide bandgap semiconductor. We use for an example of such 1D confinement a quantum well made by sandwiching a layer of GaAs between two layers of $Al_xGa_{1-x}As$, as shown first in Figure 3.59 and, simplified, reproduced in Figure 3.122. Growing two different semiconductors on top of each other, as illustrated,

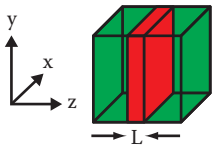


FIGURE 3.122 A quantum well can be made, for example, by sandwiching a layer of GaAs between two macroscopic layers of $Al_xGa_{1-x}As$. This creates a layer in which the electrons (and holes) behave in a 2D way (see also Figure 3.59).

forms heterojunctions. The narrower bandgap GaAs material is enclosed by a material with a considerably larger bandgap to establish a potential barrier at the surface of the confined material. Because of the potential barrier, the motion of electrons and holes is restricted in one dimension (thickness L in the z-direction in this case) and is forced to occupy discrete states of energy instead of staying arbitrarily within an energy continuum. Hence quantization of the system occurs by shrinking the thickness of the GaAs layer. A 2D electron gas in the laboratory is really a 3D electron system in which the electron motion is strongly confined in one spatial direction but in which free motion is still allowed in the other two directions. At the interface of the semiconductors GaAs and $Al_xGa_{1-x}As$ a potential well is formed. The potential well is a result of charge transfer between the two materials and their conduction band offset (E_c), confining the electrons, and the motion of electrons perpendicular to the plane of the heterointerface (z-axis in Figure 3.122) is quantized.

The composition x of the ternary semiconductor $Al_xGa_{1-x}As$ can be varied to control the electron barrier height. A good lattice match between GaAs and $Al_xGa_{1-x}As$ over a wide range of x values minimizes lattice strain at the interfaces. When the electron and holes are confined in the same layer, one talks about a type I quantum well; with electron and holes confined in different layers, one defines a type II quantum well.

For the quantum well considered here, the $Al_xGa_{1-x}As$ layers are thick enough so that tunneling through these layers remains very limited.

For an energy E smaller than the potential barrier V, the energy of an electron in the conduction band of a quantum well is given as:

$$E_n(k_x, k_y) = E_c^0 + \underbrace{\frac{n_z^2 h^2}{8m_e^* L_z^2}}_{E_{1D}(n_z)} + \underbrace{\frac{h^2(k_x^2 + k_y^2)}{8m_e^*}}_{E_{1D}(k_x, k_y)}, \quad n = 0, 1, 2, \ldots \tag{3.249}$$

with the wave function:

$$\psi_n(x, y, z) = \psi_n(z) e^{ik_x x} e^{ik_y y} \tag{3.250}$$

There is one quantized component in the z-direction and a "free" electron component in the x-y plane. The second term on the right side of Equation 3.249 represents the quantized energy in the z-direction (the thickness direction L of the GaAs film). This is the same expression we derived in Equation 3.156 for a finite-sized 1D box with infinitely high potential walls. For quantization to be important, the difference between the electron energy levels should be much larger than the thermal energy k_BT, that is, $E_n = \frac{n_z^2 h^2}{8m_e^* L_z^2} >> k_B T$, where n_z = 1, 2, 3 are the quantum numbers labeling.the energy levels. E_c^0 is the energy corresponding to the bottom of the conduction band. Strictly speaking, the above expressions apply only to an infinitely deep potential well. However, we can use the same equations as long as E_n is well below the bottom of the conduction band of the wide band material. Using this condition, we find, for example, that in GaAs where $m_e^*/m_e = 0.067$, the levels are quantized at room temperature when L_z = 150 Å. The third term on the right side of Equation 3.249 represents the kinetic energy of the electrons in the x-y plane where they are free to move. The $\mathbf{k}_z$-component is absent in the last term of Equation 3.249 because the motion in this direction is quantized. Equation 3.249 reveals that for each value of the quantum number n, the values of wave vector components $\mathbf{k}_x$ and $\mathbf{k}_y$ form a 2D band structure. The wave vector $\mathbf{k}_z$ in the z-direction, on the other hand, can only take on discrete values, $\mathbf{k}_z = n_z\pi/L_z$. For each value of n there is a sub-band with n the sub-band index as illustrated in Figure 3.123. In this figure we show energy levels (bottoms of

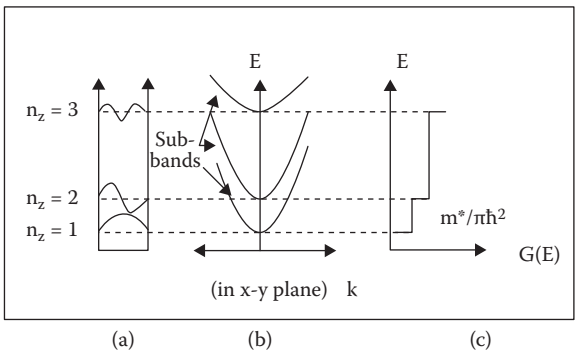


FIGURE 3.123 (a) Energy levels (bottoms of sub-bands) for a quantum well made by sandwiching a layer of GaAs between two macroscopic layers of $Al_xGa_{1-x}As$ (L_z is 150 Å). (b) Energy versus $\mathbf{k} = (\mathbf{k}_x^2 + \mathbf{k}_y^2)^{1/2}$ for 2D electron gas in GaAs quantum well. (c) Density of states for quantum well structure (Harris, 2006.[4])

sub-bands), density of states for a quantum well structure, and energy versus $\mathbf{k} = (\mathbf{k}_x^2 + \mathbf{k}_y^2)^{1/2}$ for the 2D electron gas in the GaAs quantum well.

We can carry out an analogous argument for the holes in the valence band with the difference that their quantized energy is inverted and that we need to invoke m_h^* for the effective mass of the hole. For a quantum well, the lowest-energy band-to-band (interband) transition is now different from the bandgap (E_g) transition of the bulk semiconductor. It will occur at a higher energy level (shorter wavelength) between the lowest energy state for electrons in the conduction band (n = 1) and the corresponding state for holes in the valence band. This defines the effective bandgap for a quantum well as:

$$E_g^* = E_c^0 - E_v^0 + \frac{h^2}{8L^2}\left(\frac{1}{m_e^*} + \frac{1}{m_h^*}\right) \qquad (3.251)$$

The shift to higher wavelengths is referred to as a "blue shift," caused by quantization.

Fermi Surfaces and Brillouin Zone for Quantum Wells In the case of a free 3D electron gas we appreciate that the surface of the Fermi sea is a sphere of radius $\mathbf{k}_F$ connecting points of equal energy in **k**-space (see Figure 3.111). In 2D this becomes a circle connecting points of equal energy in 2D **k**-space. In Figure 3.124 we show the Fermi circles corresponding to 2D crystals with one, two, three, and four valence electrons per atom. In this figure we also show the 2D Brillouin zone (see square in dashed line), and we see how the free electron circle of a three-valent metal (red circle) cuts the Bragg planes located at π/a. The square shown in Figure 3.124 corresponds to the first Brillouin zone of a 2D square lattice (Figure 3.112b).

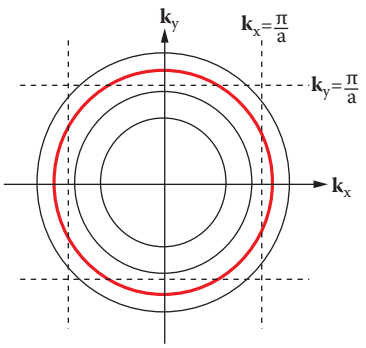


FIGURE 3.124 Free electron circles for 1, 2, 3, and 4 valence electrons. The free electron circle of a three-valent metal (red circle) cuts the Bragg planes located at π/a.

Density of States for Quantum Wells The density of states for each sub-band of a quantum well (Figure 3.123c) can be found using an approach similar to the one we used above for a 3D density of states function, that is, by counting the number of states with wave vectors **k** between **k** and $d\mathbf{k}$. In the case of 1D confinement of electrons we must find the number of **k**-states enclosed in an annulus of radius $\mathbf{k} + d\mathbf{k}$ (see Figure 3.125). Each state occupies an area A of $\left(\frac{2\pi}{L}\right)^2$. The area of the annulus, A, is given by:

$$A d\mathbf{k} = 2\pi k d\mathbf{k} \qquad (3.252)$$

Dividing the area of the annulus by the area occupied by a **k**-state, and remembering again to multiply by 2 for the electron spin states, we get for the number of states per unit area:

$$G(\mathbf{k})_{2D}\, d\mathbf{k} = 2\frac{2\pi k d\mathbf{k}}{\left(\frac{2\pi}{L}\right)^2} = \frac{L^2\mathbf{k}}{\pi} d\mathbf{k} \qquad (3.253)$$

Or, in terms of energy per unit area at an energy E (dividing by L^2), the density of states for each sub-band is given as:

$$G(E)_{2D} dE = \frac{k d\mathbf{k}}{\pi} = \sqrt{\frac{2m_e^* E}{\hbar^2}}\left(\frac{2m_e^* E}{\hbar^2}\right)^{-\frac{1}{2}} \frac{m_e^*}{\pi\hbar^2} dE$$
$$= \frac{m_e^*}{\pi\hbar^2} dE \qquad (3.254)$$

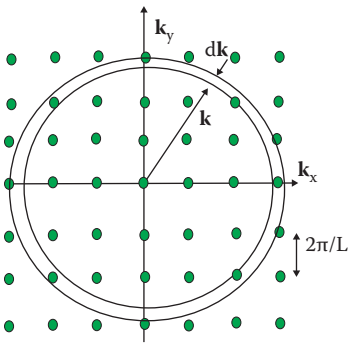


FIGURE 3.125 Density of states, $G(\mathbf{k})_{2D}$, is the number of allowed states between **k** and $\mathbf{k} + d\mathbf{k}$, i.e., the number of states between a sphere of radius **k** and a sphere of radius $\mathbf{k} + d\mathbf{k}$.

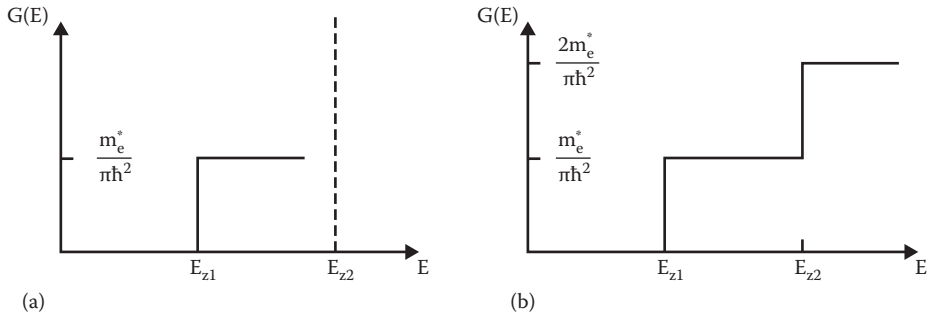


FIGURE 3.126 Filling of the first two sub-bands in a 2D structure. The first sub-band has a constant energy at $\frac{m_e^*}{\pi\hbar^2}$ (a), and the second has a constant energy at $\frac{2m_e^*}{\pi\hbar^2}$ (b).[4]

For this derivation we used Equation 3.227, which holds true for any dimensionality D of the problem. Importantly, in an ideal 2D system, the density of states is constant and does not depend on energy.

The density of states in a 2D electron gas, in which only the lowest energy sub-band (n = 1 level E_{z1}) is occupied, is illustrated in Figure 3.126a, where we assume that the confinement is in the z-direction. For all electron energies up to E_{z1} the density of states is zero because electrons cannot exist in the well at lower energies than this. Because of the freedom of motion in the plane of the hetero interface (with continuously varying wave vectors $\mathbf{k}_x$ and $\mathbf{k}_y$ in the x-y plane), the energy levels that form in the potential well are highly degenerate. If the electron density is sufficiently low, then all electrons can be accommodated in the lowest level of the well, and the freedom of motion in the transverse direction is frozen out and the electron system effectively behaves as the ideal 2D one as described by Equation 3.254 with a single step in energy only with a constant value $\frac{m_e^*}{\pi\hbar^2}$. In the regime where only the lowest sub-band is occupied, increasing energy corresponds to increasing electron motion in the x-y plane (kinetic energy).

As the electron filling of the quantum well is increased, eventually electrons begin to fill the next transverse level of the potential well (Figure 3.126b). At this point we have two so-called 2D sub-bands occupied in the quantum well. To a good approximation, electrons in the two sub-bands may be viewed as forming two independent 2D electron gas systems. Thus, the density of states is double that which we would expect in the case where just a single sub-band is occupied. For each quantum state in the quantum well, there will be a step in the density of states. The overall density of states is discontinuous, with a stepwise structure that is characteristic of quantum wells. Because of the summing over the different sub-bands, a more general description of a 2D system has Equation 3.254 modified as:

$$G(E)_{2D} = \frac{m_e^*}{\pi\hbar^2}\sum_n H(E - E_n) \qquad (3.255)$$

where $H(E - E_c)$ is the Heaviside step function.* It takes the value of zero when E is less than E_n and 1 when E is equal to or greater than E_n. E_n is the nth energy level within the quantum well.

In Figure 3.127 we summarize the characteristics of a 2D density of states function $G(E)_{2D}$. The energy spacing shown here increases with decreasing L; the thinner the 2D film, the more it approximates an ideal 2D gas with only one sub-band.

The step-like behavior of a 2D density of states function $G(E)_{2D}$ implies that the density of states in the vicinity of the bandgap is much larger than in the case of a bulk semiconductor, where the value of $G(E)_{3D}$ goes to zero (E_1 versus E_g). This makes for stronger optical transitions because a major factor

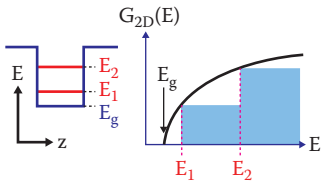


FIGURE 3.127 The density of states function for a quantum well. The solid black curve is that of a free electron. The bottom of the quantum well is at energy E_g.

* The Heaviside step function, H, also called the unit step function, is a discontinuous function whose value is zero for negative argument and one for positive argument.

in the expression for the probability of optical transitions is the density of states. The strength of an optical transition is often defined as the oscillator strength, and the oscillator strength of a quantum well in the vicinity of the bandgap is considerably enhanced compared with a bulk semiconductor. Below we will show how Fermi's golden rule for absorption describes transition rates between levels in terms of the availability of states [density of states: $G(E)$], the availability of photons (intensity E_0), and a "coupling strength" between the levels (transition matrix element $\left|H'_{v,c}\right|^2$). The enhanced oscillator strength of quantum-confined structures is put to good use in the fabrication of laser media for highly efficient and compact solid-state lasers (see Chapter 5). In Figure 3.128 we illustrate the bound-state energies for electrons in the conduction band and for holes in the valence band for a GaAs/$Al_xGa_{1-x}As$ heterojunction. If the GaAs layer is thin enough, bound states form as indicated here by dashed lines. For a perfect 2D electron gas, each bound state corresponds to a discontinuous jump in the electronic density of states function. The density of states function (DOS) can be investigated by measuring the absorption α of electromagnetic radiation associated with the excitation of an electron from the valence band to the conduction band. Because the wavelength of the radiation is long compared with the width of the well, transitions only occur between states for which the spatial variation of the wave function is similar; this leads to the selection rule $\Delta n = 0$ for the adsorption. Therefore, the allowed transitions are those indicated by the arrows in Figure 3.128a. The frequency dependence of the absorption should reflect the steps in the DOS with the steps in the absorption expected to occur at frequencies ω_n given by:

$$\hbar\omega_n = E_{Cn} - E_{Vn} \tag{3.256}$$

where $E_{Cn} - E_{Vn}$ is the energy difference between the nth bound states in the conduction and valence bands. The measured adsorption spectra for GaAs layers of thickness 140, 210, and 4000 Å are shown in Figure 3.128b, and the expected step structure is clearly visible for the two thinner layers, with arrows indicating the frequencies at which steps are expected. Peaks at energies just a little below the predicted values mark the absorption. These results from the creation of an exciton, an electron-hole bound state that is created when a photon is absorbed (see below for details). Because there is an attraction between the electron and the hole in an exciton, the photon energy required to create an exciton is lower than the predicted values that ignore such interactions. Thus, the difference in energy of the peak and the predicted absorption edge is a measure of the binding energy of the electron-hole pair. An exciton, with an energy just below the bandgap, is clearly seen in the absorption curve for the 4000-Å layers, but the step-like structure has disappeared, indicating that the DOS is a smooth curve as one expects for 3D behavior.

Electronic Conductivity of Quantum Wells The circles shown in Figure 3.124 are the Fermi circles corresponding to 2D crystals with one, two, three, and

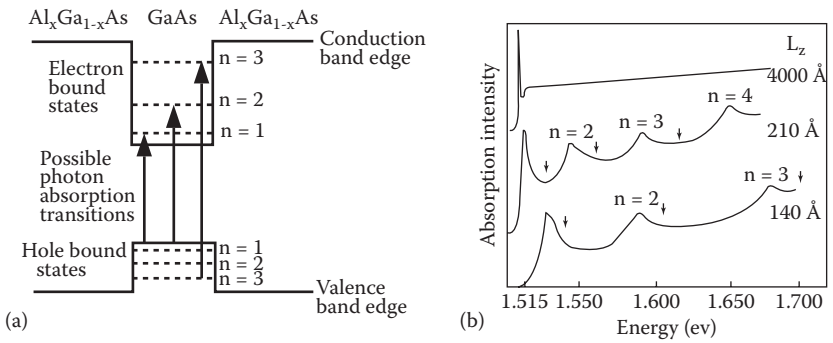


FIGURE 3.128 (a) Heterojunction of GaAs and $Al_xGa_{1-x}As$. If the GaAs layer is thin enough, bound states (dashed lines) form. When photons are absorbed, electrons are excited between these bound states. (b) The adsorption of light, measured as a function of photon energy for GaAs layers of thickness 4000, 210, and 140 Å. The arrows indicate the energies at which the onset of adsorption is expected to occur for transitions involving the nth bound state.

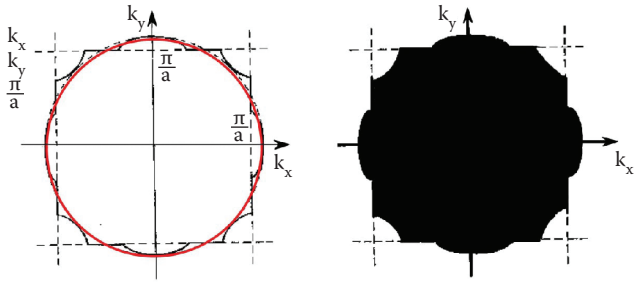


FIGURE 3.129 Deformation of the free electron circle near Bragg planes, $V(x) \neq 0$.

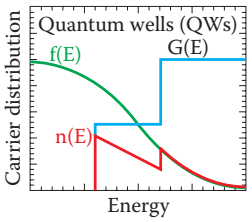


FIGURE 3.130 Quantum wells (QWs). Density of states (DOS) (blue) and occupied states (red) around the Fermi energy for a quantum well. The Fermi-Dirac function is $f(E)$, and the product of $f(E)G(E)_{2D} = n(E)_{2D}$ at $T > 0$.

four valence electrons per atom. The circles represent surfaces of constant energy for free electrons [$V_{(x)} = 0$], i.e., the Fermi surface for some particular value of the electron concentration. The total area of the filled region in **k**-space depends only on the electron concentration and is independent of any interaction of electrons with a lattice. In a solid more realistic shape of the Fermi surface depends on the lattice interaction of the electrons and will usually not be an exact circle in a lattice. There is a discontinuity introduced into a free electron Fermi circle any time it approaches a 2D Brillouin zone boundary. In Figure 3.129 we see how the shape of the Fermi circle is distorted near the surface if $V(x)$ is not zero.

Energy gaps appear at zone boundaries, and the Fermi surface intersects zone boundaries almost always perpendicularly. The crystal potential causes rounding of sharp corners on a Fermi surface. The volume enclosed by a Fermi surface only depends on electron density and not on details of the lattice interaction. In other words, the volume enclosed by a Fermi surface remains unchanged under the "deformations" just mentioned.

The number of occupied states per unit volume in the energy range E to $E + dE$ and with $f(E) = 1$ $(T = 0)$ is calculated as:

$$n_{2D} = \int_0^{\infty} n(E)_{2D} dE = \int_0^{E_F} G(E)_{2D}[f(E) = 1]dE \tag{3.257}$$

or

$$n_{2D} = \int_0^{E_F} \frac{m_e^*}{\pi\hbar^2} dE = \frac{m_e^* E_F}{\pi\hbar^2}$$

To calculate the current density for a 2D gas at a particular temperature we substitute the value for $n(E)_{2D}$ at $T > 0$ [and thus $f(T)$ can then be different from 1] in $\mathbf{J} = -n(E)_{2D}e\mathbf{v}_{avg}$ (Equation 3.3). The function $n(E)_{2D}$ for a given temperature T (>0) is shown as a red line in Figure 3.130. In the same graph we also show the Fermi-Dirac function and the DOS function (blue).

The form of the density of states function of a 2D gas can also be dramatically modified by a magnetic field, giving rise to very pronounced behavior in the conductance with the appearance of the so-called Landau levels.

Quantum Wires: $L_1 > \lambda_{DB} > L_{2,3}$

Quantization We saw earlier that when L becomes very small along two directions (2D confinement)—of the order of the de Broglie wavelength of an electron—one obtains a quantum wire where electrons can only move freely in one direction, i.e., along the length of the quantum wire as shown first in Figure 3.63 and, simplified, reproduced in Figure 3.131. These 1D electronic structures with 2D quantum confinement comprise nanowires, quantum wires, nanorods, and nanotubes.

The starting point for the fabrication of one type of quantum wire is a 2D electron gas confined in one direction as discussed above. A 2D quantum gas that is very strongly confined at some interface and where we can assume that only the lowest subband of the electron gas is occupied, so that the motion transverse to that interface is frozen out. This is the situation we encountered for a

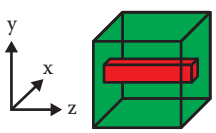


FIGURE 3.131 A rectangular quantum wire.

heterojunction with a very small L, so that Equation 3.254 is applicable [$G(E)_{2D} = \frac{m_e^*}{\pi\hbar^2}$ (no summation of steps)]. To this strong confinement, a typically weaker, lateral confinement of the electrons is added by etching nanowires in the 2D quantum well. In Figure 3.131 we drew a rectangular quantum wire with a square cross-section; in reality nanowires will typically be much wider than they are thick (the x-direction is assumed to be the film thickness direction), so the quantum confinement is most severe in the x-direction. The reason for this discrepancy is that in micro- and nanotechnology it is much easier to control a film's thickness (x-direction) than it is possible to control the lateral dimensions of a structure (y direction). Let us consider the 75-nm-wide quantum wires etched in a $GaAs/Al_xGa_{1-x}As$ heterojunction in Figure 3.132. The 2D electron gas formed at the heterointerface is confined here in a scale of just a few nanometers (in the thickness or x-direction), and so its quantized energies are large in the x direction. The lateral confinement (y direction) of electrons in the 75-nm-wide wire is much weaker than this and its quantized energies are consequently much smaller. The weaker lateral confinement in the wires gives rise to a series of relatively closely spaced energy levels (see E_Y in Figure 3.132). Therefore, transport through the wire in the z-direction will involve electrons that occupy many of these lateral sub-bands. Thus, these structures in reality are quasi-2D confined systems with free electron motion in one direction and two different types of confinement in the other two.

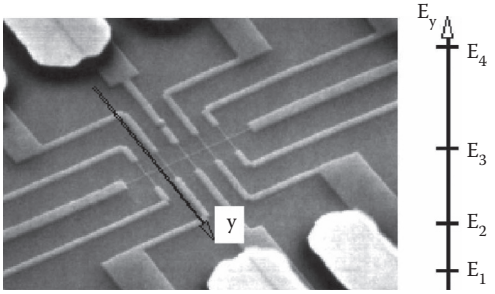


FIGURE 3.132 75-nm-wide quantum wires etched in a $GaAs/Al_xGa_{1-x}As$ heterojunction. The confinement of electrons in the x- and y-directions quantizes the electron energy into a set of discrete energies. The confinement in the x-direction is much stronger than in the y-direction.

For an idealized nanowire with a square cross section and the x = y dimensions in the nanoscale but continuous along the wire axis (z-direction), the energy dispersion function may be written as:

$$E = E_{n_1,n_2}(k_z) + \frac{\hbar^2 k_z^2}{2m_e^*} \qquad (3.258)$$

with the wave function:

$$\psi(x, y, z) = \psi_{n_1,n_2}(x, y)e^{ik_{yz}} \qquad (3.259)$$

with n_1 and n_2 the quantum numbers labeling the eigenstates in the x-y plane and $\mathbf{k}_z$ the wave vector in the z-direction, and where we assume that the dispersion relation for the electron energy is parabolic. The energy term has again two contributions: one is caused by the continuous band value $\left(\frac{\hbar^2 k_z^2}{2m_e^*}\right)$, and the other term [$E = E_{n_1, n_2}(k_z)$] is that of the quantized values of electrons confined in two dimensions as derived in Equation 3.158 (unconstrained direction is along the z-axis). Referencing energies to the conduction band edge, the latter expression may be rewritten as:

$$E_{n_1,n_2}(\mathbf{k}_z) = E_c^0 + \frac{h^2}{8m_e^*}\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2}\right) + \frac{\hbar^2 k_z^2}{2m_e^*} \qquad (3.260)$$

where E_c^0 is the bottom of the conduction band. The lowest sub-band is obtained for n_1 and $n_2 = 1$. The energy at the bottom of each sub-band ($\mathbf{k}_z = 0$) is simply given by:

$$E_{n_1,n_2} = E_c^0 + \frac{h^2}{8m_e^*}\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2}\right) \qquad (3.261)$$

If we choose the cross-section of the GaAs quantum wire containing the 2D confined electron gas to be equal to 100 × 100 Å, then the lowest energies in the two lowest sub-bands are equal to 0.112 eV and 0.280 eV. This is determined from Equation 3.261 for $n_1 = 1$, $n_2 = 1$ and $n_1 = 1$, $n_2 = 2$, respectively. In the case that the nanowire is much wider (y-direction) than it is thick (x-direction), the quantum confinement is most severe in the x-direction. Because in this case $L_y >> L_x$, the n_2 levels form a staircase of small steps in the widely separated sub-bands corresponding to the various values for n_1. Thus, confinement that is different in the x and y directions splits each

sub-band further up into a set of more narrowly spaced sub-bands.

In Figure 3.133 we show band structure and density of states for a quantum wire with a square cross section; similar to the situation in quantum wells, sub-bands are formed, but in quantum wires with a rectangular cross-section multiple sub-bands form at each eigenvalue $E(n_x,n_y)$, spread out as $\frac{\hbar^2 k_z^2}{2m_e^*}$.

Fermi Surface of Quantum Wires Whereas Fermi surfaces in 3D and 2D electronic structures consist of a sphere and a circle, respectively, the Fermi surface of a strictly 1D electronic structure consists of just two points at $+\mathbf{k}_F$ and $-\mathbf{k}_F$ or $\int d\mathbf{k} = 2\mathbf{k}_F$ (where $\mathbf{k}_F$ is the Fermi wave vector) or $n_{1D} = \frac{2\mathbf{k}_F}{\pi}$. This unusual Fermi surface has some pretty significant consequences. In 1981, Hiroyuki Sakaki predicted that ideal 1D electrons moving at the Fermi level in quantum wires would require very large momentum changes ($\Delta k = 2\mathbf{k}_F$) to undergo any scattering. The result is that electron scattering is strongly forbidden. This is a consequence of the fact that in one dimension, electrons can scatter only in one of two directions: forward and 180° backward. With this large reduction in scattering, electrons should achieve excellent transport properties (e.g., very high mobility).

Density of States Function for Quantum Wires For calculating the density of states for a quantum wire, we use the same approach we took for the 3D and 2D cases. The **k**-state has now a length l of $2\pi/L$, and we must find the number of **k**-states lying in a length of $\mathbf{k} + d\mathbf{k}$. The wire length difference, l, is given by:

$$l = 2 \tag{3.262}$$

The factor of two appears because the wave number could be either positive or negative, corresponding to the two directions along the wire.

The resulting density of states per unit length of 1D **k**-space is obtained by multiplying by 2 for spin degeneracy and dividing by $2\pi/L$:

$$G(\mathbf{k})_{1D}\,d\mathbf{k} = 2\left(\frac{2}{2\pi/L}\right)d\mathbf{k} = \frac{2L}{\pi}d\mathbf{k} \tag{3.263}$$

To obtain the density of states in terms of the energy $[G(E)_{1D}]$, we use again the relation between E and **k**, derived earlier as $\mathbf{k} = \left(\frac{2m_e^* E}{\hbar^2}\right)^{\frac{1}{2}}$ (Equation 3.150) and differentiate the latter expression with respect to energy:

$$d\mathbf{k} = \left(\frac{2m_e^* E}{\hbar^2}\right)^{-\frac{1}{2}} \frac{m_e^*}{\hbar^2} dE \tag{3.227}$$

The density of energy states $G(E)_{1D}$, the number of allowed states between E and $E + dE$, per unit energy and unit length (divide by L) is then obtained using the chain rule:

$$G(E)_{1D}dE = \frac{2d\mathbf{k}}{\pi} = \frac{1}{\pi}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{1}{2}} \frac{1}{E^{\frac{1}{2}}} dE = \frac{1}{h\pi}\sqrt{\frac{2m_e^*}{E}}dE \tag{3.264}$$

This is in sharp contrast with the behavior of a 3D electron gas where $G(E)_{3D}$ goes to zero at low energies, and two dimensions, where $G(E)_{2D}$ steps up to a constant value at the bottom of each 2D sub-band. Remembering that the group velocity is given as:

$$\mathbf{v}_g = \frac{d\omega}{d\mathbf{k}} \tag{3.101}$$

which we found may also be written as:

$$\mathbf{v}_g = \frac{1}{\hbar}\left[\frac{dE}{d\mathbf{k}}\right] \text{ or more generally } \mathbf{v}_g = \frac{1}{\hbar}\nabla_k E(\mathbf{k}) \tag{3.202}$$

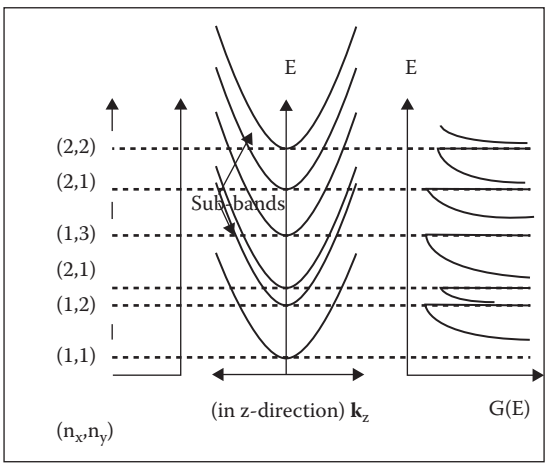


FIGURE 3.133 Band structure and density of states for a quantum wire with a square cross-section.[4]

we can rewrite Equation 3.264 now also in terms of the group velocity as:

$$G(E)_{1D}dE = \frac{2dk}{\pi} = \frac{2}{\pi}\left(\frac{dE}{dk}\right)^{-1} dE = \frac{2}{\pi\hbar v_g(E)}dE \quad (3.265)$$

Thus, for a 1D system the density of states is inversely proportional to the velocity! We will use this expression in v_g when we derive the expression for the current through a nanowire. We will learn that in a nanowire the current is constant and proportional to the velocity and density of states (see further below Equation 3.269). If we add electrons to the nanowire, they initially fill only the lowest of the levels, say level (1,1) in Figure 3.133. As the energy is increased and remains less than level (1,2), the lateral motion remains frozen, and the increase in energy is transferred into motion along the length of the wire. The density of states in this level will then take the form predicted by Equation 3.265. However, in most situations the Fermi energy of the electrons in the wire is several times larger than the average spacing between the lateral energy levels in Figure 3.133, so several of these levels will be occupied, and each of these levels defines a corresponding 1D sub-band. The density of states within each sub-band is 1D, but the total density of states is obtained by summing over all sub-bands n. The summing over the individual sub-bands can be formulated mathematically as:

$$G(E)_{1D} = \frac{1}{\pi}\left(\frac{m_e^*}{\hbar^2}\right)^{\frac{1}{2}} \sum_n L\left(\frac{1}{E - E_n}\right)^{\frac{1}{2}} \quad (3.266)$$

L is the unit step function (same as H in Equation 3.255). The density of states of a quantum wire diverges as $\left(\frac{1}{E - E_n}\right)^{\frac{1}{2}}$ at each sub-band threshold and has the inverse energy dependence $E^{-1/2}$ compared with the $E^{1/2}$ dependence of a 3D electron gas (bulk semiconductor). The DOS of a quantum wire obviously has a more pronounced structure than does a 2D well, with a large number of sub-bands, each one starting as a peak. An immediate manifestation of a large $G(E)_{1D}$ at the bottom of each sub-band is again an increase of the strength of optical transitions or oscillator strength as compared with 3D and 2D electronic structures.

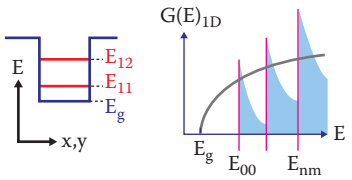


FIGURE 3.134 The density of states function for a rectangular quantum wire. The solid black curve is that of a free electron.

In Figure 3.134 we summarize the characteristics of a 1D density of state function $G(E)_{1D}$. The axis of the rectangular wire is again in the z-direction, and quantization is in the x- and y-directions. The peaks show the calculated density of states of a quantum wire over a range of energies where several different sub-bands become occupied. For comparison, the solid line shows the monotonic variation of the density of states expected for a 3D system.

Electronic Conductivity of Quantum Wires Ohm's law for macroscopic systems is given by V = IR, and in terms of conductance σ this is equivalent to σ = J/E (Equation 2.2). From the previous sections, we know that only electrons close to the Fermi level contribute to the conductance. From Bloch's theorem, we also recall that in an ideal periodic potential, electrons propagate without any scattering—and thus no resistance at all—but that electron propagation in real materials does involve scattering. The origin of such scattering can be any source of disorder that disturbs the perfect symmetry of the lattice. Examples include defects and impurities, scattering from other electrons, and lattice vibrations (phonons). Because an electric current constitutes a movement of those "bumbling" electrons under an electric field, one expects that with nanostructures featuring dimensions comparable with the fundamental size of the electron, electrical properties will be strongly influenced by quantum-mechanical transport effects.

Each of the discrete peaks in the density of states (DOS) in Figure 3.134 is caused by the filling of a new lateral sub-band. The peaks in the density of states functions at those energies where the different sub-bands begin to fill are called criticalities or Van Hove singularities. These singularities (sharp peaks) in the density of states function lead to sharp

peaks in optical spectra and can also be observed directly with scanning tunneling microscopy (STM). The STM tunneling current J as a function of bias V gives spatial and spectroscopic information about the quantum states of nanostructures. At low temperatures, the derivative of the current with respect to voltage is roughly proportional to the density of states of the sample weighted by the density of states at the STM tip position in other words a convolution of the DOS of the tip and the sample. Because dJ/dV ~ DOS, the DOS is directly observed from an STM scan. This is illustrated in Figure 3.135 for a metallic and semiconducting carbon nanotube. The DOS is shown in Figure 3.135a and b, and the differential conductance is illustrated in Figure 3.135c and d. Metallic carbon tubes have a constant DOS around E = 0, whereas for semiconducting ones there is a gap at E = 0 (DOS = 0).

Singularities occur because electron transitions accumulate at frequencies for which the bands n_1,n_2 are parallel, i.e., at frequencies where the gradient is equal to zero:

$$\nabla_k[E_{n_1}(\mathbf{k}) - E_{n_2}(\mathbf{k})] = 0 \qquad (3.267)$$

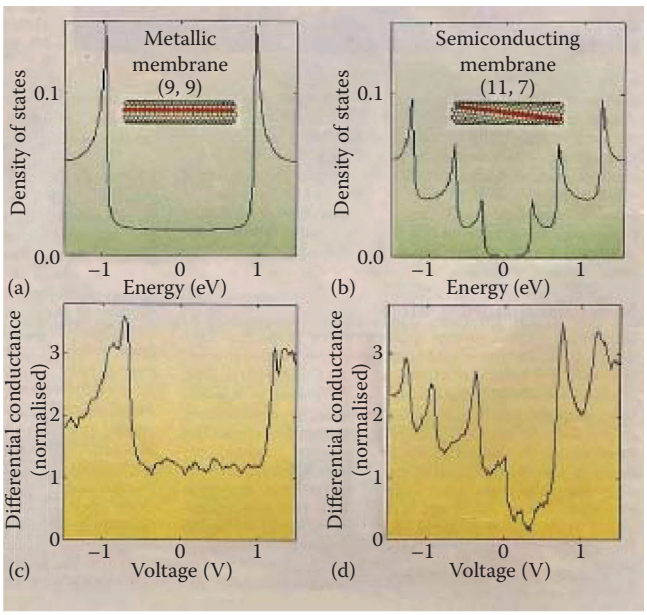


FIGURE 3.135 Van Hove singularities can be observed directly with scanning tunneling microscopy (STM). This is illustrated here for a metallic and semiconducting carbon nanotube. The DOS is shown in (a) and (b) and the differential conductance in (c) and (d). (From Dekker, C., *Physics Today*, 52, 22–28, 1999. With permission.[24])

where n_1 is an empty band and n_2 is a filled band. This can also be appreciated by considering that the density of states according to Equation 3.228 is given as $G(E) = G(k)(dE/d\mathbf{k})^{-1}$, so that singularities emerge when $d\mathbf{k}/dE = 0$ (for an example of such states, see the band diagram of Si in Figure 3.116, states at E_1 and E_2).

The number of occupied states at T = 0 [where f(E) = 1] per unit volume in the energy range E to $E + dE$ is again calculated from $n(E)_{1D}dE = G(E)_{1D}[f(E) = 1]dE$ and one obtains:

$$n_{1D} = \int_0^\infty n(E)_{1D}\,dE = \int_0^{E_F} G(E)_{1D}[f(E) = 1]dE \text{ or } n_{1D} = \frac{2\mathbf{k}_F}{\pi} \qquad (3.268)$$

that $n_{1D} = \int_0^{E_F} G(E)_{1D}dE$ indeed equals $\frac{2\mathbf{k}_F}{\pi}$ is easily recognized from the fact that:

$$G(E)_{1D}dE = \frac{2d\mathbf{k}}{\pi} \qquad (3.265)$$

To calculate the number of occupied states at T > 0 per unit volume in the energy range E to $E + dE$, we use again $n(E)_{1D}dE = G(E)_{1D}f(E)dE$, but because T is now > 0, $f(E)$ might be different from 1. That function is shown as a red line in Figure 3.136.

Consider now a quantum wire held between ideal contacts, as shown in Figure 3.137. In this very short 1D wire, the charge transport is assumed to be ballistic, i.e., electron transport without scattering, or the mean free path (λ) and device size (L) are comparable. Because the current carrying capacity of a 1D wire is finite, there must be a finite conductance even in the absence of scattering. Applying a voltage V

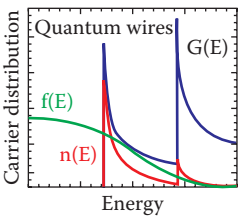


FIGURE 3.136 Quantum wires. Density of states (DOS) (blue) and occupied states (red) around the Fermi energy for a quantum wire. The Fermi-Dirac function is $f(E)$, and $n(E)$ is the product of $f(E)$ and $G(E)_{1D}$ at $T > 0$.

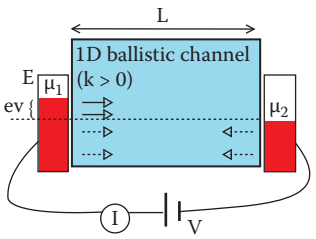


FIGURE 3.137 A quantum wire held between ideal contacts. In this very short one-dimensional wire, the charge transport is assumed to be ballistic with the mean free path (λ) and device size (L) comparable. Fermi level $E_F(1) = \mu_1$ and Fermi level $E_F(2) = \mu_2$ with $E_F(2) - E_F(1) = qV$, where $q = -e$ for electrons and $+e$ for holes. A net current flows between the two contacts for an applied bias V. A positive wave vector $\mathbf{k}$ is assumed with excess right-moving electrons constituting the current.

over the wire has an excess of electrons moving in one direction only (left to right in the current case), so we assume a positive wave vector $\mathbf{k}$ with excess right-moving electrons constituting the net current.

The chemical potential μ of the electrons in the contacts, we recall, is equivalent to the Fermi level, and here we assign the contacts Fermi level $E_F(1) = \mu_1$ and Fermi level $E_F(2) = \mu_2$ with $E_F(2) - E_F(1) = qV$, with $q = -e$ for electrons and $+e$ for holes. To calculate the current density for this 1D electron gas at $T = 0$, we must substitute the value for $n(E)_{1D}$ at $T = 0$ [where $f(E) = 1$] in $\mathbf{J} = -n(E)_{1D}\, e\mathbf{v}_{avg}$ (Equation 3.3), and because $n(E)_{1D} = \int_0^{\infty} n(E)_{1D}\, dE = \int_0^{E_F} G(E)_{1D} dE$, we can use the expression for $G(E)_{1D}$ in terms of the group velocity, $G(E)_{1D} = \dfrac{2}{\pi\hbar v_g(E)}$ (Equation 3.265), resulting in:

$$\mathbf{J} = -n(E)_{1D}\, e\mathbf{v}_{avg}(E) = -e\int_{E_F(1)}^{E_F(1)} \mathbf{v}_{avg}(E) n(E)_{1D}\, c$$

$$= -e\int_{E_F(1)}^{E_F(2)} \mathbf{v}_{avg}(E) G(E)_{1D}\, dE$$

$$\mathbf{J} = -e\int_{E_F(1)}^{E_F(2)} \mathbf{v}_{avg}(E)\frac{2}{\pi\hbar v_g(E)}\, dE \quad \text{and}$$

$$\text{since } \mathbf{v}_{avg}(E) = \mathbf{v}_g(E)$$

$$\mathbf{J} = -\frac{2e}{h}[E_F(2) - E_F(1)] = \frac{2e^2}{h}V \qquad (3.269)$$

since $E_F(2) - E_F(1) = qV$, with $q = -e$ for electrons and $+e$ for holes.

For a macroscopic metal wire, the conductance decreases continuously as the diameter of the constriction is reduced. However, as the diameter becomes of the order of the Fermi wavelength, the conductance takes on integer values of the fundamental unit of conductance $\sigma\ (= J/V) = 2e^2/h$. The conductance $\sigma\ (= J/V)$ is fixed, regardless of length L of the conductor, and as a consequence there is no well-defined conductivity σ. This phenomenon is called quantized conductance (QC). What is remarkable is that in the case of a quantum wire the velocity cancels with the density of states, producing a current that depends only on the voltage and fundamental constants. As a consequence, the current carried by a sub-band is independent of the sub-band shape; it need not be parabolic, and the relation holds for all cases! The amount of current is dictated only by the number of modes (also called sub-bands or channels) $M(E_F)$, which are filled between μ_1 and μ_2 with each mode contributing $2e^2/h$ or:

$$\sigma = \frac{2e^2}{h}M(E_F) \qquad (3.270)$$

The corresponding quantized resistance is:

$$R = \frac{h}{2e^2}\frac{1}{M(E_F)} \qquad (3.271)$$

As we increase the channel cross-section, the number of modes increases, so the resistance goes down. This quantized resistance R has a value of 12.906 kΩ and has allowed for the definition of a new practical standard for electrical resistance: the resistance unit h/e^2, roughly equal to 25.8 kΩ; the latter number is referred to as the von Klitzing constant R_K (after Klaus von Klitzing, the discoverer of quantization of the Hall effect), and since 1990, a fixed conventional value R_K –90 has been used in resistance calibrations worldwide.

Resistance of a conductor, of course, decreases as the length of the channel decreases. But at some point it reaches a minimum, given by Equation 3.271, and decreases no further. This minimum resistance comes from the contact resistance, and the smaller the device, the more important this contact resistance. One might assume that this resistance can be eliminated by improving the contacts, but this simply is not true: there is a minimum contact resistance.

Quantized conductance in a quantum wire is dramatically illustrated in Figure 3.138, where we show experiments involving quantum wires in a GaAs/$Al_xGa_{1-x}As$ heterostructure at very low temperatures (close to 0 K). A short quasi-1D "wire" is formed here between two regions of a 2D electron gas as a negative gate voltage V_g applied to the metallic gates on the surface depletes the carriers in the underlying 2D electron gas in the GaAs (Figure 3.138a). Once the narrow constriction is reduced to a width d comparable with the electron Fermi wavelength λ_F (see Figure 3.138b) 1D quantization is obtained regardless of the wire length L. In the conductance voltage plot shown in Figure 3.138c, 14 quantized conductance steps of height $2e^2/h$ are observed for a wire width d of 0.5 μm (see also Figure 3.138d). Each step corresponds to the occupancy of an additional sub-band or additional conductance channel in the quantum wire. In other words, $M(E_F) = 14$.

The conductance voltage plots in Figure 3.138d show that quantized steps are still clearly present for the 5-μm wire, indicating ballistic transport, but wash away for 10-μm wires because of crossover to diffuse transport. Thus, for this GaAs 1D system, the electron Fermi wavelength λ_F mean free path is between 5 and 10 μm. Long electron Fermi wavelengths λ_F are obviously key to integrating these 1D electronic nanostructures.

An important experimental finding is that the conductance quantization is typically destroyed in narrow constrictions longer than a few hundred nanometers; ballistic transport cannot be preserved on length scales much longer than this. The reason for this is that as the constriction length is made longer, there is a greater probability of finding some disordered region (a barrier) that can disrupt ballistic transport by causing electron scattering. When we derived the conductance quantization

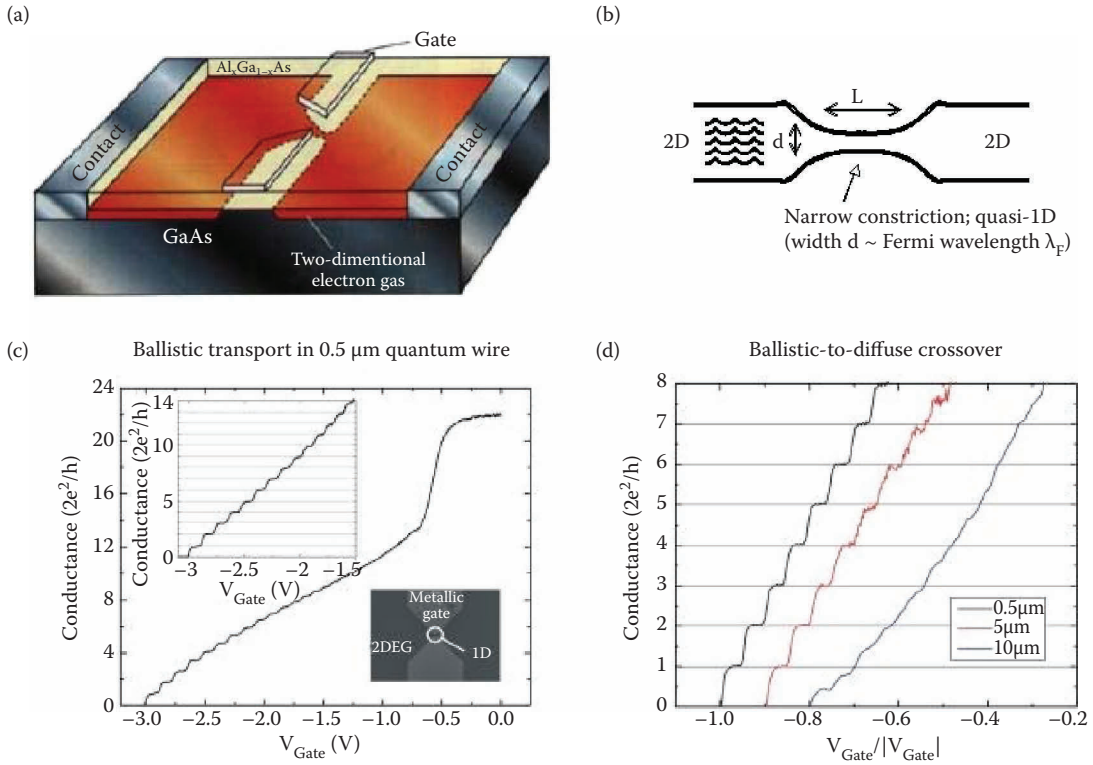


FIGURE 3.138 Conductance quantization in a short channel electrostatically defined in a GaAs/$Al_xGa_{1-x}As$ heterostructure. Measurements are carried out at very low temperatures (close to 0 K). (a) A negative gate voltage V_g applied to the metallic gates depletes the carriers in the underlying 2D electron gas. (b) Once the narrow constriction is reduced to a width d comparable to the electron Fermi wavelength λ_F, 1D quantization is obtained regardless of the wire length L. (c) In the conductance voltage plot, 14 quantized conductance steps of height $2e^2/h$ are observed for a wire width d of 0.5 μm. (d) Ballistic to diffuse transport crossover in GaAs quantum wires. The conductance voltage plots show that quantized steps are still clearly present for the 5-μm wire, indicating ballistic transport, but wash away for 10-μm wires because of crossover to diffuse transport. (From Sandia National Laboratory.)

for a ballistic system we assumed a total absence of any scattering between the different 1D sub-bands. This allowed us to make a definite association between the direction of the carriers in the wire and the reservoirs from which they originated (see Figure 3.139).

Such a distinction is no longer possible in the case where scattering occurs in the wire because electrons traveling in any particular direction may have originated from either reservoir. To compute the current that flows through such a disordered conductor, we now consider the probability that an electron incident in a given channel in one lead is transmitted into the other lead. Because of scattering in the sample an electron incident in some particular channel in the input lead will be transmitted into the different channels of the output lead with different probabilities. In the case in which a channel is not perfectly conducting, e.g., there is some type of barrier in the channel, the conductance of that channel is then given by the quantum conductance times the electron transmission probability at the Fermi energy $T_n(E_F)$ through the barrier in that channel:

$$\sigma = \frac{2e^2}{h} T_n(E_F) \tag{3.272}$$

where $T_n(E_F)$ is given as:

$$T_n(E_F) = \sum_{m=1}^{N} T_{nm}(E_F) \tag{3.273}$$

Here T_{nm} is the probability that an electron incident in mode n is transmitted into mode m in the output lead and the summation runs over the number (N) of occupied modes in the leads. Therefore, the probability $T_n(E_F)$ provides a measure of the degree of disorder in the conductor. This leads us to the so-called Landauer (1927–1999) formula, wherein the total current flow through the conductor is just the sum over all the occupied channels, which leads to the following expression for the conductance:

$$\sigma = \frac{J}{V} = \frac{2e^2}{h} \sum_{n=1}^{N} T_n = \frac{2e^2}{h} T \tag{3.274}$$

where T is the total transmission coefficient of the conductor. A simple application of the Landauer formula is the problem of the ballistic quantum point contact that we considered earlier. In this case, each channel is perfectly transmitted, and the sum over the transmission probabilities in Equation 3.274 is replaced by the total number of channels $M(E_F)$; therefore, we recover the usual quantization condition given by Equation 3.270:

$$\sigma = \frac{2e^2}{h} M(E_F) \tag{3.270}$$

So we sum over the contributions of each sub-band and for perfect transmission of all the sub-bands $T(E_F) = M(E_F)$ (which was 14 in Figure 3.138). When scattering is possible, however, the transmission coefficients T_n will no longer be equal to unity, and so the conductance quantization will not be observed.

The Landauer equation can also be written in terms of transmission and reflection coefficients.

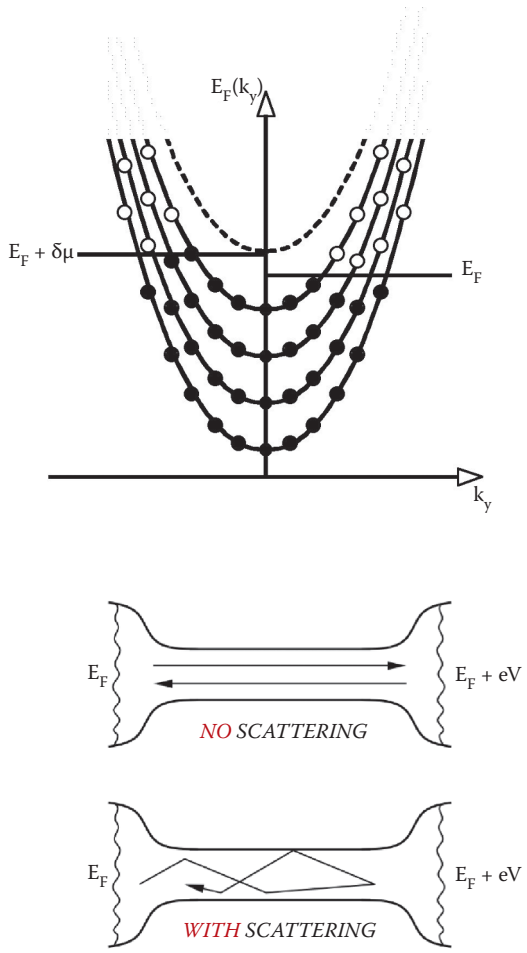


FIGURE 3.139 When there is no scattering in the wire, states are filled to higher energies in one direction than in the other.

Based on Equation 3.274, the resistance of a one-channel system [$M(E_F) = 1$] can be rewritten as:

$$\text{Resistance}\,(E_F) = \frac{h}{2e^2}\frac{1}{T(E_F)} = \frac{h}{2e^2}\left[1+\frac{1-T(E_F)}{T(E_F)}\right]$$
$$= \frac{h}{2e^2} + \frac{h}{2e^2}\frac{R(E_F)}{T(E_F)}$$
$$= \text{quantized contact resistance} + \text{scattering from barriers inside channel} \qquad (3.275)$$

where $R(E_F) = 1 - T(E_F)$ is the reflection coefficient, and $T(E_F) + R(E_F) = 1$. Therefore, the resistance of the quantum wire is the sum of the quantized contact resistance and a second term, as a result of scattering from barriers in the channel. It is the latter term that is zero for a perfect conductor.

At finite temperatures a thermal smearing out of the conductance staircase is observed because the Fermi-Dirac energy distribution of electrons in the left and right leads must now be taken into account. This smearing out effect is observed in Figure 3.140, where the conductance voltage plots for the same type of quantum wires of Figure 3.138 are shown at 0.3 K, 0.6 K, 1.6 K, and 4.2 K. At the highest temperatures the staircase profile is almost lost as transport becomes more and more incoherent because of dephasing, resulting from inelastic scattering, phonons, etc.

Quantum Dot: $\lambda_{DB} > L_{1,\,2,\,3}$

Quantization In quantum dots (QDs), generally consisting of 10^3 to 10^5 atoms and ranging from nanometers to tens of nanometers in size, dimensions are smaller than the de Broglie wavelength of thermal electrons in all three directions (L_1, L_2, $L_3 < \lambda_{DB}$). A quantum dot was first illustrated in Figure 3.67 and, simplified, is reproduced in Figure 3.141. Charge carriers in these "artificial atoms" are dimensionally confined in three directions (zero degrees of freedom), just like in an atom, and similarly have only discrete energy levels. The energy levels for an electron in a quantum dot are based on Equation 3.160, where we introduce the effective electron mass m_e^* and, for generality, make the dimensions $L_1 \# L_2 \# L_3$, to obtain:

$$E_{n_1,n_2,n_3} = \frac{h^2}{8m_e^*}\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2} + \frac{n_3^2}{L_3^2}\right) \qquad (3.276)$$

where the quantum numbers n_1, n_2, and n_3, each assuming the integral values 1, 2, 3,…, characterize quantization along the x-, y-, and z-axes, respectively. When considering the density of states for a zero-dimensional (0D) electron gas, no free motion is possible: there is no **k**-space to be filled with electrons, and all available states exist only at discrete energies. Because carriers cannot move there is no kinetic energy term $\left(\frac{\hbar^2\mathbf{k}^2}{2m_e^*}\right)$ here, only the quantized energy terms survive.

In Figure 3.142 we show the band structure and the density of states for a quantum dot.

Quantum dots exhibit mesoscopic phenomena that are not observed in "simple" atoms. Mesoscale describes structures and behaviors at lengths between nanometers and 0.1 micrometers—too large for molecular or atomic characterization methods but too small to be fully characterized by their bulk behavior or to be modeled using continuum models. Phenomena at that scale, however, determine the properties of many scientifically and commercially important materials.

Fermi Surfaces and Brillouin Zone for Quantum Dots There is no Fermi surface for quantum dots

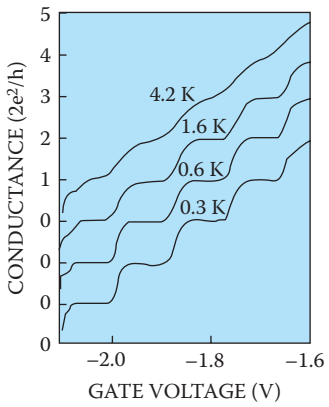


FIGURE 3.140 Conductance in number of units of $2e^2/h$ as a function of gate voltage (V) and at different temperatures. Same quantum wires as studied in Figure 3.138.

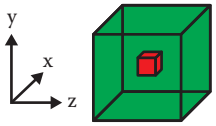


FIGURE 3.141 A cubic quantum dot.

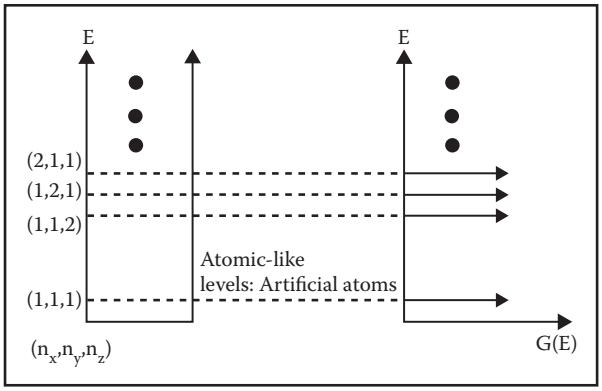


FIGURE 3.142 Energy levels and DOS of a quantum dot.[4]

because electrons are confined in all three directions (L_1, L_2, $L_3 < \lambda_{DB}$) and cannot freely move. It is important to figure out for different solids at what length scale charge carriers become frozen in fixed energy states. The Fermi wavelength is a most important parameter for this as it tells us how finely a solid needs to be structured to have an impact on its electronic properties. The size of an electron in a solid is essentially its Fermi wavelength, $\lambda_F = 2\pi/\mathbf{k}_F$, with $\mathbf{k}_F$ given by $(3n\pi^2)^{1/3}$ (Equation 3.238), where n is the electron density. In most metals the electron density is very large ($n > 10^{21}$ cm^{-3}), and the value of the Fermi wavelength is consequently of the order of a few nanometers. In semiconductors, however, the lower carrier density ($n < 10^{21}$ cm^{-3}) gives rise to larger electron wavelengths of several tens of nanometers (see also Table 3.12). Therefore, confinement and quantization phenomena are visible in semiconductors already at dimensions greater than 200 nm, whereas in metals they typically are seen at 1–2 nm. This is one important reason why quantum effects have been observed mainly in semiconductor nanostructures, which have large Fermi wavelengths, rather than in metals, which have Fermi wavelengths comparable with a few crystal lattice periods only. We can also approach the difference in quantization effects between a metal and semiconductor nanoparticle from the point of view of the separation between adjacent energy levels. The average level spacing at the Fermi level between subsequent levels in a metal can be estimated from:

$$\Delta E_{n+1-n} = \frac{1}{G(E)_{3D}(E = E_F)} = \frac{2E_F}{3n_{3D}} \qquad (3.237)$$

For a spherical Au nanoparticle with a radius R = 2 nm, the average level spacing $\Delta E = E_{n+1} - E_n$ is 2 meV only. In contrast, we calculate the energy separation between adjacent levels for a 2-nm CdSe particle as 0.76 eV (see Volume III, Chapter 3). So obviously energy-level quantization is far more pronounced in semiconductor dots than in metal dots. Beyond the Fermi wavelength difference (or electron density difference) there are other reasons why semiconductors show quantum effects at a larger length scale: first, the energy level spacing—the energy difference $E_{n+1} - E_n$—increases for low-lying states in a 3D potential well; second, the effective mass for electrons is smaller in semiconductors than in metals, and according to Equation 3.235 this also decreases the effective electron density n; and, third, the appearance of the plasmon oscillations of metal particles, described in Chapter 5 on photonics, often obscures quantization effects in metals.

A Brillouin zone describes a material's periodicity and is not defined for a quantum dot.

Density of States Function for Quantum Dots

Because band structure is no longer a term that has any relevance in a quantum dot, one describes the density of states for 0D (0 degrees of freedom) with a summation of delta functions δ(E) (sharp DOS peaks):

$$G(E)_{0D} = 2\sum_n \delta(E - E_n) \qquad (3.277)$$

where 2 is for the spin degeneracy. Such discrete atom-like densities of states are illustrated in Figure 3.143. The solid black curve is that of an electron in a large 3D semiconductor.

As we just saw above, the relative importance of the quantum effect in a given system is determined

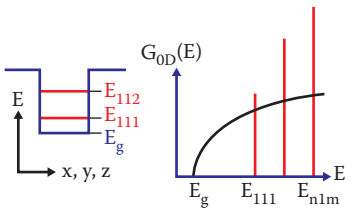


FIGURE 3.143 The density of states function for a quantum dot. The solid black curve is that of a free electron (see also Figure 3.119, full line), for example, in self-organized III-V quantum dots (e.g., InAs/GaAs).

by the energy differences between the energy levels: $\Delta E = E_{n+1} - E_n$. Quantum size effects are only observable when this energy separation is larger than the thermal energy of the carriers (kT), so that adjacent sub-bands (e.g., the E_{111} and E_{112} levels in Figure 3.143) are differently populated.

Electronic Conductivity of Quantum Dots To illustrate conduction through a quantum dot we introduce a type of quantum dot that is made by the electrostatic confinement of a 2D electron gas through a set of electrodes whose voltage can be varied at will. The specific example shown in Figure 3.144 is that of a lateral quantum dot (Figure 3.144a) made by applying a voltage on a set of gate electrodes (Figure 3.144b) to deplete the electrons in the underlying $Al_{0.3}Ga_{0.7}As$ layer and thereby defining the lateral size of a quantum dot in the 2D electron gas at the $Al_{0.3}Ga_{0.7}As$/GaAs heterojunction (Figure 3.144c). The potential well forms as a result of charge transfer between the

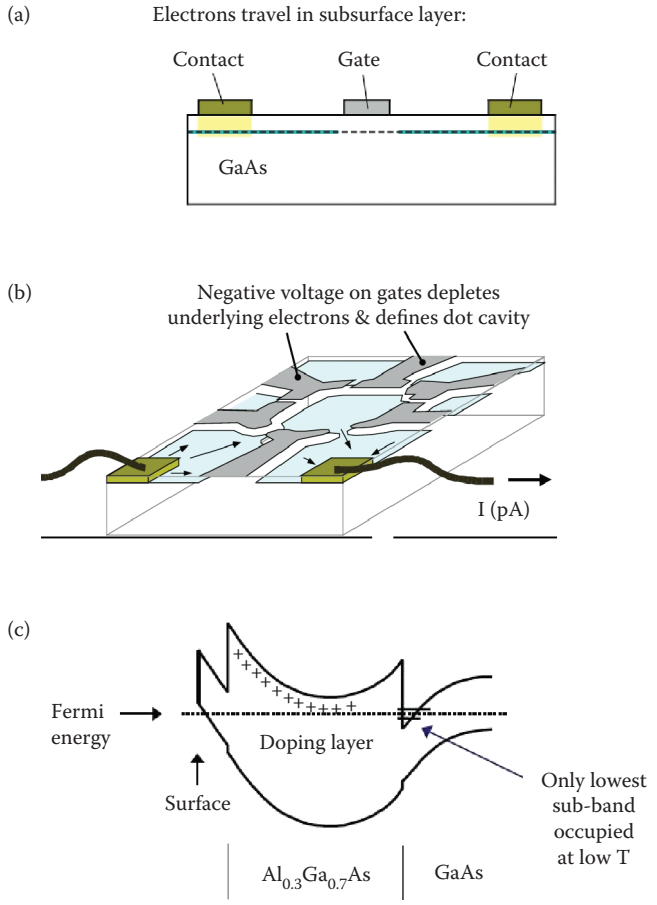


FIGURE 3.144 Electrostatically confined quantum dot in a 2D electron gas. (a) physical configuration. (b) gate electrodes and contacts. (c) energy diagram.

two materials and their conduction band offset (E_c), and it confines the electrons so that the motion of electrons perpendicular to the plane of the heterointerface is quantized. The gate electrodes then further confine this 2D gas in the lateral direction to make a quantum dot (3D confinement).

Any movement of electrons through a quantum dot requires the number of electrons in the dot to change one at a time. The Coulomb repulsion between electrons in the dot means that the energy of a dot containing $N+1$ electrons is higher than one containing N electrons. This in turn requires extra energy to add an electron to the dot, and no current flows until increasing the voltage provides this energy. This is known as the Coulomb blockade.

To better visualize the electronic transport through a quantum dot, consider a classical capacitor and compare it with a quantum dot between contact pins (Figure 3.145a). In the case of the macroscopic capacitor, the charge on the plates is given as $Q = CV$ and the displacement current as $I(t) = dQ/dt$, and no charge can flow through this capacitor because the dielectric is an insulator. The electrostatic energy in the capacitor is then given as:

$$E = \int_0^\infty V(t)I(t)dt = \frac{Q^2}{2C} \tag{3.278}$$

In the case of a quantum particle contacted by two measuring electrodes, one can envision that the contact points form a capacitor on either side of the quantum dot; if an electron is transferred by tunneling across these capacitors (Figure 3.145a), its Coulomb energy is changed by:

$$E_c = e^2/2C \tag{3.279}$$

where it is assumed that the capacitors on either side of the quantum dot are identical. To ensure that thermal motion does not initiate a charge flow and an accompanying change in E_c, kT should be much smaller than the Coulomb blockade ($kT \ll e^2/2C$). This means that C must be less than an attofarad (10^{-18} F). The tunneling current is given as:

$$I = V_c/R_T \tag{3.280}$$

Current starts to flow at the Coulomb voltage, or:

$$V_c = e/2C \tag{3.281}$$

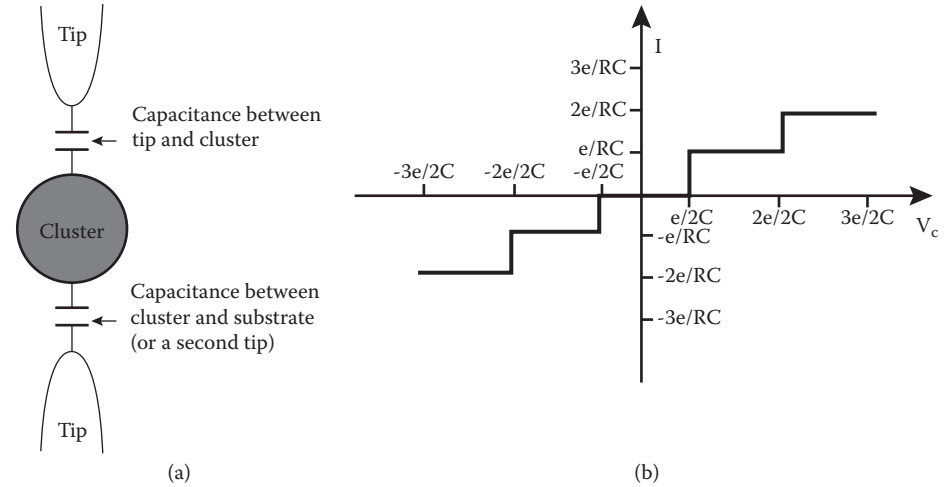


FIGURE 3.145 (a) Quantum dot probed by two pins. (b) Current-voltage plot of a quantum dot.

The electrostatic potential of the quantum dot increases significantly by introducing just one electron. Further electron transfer happens if the Coulomb energy of the quantum dot is compensated by an external voltage $V_c = \pm ne/2C$, where n is an integer. Repeated tunneling results in a "staircase" with step height in current, e/RC, as shown in Figure 3.145b.

The electron density function of a quantum dot (Equation 3.277) is illustrated in Figure 3.146: because $G(E)$ is peaked in this case (blue line spikes), so is the electron density function n (red spikes).

Generalizations and Summary

Density of States Functions From the previous sections, quantum confinement in different dimensions gives rise to very different densities of states functions compared with the situation in a bulk material, and this causes dramatic modification of the electrical and optical behavior. The higher the degree of the confinement, the fewer degrees of freedom, and the more the density of states becomes singular. Generalizing the density of states $G(E)_D$ varies with the dimension D as:

$$G(E)_D = \int_{\sum(E)} \frac{d\sum}{4\pi^3} \frac{1}{|\nabla E(\mathbf{k})|} \qquad (3.282)$$

where

$\sum(E)$ is the equi-energetic as described for each D and

$\Delta E(\mathbf{k})$ is the gradient of E with respect to **k**. As we have seen the dependence of the density of states on the inverse of the gradient of E results in divergences in the DOS functions for lower dimensions.

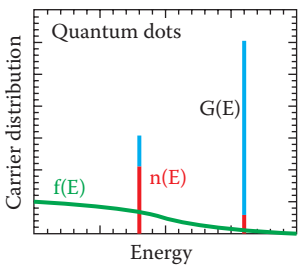


FIGURE 3.146 Density of states (DOS) (blue) and occupied states (red) around the Fermi energy for a quantum dot. The Fermi-Dirac function is $f(E)$, and the product of $f(E)G(E) = n_D$.

For a computer program using Mathematica to calculate the densities of states in 1D, 2D, and 3D, visit Dr. Fretwell's web site at http://pyweb.swan.ac.uk/~fretwell/DOS.pdf and download densities of states.

Fermi Quantities We can also generalize Equation 3.237 for $G(E)_{3D}(E = E_F)$ the density of single-particle states available per unit volume and unit energy at the Fermi level in 3D, to:

$$G(E)_{nD}(E = E_F) = \left.\frac{dn_{nD}(E)}{dE}\right|_{E=E_F} \qquad (3.283)$$

where n_{nD} is now the spatial electron density in nD dimensions: $n_{3D} = N/V$ for 3D (volume), $n_{2D} = N/A$

for 2D (area), and $n_{1D} = N/L$ (length) for 1D. From these definitions, we can find the spacing of single-particle levels in a piece of material, as well as all the other Fermi quantities described earlier. There is a qualitative change in energy dependence as the dimensions of a block of material are varied! Other things maintained equal, a higher $G(E)$ means levels are more closely spaced.

Fermi's Golden Rule in Solids

In quantum physics, Fermi's golden rule is a way to calculate the probability of a transition between two eigenstates of a quantum system, like the simple one illustrated in Figure 3.147. As sketched here, electronic levels in a solid are broadened because of atom-atom interactions. Consequently, when shining light on a semiconductor, as shown here, there is no single absorption line, but many transitions are feasible.

Fermi's golden rule (FGR), which was derived by Dirac, originates in the time-dependent perturbation theory. Time-dependent perturbation theory enables one to treat excitations that explicitly depend on time. This is of importance, for example, in discussing optical transitions where one can view the solid, with unperturbed Hamiltonian H_0, as being perturbed by a time-dependent electromagnetic field $H'(t)$ generated by an incident photon flux as:

$$\underset{\text{solid}}{H} = \underset{\text{solid}}{H_0} + \underset{\text{photon}}{H'(t)} \tag{3.284}$$

Fermi's golden rule for absorption describes transition rates between energy levels in terms of the availability of states [the density of states (DOS) $G(E)$], the availability of photons [intensity $(\mathbf{E}_0^2)$], and a "coupling strength" or transition matrix element between the levels ($|\langle n|H'|n+1\rangle|^2$). The transition rate (transitions per second) between an initial state n and a final state $n + 1$ on illumination with photons is given by the FGR as:

$$\Gamma_{n\to n+1} = \frac{2\pi}{\hbar}\mathbf{E}_0^2\left|\langle n|H'|n+1\rangle\right|^2 \times \delta(E_n - E_{n+1} - \hbar\omega) \tag{3.285}$$

where E_0 = the intensity of the photon radiation
$\hbar\omega$ = the photon energy
$\delta(E_n - E_{n+1} - \hbar\omega)$ = density of states (units C·m) at an energy $E_n - E_{n+1} - \hbar\omega$.

Note that $\langle n|H'|n+1\rangle|^2 = 0$ corresponds to a forbidden transition. Let us apply this expression to the transition of electrons from a filled valence to an empty conduction band, where $|\langle n|H'|n+1\rangle|$ becomes $|\langle v|H'|c\rangle|$. In theory, we need to calculate this matrix element for each set of initial and final states. In practice in solids, the matrix element does not vary strongly as we consider successive band states, so the total transition rate can be approximated as:

$$\begin{aligned}\Gamma &= \sum_{v,c}\Gamma_{v,c} \\ &= \mathbf{E}_o^2\sum_{v,c}\frac{2\pi}{\hbar}\left|\langle v|H'|c\rangle\right|^2\delta_{v,c}\,[E_c(\mathbf{k}) - E_v(\mathbf{k}) - \hbar\omega] \\ &\cong \mathbf{E}_0^2\frac{2\pi}{\hbar}\left|\langle v|H'|c\rangle\right|^2\sum_{v,c}\delta_{v,c}[E_c(\mathbf{k}) - E_v(\mathbf{k}) - \hbar\omega]\end{aligned} \tag{3.286}$$

The joint density of states in this expression is calculated from:

$$G_{v,c}(E) = \frac{2}{8\pi^3}\int d\mathbf{k}\,\delta_{v,c}[E_c(\mathbf{k}) - E_v(\mathbf{k}) - \hbar\omega] \tag{3.287}$$

and thus:

$$\Gamma = \mathbf{E}_o^2\frac{2\pi}{\hbar}\,|\langle v|H'|c\rangle|^2\;G_{v,c}[E_c(\mathbf{k}) - E_v(\mathbf{k}) - \hbar\omega] \tag{3.288}$$

Because of the form of the expression of joint density of states in Equation 3.287, optical measurements provide information about particular points in the Brillouin zone: usually high symmetry points near the band extremes. To get a better insight into this, let us perform the integral in Equation 3.287 in terms of energy dE instead of $d\mathbf{k}$. Consider a surface

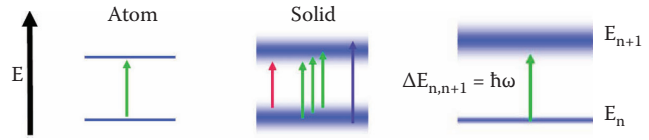


FIGURE 3.147 In solids, electronic levels broaden because of atom-atom interactions. Here we want to calculate the probability of a transition between two eigenstates E_n and E_{n+1} by using Fermi's golden rule.

FIGURE 3.148 Energy surface *S* in **k**-space.

S in **k**-space such that $E_c(\mathbf{k}) - E_v(\mathbf{k}) = \hbar\omega$ as sketched in Figure 3.148, where we can write:

$$dE = \left|\nabla_k E\right| d\mathbf{k}_n$$

or

$$\left|\nabla_k (E_c - E_v)\right| d\mathbf{k}_n = d(E_c - E_v) \text{ and thus}$$

$$d\mathbf{k} = dS d\mathbf{k}_n = dS \left\{ \frac{d(E_c - E_v)}{\left|\nabla_k (E_c - E_v)\right|} \right\} \tag{3.289}$$

The joint density of states is then:

$$G_{v,c}(E) = \frac{2}{8\pi^3} \int_{\text{k-space}} \frac{dS\, d(E_c - E_v)\delta[E_c(\mathbf{k}) - E_v(\mathbf{k}) - \hbar\omega]}{\left|\nabla_k (E_c - E_v)\right|} \tag{3.290}$$

Integrating over $d(E_c - E_v)$ results in:

$$G_{v,c}(E) = \frac{2}{8\pi^3} \int \frac{dS}{\left|\nabla_k (E_c - E_v)\right|_{E_c - E_v = \hbar\omega}} \tag{3.291}$$

At critical points, i.e., at Van Hove singularities, where $|\nabla_k (E_c - E_v)| \to 0$, large contributions to $G_{v,c}(E)$ occur. At these points the bands are parallel to each other; this can be recognized more simply by considering that in the density of states $G(E) = G(k)(dE/d\mathbf{k})^{-1}$ singularities emerge when $d\mathbf{k}/dE = 0$ (see "Brillouin Zones for Bulk Materials"). Critical points fall then in different categories depending on whether the band separations are increasing or decreasing as we move away from the critical point. This is illustrated in a detailed Si absorption spectrum in Figure 3.149. The peaks E_1 and E_2 in Figure 3.149a correspond to the **k**-values where the conduction and valence bands are parallel to each other as shown in Figure 3.149b.

For a direct transition between parabolic bands as sketched in Figure 3.150a, the difference $E_c(\mathbf{k}) - E_v(\mathbf{k})$ is given as:

$$E_c(\mathbf{k}) - E_v(\mathbf{k}) = E_g + \frac{\hbar^2 \mathbf{k}^2}{2}\left(\frac{1}{m_e^*} + \frac{1}{m_h^*}\right) = \frac{\hbar^2 \mathbf{k}^2}{2m_r^*} \tag{3.292}$$

where we introduced the reduced effective mass:

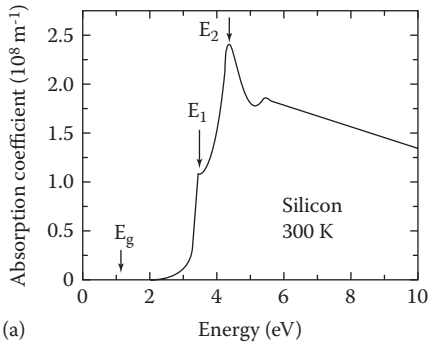
$$\frac{1}{m_r^*} = \frac{1}{m_e^*} + \frac{1}{m_h^*} \tag{3.293}$$

The gradient of $E_c(\mathbf{k}) - E_v(\mathbf{k})$ is then:

$$\nabla E_c(\mathbf{k}) - E_v(\mathbf{k}) = \frac{\hbar^2 \mathbf{k}}{m_r^*} \tag{3.294}$$

Substituting Equation 3.294 in Equation 3.291, we find:

$$G_{v,c}(E) = \frac{2}{8\pi^3} \int \frac{dS}{\left|\nabla_k (E_c - E_v)\right|_{E_c - E_v = \hbar\omega}} = \frac{2}{8\pi^3} \left[\frac{4\pi k^2}{\frac{\hbar^2 \mathbf{k}}{m_r^*}} \right] = \frac{m_r^*}{\pi^2 \hbar^2} \mathbf{k} \tag{3.295}$$



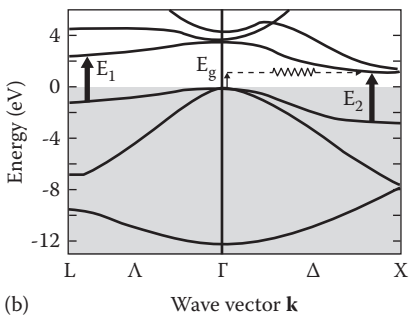


FIGURE 3.149 (a) Absorption spectrum for Si. (b) Si band structure. The indirect bandgap is at 1.1 eV, and critical points are at E_1 (3.2 eV) and E_2 (4.3 eV) semiconductors compared.

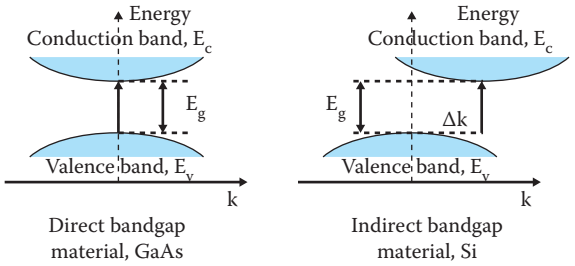


FIGURE 3.150 Direct and indirect bandgaps compared in an *E*-**k** diagram.

To evaluate **k**, note that $E_c(\mathbf{k}) - E_v(\mathbf{k}) = E_g + \frac{\hbar^2\mathbf{k}^2}{2m_r^*}$ (Equation 3.292), so that:

$$\mathbf{k} = \left[\frac{2m_r^*}{\hbar^2}(\hbar\omega - E_g)\right]^{\frac{1}{2}} \tag{3.296}$$

And, finally, the joint density of states for a two-parabolic band system is:

$$G_{v,c}(E) = \frac{1}{2\pi^2}\left[\frac{2m_r^*}{\hbar^2}\right]^{\frac{3}{2}}(\hbar\omega - E_g)^{\frac{1}{2}} \tag{3.297}$$

Substituting this result in Equation 3.288 yields:

$$\Gamma = \mathbf{E}_0^2\frac{2\pi}{\hbar}\left|H'_{v,c}\right|^2\frac{1}{2\pi^2}\left[\frac{2m_r^*}{\hbar^2}\right]^{\frac{3}{2}}(\hbar\omega - E_g)^{\frac{1}{2}} \tag{3.298}$$

We can relate the transition rate Γ to the absorption coefficient α of a material. Because the absorption rate Γ of photons (with an energy $E = \hbar\omega$) by a material corresponds to an absorbed power per volume, *W*, we obtain:

$$\alpha = W = \hbar\omega\Gamma = -dI/dV \tag{3.299}$$

where *I* is the irradiance, and because $\mathbf{E}_0^2 = \frac{2I}{nc\varepsilon_0}$, we obtain:

$$\frac{dI}{dV} = -\frac{2\pi}{\hbar}\left|H'_{v,c}\right|^2\hbar\omega\frac{2I}{2\pi^2nc\varepsilon_0}\left[\frac{2m_r^*}{\hbar^2}\right]^{\frac{3}{2}}(\hbar\omega - E_g)^{\frac{1}{2}}$$

or

$$\alpha(\omega) = \frac{2\omega}{\pi nc\varepsilon_0}\left|H'_{v,c}\right|^2\left[\frac{2m_r^*}{\hbar^2}\right]^{\frac{3}{2}}(\hbar\omega - E_g)^{\frac{1}{2}} \tag{3.300}$$

This shows that the absorption coefficient α depends on the transition matrix and on the density of states. The transition matrix element is very difficult to calculate; however, in practice, sufficiently close to the band edge, the product of ω and the matrix element remains approximately constant over a significant frequency range, so that we may write:

$$\alpha(\omega) \propto G(E) \propto (\hbar\omega - E_g)^{\frac{1}{2}} \tag{3.301}$$

In other words, a plot of the absorption coefficient versus frequency is a good reflection of the density of states function. In the case of an indirect bandgap semiconductor, absorption is a second-order process involving both a phonon and a photon. In this case, one obtains for the absorption:

$$\alpha(\omega) \propto G(E) \propto (\hbar\omega - E_g \pm \hbar\nu_p)^2 \tag{3.302}$$

with ν_p the phonon frequency and with + for absorption and – for emission.

In Chapter 5, we will see that for many semiconductors, the effect of excitons (a bound electron-hole state) causes the density of states to be enhanced for $\hbar\omega \approx E_g$, so that the absorption coefficient in reality might have a rather different shape than the predicted dependence (see the absorption curves for various semiconductors in Figure 5.93).

Fermi's golden rule applies to all types of quantum transitions, not only simple photon absorption. The expression can be expanded to cover spontaneous emission, stimulated emission, tunneling, scattering phenomena like phonon-electron scattering, and so on. The golden rule is especially useful in the case of isolated atoms where electron wave functions have been calculated theoretically or in cases where the density of states is the most rapidly varying function, so that the shape of the absorption spectrum is dominated by the DOS. Apart from excitonic effects (see Chapter 5) and selection rules, the dimensionality of the structure has little influence on transition probability.

Heat Transport and Heat Capacity

Introduction

Using quantum mechanics, we now understand the photoelectric effect, avoided the ultraviolet catastrophe, clarified the light spectra emitted from a hot gas, and gave a firm scientific base to Dmitri

Mendeleev's periodic table. With the band model of solids, the difference between conductors, semiconductors, and insulators also became clear. Curiously, we also found that in quantum mechanics, there is a finite probability of finding a particle inside a "classically forbidden" region, i.e., inside a "barrier." We still need to demonstrate how quantum mechanics correctly explains thermal transport, heat capacity, and their temperature dependencies.

Specific Heat and Thermal Transport

Electron Gas Contribution to Specific Heat Earlier in this chapter we described how, experimentally, C_v, the sum of vibrational (phonon) and electronic contributions (electrons) to the heat capacity ($C_v = C_{v,el} + C_{v,lat}$) is proportional to T^3 at low temperatures while approaching a constant 3R at high T, and how the electronic contribution $C_{v,el}$ is found to vary linearly with temperature (see Figure 3.16). From classical theory, at high temperature, electrons are expected to contribute to the lattice heat capacity for an amount of $3/2R$, something that is not observed experimentally. The total C_v for metals is found to be only slightly higher than that for insulators, which feature lattice contributions only! We explore here how QM comes to the rescue in this case. We first consider the electronic contribution to the specific heat and then consider Einstein and Debye's calculations for the lattice contribution. Along the way we reveal details about the nature of motions of atoms in a lattice, i.e., about phonons.

In the classical model for specific heat capacity of an electron gas, we used the principle of equipartition of the energy from the classical ideal gas theory. This led to the wrong expectation of an electronic contribution to the specific heat of $3/2R$. The reason that electrons do not contribute to the heat capacity at room temperature is again, just like in Drude's conductivity case, because they are fermions and cannot be treated as an ordinary Maxwell-Boltzmann gas. At room temperature only the very few electrons in the Maxwell-Boltzmann tail of the Fermi distribution can contribute to the heat capacity. Thus, at ordinary temperatures the electronic heat capacity is almost negligible, and it is the vibrations of the atoms, i.e., the phonon contributions, that dominate (see Figure 3.16)!

Only electrons within an energy range of kT of the Fermi level can be thermally excited, and the minimum energy required for exciting all the free electrons is E_F. The fraction of electrons that can be excited at T = 300 K, say in the case of copper is given by $kT/E_F = 0.025/5 = 0.005 = 0.5\%$. Therefore, at room temperature less than 1% of the available electrons contribute to the heat capacity, whereas in the classical view all electrons are supposed to contribute. Because each electron can absorb a kinetic energy of 3/2 kT, the energy of a kilomole of metal with some simplifications can be written down as:

$$U = U_0 + n\left(\frac{kT}{E_F}\right)\left(\frac{3}{2}kT\right)$$
$$= U_0 + \left(\frac{nk}{E_F}\right)\left(\frac{3}{2}kT^2\right) \qquad (3.303)$$

Or the electronic contribution to the heat capacity is given by:

$$C_{v,el} = \left(\frac{dU}{dT}\right) = 3nk\left(\frac{k\,T}{E_F}\right) \qquad (3.304)$$

This expression has the correct linear temperature dependence (see Figure 3.16), and in the case of copper one obtains a room temperature value for $C_{v,el}$ of $3R \times 0.005$ or $C_{v,el} = 0.015R$, whereas the lattice contribution $C_{v,lat}$ is $3R$. The value of $C_{v,el} = 0.015R$ agrees well with experimental data.

At very low temperatures, the lattice contribution falls off as T^3 and becomes small, whereas the electronic contribution turns significant. A more detailed analysis (for example, see Pillai[23]) leads to the following expression for U in a 3D structure:

$$U = U_0 + \frac{\pi^2}{6}(k_B T)^2 G(E = E_F)_{3D} \qquad (3.305)$$

Or the specific heat for a 3D case is:

$$C_{v,el} = \frac{dU}{dT} = \frac{\pi^2}{3}k_B^2 T G(E = E_F)_{3D} \qquad (3.306)$$

Because $G(E = E_F)_{3D}$ was derived earlier as:

$$G(E)_{3D}(E = E_F) = \left.\frac{dn_{3D}}{dE}\right|_{E = E_F} = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} E_F^{\frac{1}{2}}$$

(Equation 3.236), and the Fermi energy was calculated as:

$$E_F = \frac{\hbar^2}{2m_e^*}\left(3\pi^2 n_{3D}\right)^{\frac{2}{3}} = (0.365\ \text{eV.nm}^2)(n_{3D})^{\frac{2}{3}}$$

(see Equation 3.232), or also:

$$G(E)_{3D}(E = E_F) = \left.\frac{dn_{3D}}{dE}\right|_{E=E_F} = \frac{3n_{3D}}{2E_F} \qquad (3.237)$$

Thus, for the quantum electronic heat capacity we derive:

$$C_{v,el} = \frac{dU}{dT} = \frac{\pi^2}{3}k_B^2 TG(E = E_F)_{3D}$$

$$= \frac{\pi^2}{2}\left(\frac{k_B T}{E_F}\right)n_{3D}k_B \qquad (3.307)$$

again with the expected linear temperature dependence (see Figure 3.13). Different dimensionalities lead to different results as U in Equation 3.306 depends on the density of states $G(E)_{\#D}$. Classically, the heat capacity $C_{v,el}$ was given by $3/2R$ or $3/2nk_B$; from the above it follows that Fermi-Dirac statistics depress $C_{v,el}$ by a factor of $\frac{\pi^2}{3}\left(\frac{k_B T}{E_F}\right)$, or for 1 kmol of monovalent metal the heat capacity is:

$$C_{v,el} = \left(\frac{\pi^2}{2}\right)\left(\frac{kT}{E_F}\right)R = \left(\frac{\pi^2}{2}\right)\left(\frac{kT}{kT_F}\right)R$$

$$= \left(\frac{\pi^2}{2}\right)\left(\frac{T}{T_F}\right)R = AR$$

$$\text{with } A = \left(\frac{\pi^2}{2}\right)\left(\frac{T}{T_F}\right) \qquad (3.308)$$

The ratio of electronic specific heat calculated classically and quantum mechanically is:

$$\frac{C_{v,el}(QM)}{C_{v,el}(\text{Classical})} = \frac{\left(\frac{\pi^2}{2}\right)\left(\frac{T}{T_F}\right)R}{\frac{3}{2}R} = 0.5710^{-4}T \qquad (3.309)$$

The total specific heat C_v is the sum of vibrational and electronic contributions ($C_v = C_{v,el} + C_{v,lat}$). Equation 3.307 reveals a linear dependence of electronic specific heat with temperature, whereas the lattice specific heat varies as the cube of the absolute temperature. In other words:

$$C_v = AT + BT^3$$

$$\text{electron + phonon contributions to specific heat} \qquad (3.310)$$

with A and B constants. Because the lattice specific heat at high temperature is $3R$ (Dulong-Petit law) and is independent of temperature, Equation 3.310 may be rewritten as:

$$\frac{C_v}{T} = A + BT^2 \qquad (3.311)$$

Lattice Contribution to Specific Heat

Lennard-Jones Potential In the classical model for heat capacity C_v (Dulong-Petit law), it was assumed that a solid consists of a large number of atomic oscillators executing simple harmonic oscillations about their equilibrium positions (see Figure 3.14). Atoms vibrate about their ideal lattice positions because of the thermal energy in the lattice, and the vibration amplitude is about 10% of the equilibrium atomic spacing. The equilibrium distance between atoms in a lattice can be derived from the Lennard-Jones relationship as depicted in Figure 3.151 and detailed in Chapter 7 (see Figure 7.11). According to the Lennard-Jones model, neighboring atoms are subject to two distinct forces: a repulsive force at short ranges (Pauli repulsion from Pauli's exclusion principle) and in the limit of large distances an attractive force (van der Waals force). In 1931, taking into account both repulsive and attractive interactions, John Lennard-Jones of Bristol University made a good approximation of the total potential energy

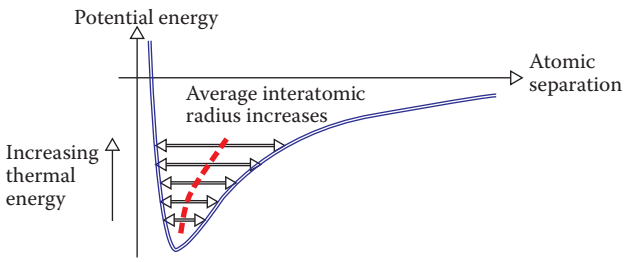


FIGURE 3.151 Lennard-Jones model approaching neutral atoms are subject to two distinct forces: a repulsive force at short ranges (Pauli repulsion from Pauli's exclusion principle) and in the limit of large distances an attractive force (van der Waals force).

of that interaction with the so-called Lennard-Jones (L-J) potential (also referred to as the L-J potential or the 6-12 potential) (see Equation 7.22). Close to equilibrium the L-J potential takes on a parabolic curve, typifying a simple harmonic oscillator.

At higher temperatures, the L-J curve is not symmetric, and the increased energy of the atoms leads to a change of average atomic spacing, which explains the thermal expansion of materials characterized by a thermal expansion coefficient α. If the curve is more symmetric, the effect is reduced and α is lower.

Classical and Quantum Oscillators The internal energy of a crystal, *U*, is caused by the vibrational energy of all the atomic oscillators. In the case of a classical harmonic oscillator, energy can take on a continuum set of values from zero to infinity. Using the classical oscillator model, Dulong and Petit derived that C_v must be $3R$, independent of temperature because each atom oscillator in the solid has three degrees of freedom. But we know from experiments that the specific heat for all solids decreases sharply at low temperatures and approaches zero at 0 K (see Figure 3.15). Clearly, the Dulong-Petit prediction is wrong for low temperatures.

We will see next that Einstein, as usual, came up with the solution to this difficulty, and in 1903 he proposed that the energies of the "atomic oscillators" were quantized. He considered a solid of *N* atoms to behave as $3N$ independent harmonic oscillators and that, unlike electrons, the number of these oscillators increases with temperature or, conversely, phonons are eliminated with decreasing temperature and the energy of each phonon is constant: for electrons it is their energy that increases with temperature, not their number. Like Planck before him, he assumed that the atom oscillators were quantized, with discrete values given by Planck's formula $E_n = nh\nu = n\hbar\omega$. Debye's model, an improvement on Einstein's model, retained the quantization of the atom oscillator energies but regarded them as coupled, with $3N$ modes belonging to the whole system rather than the *N* single atoms. Lattice vibrations play a most important role in a material's thermal and electrical conductivities. In particular, long-wavelength phonons give rise to sound in solids—hence the name phonon (voice in Greek). Phonons are quantum mechanical particles with integer spin, i.e., bosons that are found to obey the Bose-Einstein distribution. Classical oscillators were compared with quantum mechanical oscillators earlier in this chapter. For a classical harmonic oscillator the energy values are continuous from zero to infinity. Revisiting the harmonic oscillator from Figure 3.51 and solving the time-independent Schrödinger equation (TISE) for a quantum harmonic oscillator $-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} = (E - \frac{1}{2}kx^2)\psi(x)$ (Equation 3.189), we saw how energy quantization came about with energy steps given by $E_n = (n + \frac{1}{2})\hbar\omega$, n = 0,1,2,3, . . . (Equation 3.191). This energy quantization results in the evenly spaced energy levels shown in Figure 3.78. The horizontal lines cutting through the parabolic part of the curve represent the quantization of the vibration states. Also different from the classical oscillator is the fact that a quantum oscillator has a zero-point energy (ZPE) of $E_0 = \frac{1}{2}\hbar\omega$ (Equation 3.193). To describe phonons in solids, Debye went beyond the treatment of a set of these single oscillators and considered a network of interconnected oscillators. Phonons constitute a quantum mechanical version of the normal modes in classical mechanics, in which each part of a lattice oscillates with the same frequency.

Phonons

1D Monoatomic Lattice—No Light Interaction Possible In a crystal, the individual molecular vibrations are replaced by collective vibration modes of the whole lattice. Let us consider phonons in the simplest case of a 1D mechanical harmonic chain of length *L* with *N* identical atoms in a crystalline monoatomic 1D lattice as sketched in Figure 3.152. Here, a unit cell (length *a*, so that $N \times a = L$) contains

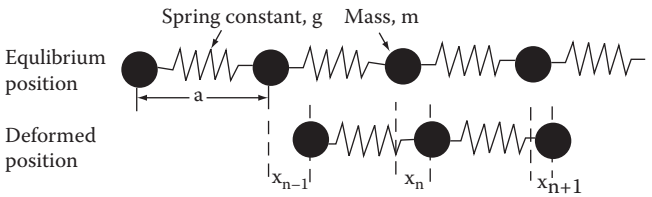


FIGURE 3.152 A 1D array of springs with all identical atoms with mass *M*. Note that *K* is a spring constant here, not a wave vector.

only one atom with mass m, and there is only one spring constant K (capitalized to distinguish it from the phonon wave vector $\mathbf{k}_{ph}$). This is the simplest mechanical model of a lattice, and we will analyze how phonons arise from it. The formalism that we develop for a 1D lattice is readily generalizable to 2D and 3D.

The atoms' equilibrium positions are na, where integer n labels atoms, but each atom executes a motion from equilibrium, which we denote as x_n. Classically the situation depicted in Figure 3.153 is described through an equation of motion where, for simplicity, we only consider nearest-neighbor interactions:

$$m\frac{d^2x_n}{dt^2} = K(x_{n+1} + x_{n-1} - 2x_n) \qquad (3.312)$$

This equation describes longitudinal (L) and transverse (T) modes for small amplitude vibrations with a restoring force that depends linearly on the atom's excursion. This expression has traveling wave solutions where the displacement of an atom relative to its center position is given as:

$$x_n = x_0 \exp(-i\omega t)\exp(in\mathbf{k}_{ph}a) \qquad (3.313)$$

with $\mathbf{k}_{ph}$ the phonon wave number (just like the electron case). This same result may be obtained from quantum mechanics. In the latter case $\mathbf{k}_{ph}$ takes on quantized values ($\mathbf{k}_{ph,n}$ is then the wave vector of the n^{th} mode) because the total number of atoms N is finite (Figure 3.153). For these modes, a single quantum of energy is called a phonon, and the total number of phonon modes depends on the total number of atoms N in crystal length L. The form of the quantization depends on the choice of boundary conditions; for simplicity, we impose periodic boundary conditions, defining the $(N+1)^{th}$ atom as equivalent to the first atom. Physically, this corresponds to joining the chain at its ends (reread the section on Born and von Karman's periodic boundary condition, p. 124 "The Born and von Karman's Periodic Boundary Condition"). The resulting quantization is:

$$\begin{aligned}
&\mathbf{k}_{ph,n}(\min) = \frac{\pi}{Na} = \frac{\pi}{L} \\
&\mathbf{k}_{ph,n} = \frac{\pi}{Na} = \frac{n\pi}{L} \\
&\text{for } n = 0, = 1,2,\ldots,N \\
&\mathbf{k}_{ph,n}(\max) = \frac{N\pi}{L} = \frac{\pi}{a}
\end{aligned} \qquad (3.314)$$

with the corresponding wavelengths:

$$\lambda_{\min} = 2a,\ \lambda_n = \frac{2L}{n},\ \text{and}\ \lambda_{\max} = 2L \qquad (3.315)$$

These simplest ("normal mode") cases are depicted in Figure 3.154, and any motion can be written as a superposition of the normal modes. Mode wavelengths are easy to find, but finding the energy for each these modes is a more complex issue.

In real crystals the number of atoms N is so large that the $\mathbf{k}_{ph,n}$-spacing and the corresponding frequency spacing are irrelevant so that vibration spectra look continuous.

Substituting the wave solutions for x_n, x_{n-1}, and x_{n+1} (Equation 3.313) in Equation 3.312, we obtain, after a lot of convenient cancellations, a relation of the vibration frequency as a function of the wave vector:

$$\omega_n^2 = \frac{2K}{m}(1 - \cos\mathbf{k}_{ph,n}a) \qquad (3.316)$$

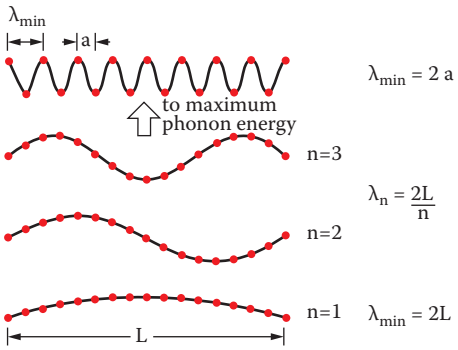


FIGURE 3.153 The simplest normal modes of lattice vibrations in a 1D crystal lattice of length L.

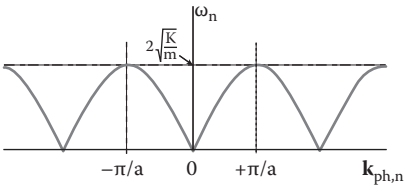


FIGURE 3.154 Phonon dispersion relation for a 1D monoatom crystal.

The vibration frequency ω_n depends only on the relative phase of the neighbors, and we now have a series of allowed modes with wave vector $\mathbf{k}_{ph,n}$ and angular frequency ω_n according to:

$$\omega_n = \sqrt{\frac{2K}{m}(1-\cos \mathbf{k}_{ph,n}a)} = 2\sqrt{\frac{K}{m}}\left|\sin\left(\frac{\mathbf{k}_{ph,n}a}{2}\right)\right| \quad (3.317)$$

This expression represents the dispersion relation for a 1D monoatomic lattice as shown in Figure 3.154. At low values of $\mathbf{k}_{ph,n}$ the oscillation frequencies are low as well, and the maximum oscillation frequency that a collective mode can have is $2\sqrt{\frac{K}{m}}$.

The maximum oscillation frequency occurs at $\mathbf{k}_{ph} = \pm\pi/a$; in other words, when adjacent atoms are 180° out of phase (the wave number π/a corresponds to a vibration wavelength of $2a$). As in the case of electron wave functions, a first Brillouin zone ranges from $-\pi/a < \mathbf{k}_{ph,n} < \pi/a$. For any wave numbers $> \pi/a$, the same physical configuration can be described by a wave number that is less than π/a: there is no new physical situation outside this zone.

In Figure 3.155 we focus for a moment on the positive $\mathbf{k}_{ph,n}$ side in the first Brillouin zone of the phonon dispersion relation. At $\mathbf{k}_{ph,n} = 0$ all atoms move in phase, and at $\mathbf{k}_{ph,n} = \pi/a$ neighboring atoms move in opposite directions. In Figure 3.155 we show both the longitudinal (LA) and transverse (TA) acoustic branches. The longitudinal branch lies above the transverse branch. For a 3D monoatomic crystal there is one longitudinal vibration mode and two transverse direction modes. In the case illustrated the two transverse waves are "degenerate," i.e., they have the same frequency. The energies involved here are typically in the range of hundreds of meV [$\omega_{max} \sim (10^{12}–10^{14})$ Hz or $\lambda \sim$ (5–500) μm and $h\nu \sim$ (2–200) meV]. These low energies are possible because the restoring force depends on an atom position difference, which can become very small for long wavelengths. With a thermal energy at room temperature of ~24 meV, many phonon modes are excited even at room temperature.

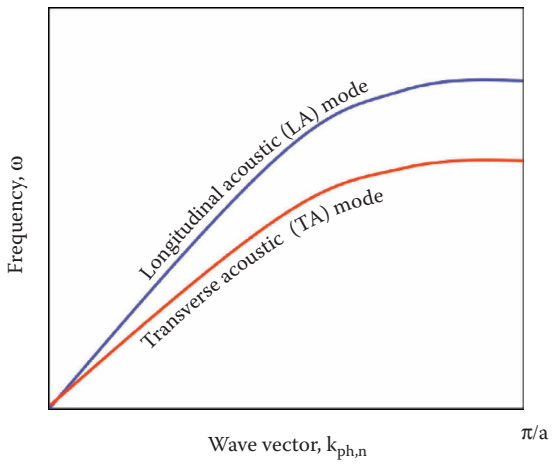


FIGURE 3.155 Positive **k** values in the first Brillouin zone for phonons.

In three dimensions, as mentioned above, there are three branches of the dispersion relation: two transverse and one longitudinal. Different crystal structures have different propagation speeds along different directions. In a monatomic lattice, all the phonon modes are called "acoustic" modes.

As illustrated in Figure 3.156, the slope of the dispersion relation is the group velocity v_g ($v_g = \partial\omega/\partial\mathbf{k}_{ph}$), whereas the phase velocity $v_{ph} = \omega/\mathbf{k}_{ph}$. At low values of $\mathbf{k}_{ph} \sim 0$ (i.e., long wavelengths), the dispersion relation is almost linear, and the speed of sound v_s through the material $v_s = v_g = v_{ph}$ is then represented as:

$$v_s = \left(\frac{d\omega}{d\mathbf{k}_{ph}}\right)\mathbf{k}_{ph\to 0} \quad (3.318)$$

At $\mathbf{k}_{ph} = \pm\pi/a$, the Brillouin zone boundary (BZB), the group velocity = 0. As a result, phonons with long wavelengths can propagate for large distances so that sound propagates through solids without significant distortion. This behavior fails at short

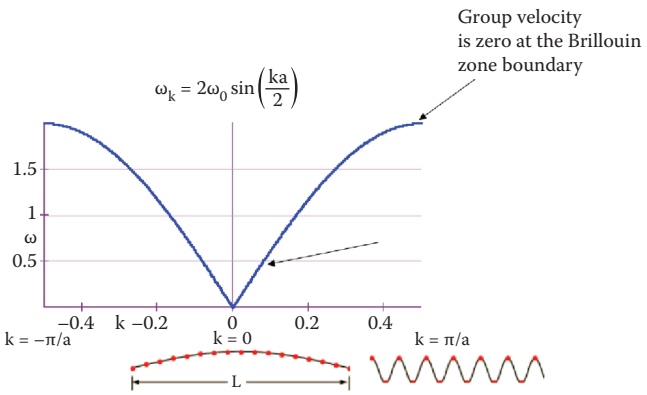


FIGURE 3.156 The slope of the dispersion relation is the group velocity v_g ($v_g = \partial\omega/\partial\mathbf{k}_{ph}$), whereas the phase velocity $v_{ph} = \omega/\mathbf{k}_{ph}$.

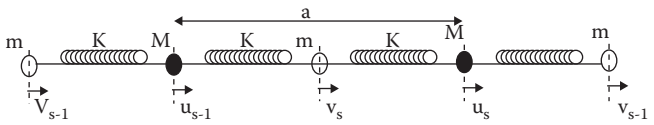


FIGURE 3.157 Vibrational modes in a crystalline diatomic 1D lattice.

wavelengths (large values of $\mathbf{k}_{ph}$) because of the microscopic details of the lattice. Sound waves in fluids only have longitudinal components, whereas sound waves in solids have longitudinal and transverse components. The reason behind this is that fluids cannot support shear stresses. Monatomic crystals only have acoustic branches. Light does not interact with the lattice modes of a monoatomic crystal; for that we need a lattice that has at least two different atoms (diatomic) so that a polarization **P** can be generated.

1D Diatomic Lattice—Light Interaction Possible A diatomic lattice is illustrated in Figure 3.157. Here we consider two types of atoms with masses m (light) and M (heavy) characterized by displacement amplitudes u_s and v_s, respectively. In this case, a is the length of the unit cell, and each unit cell contains one of each kind of atom. A wave vector of π/a here implies a phase difference of 180° between the behavior inside adjacent unit cells, not between adjacent atoms as in the 1D monoatomic lattice! This internal structure of the unit cell adds an internal degree of freedom, giving rise to two phonon modes for each value of the wave vector.

Assuming again that only next neighbors interact, the equations of motion lead to an expression of the order ω^4, resulting in two independent solutions in ω^2 for the dispersion relationship:

$$mM\omega^4 - 2K(m+M)\omega^2 + 2K^2(1-\cos\mathbf{k}_{ph}a) = 0$$

or

$$\omega^2 = \frac{K}{mM}\left\{(m+M) \pm \sqrt{(m+M)^2 - 2mM(1-\cos\mathbf{k}_{ph}a)}\right\} \quad (3.319)$$

Each solution is twofold degenerate because the electrical field E can be chosen in an arbitrary direction perpendicular to the propagation vector. If the masses and spring constants are known, this equation can be solved. The dispersion relation for a diatomic linear lattice is schematically shown in Figure 3.158, where we see that in solids with more than one atom in the smallest unit cell, there is a "phonon branch" for "acoustic" phonons (red curve) and an "optical branch" phonon (blue curve). Both types of phonons come with transverse and longitudinal modes. This makes for four possible types of phonons: LA, TA, longitudinal optical (LO), and transverse optical (TO), as summarized in Figure 3.159. Figure 3.158b shows the phonon dispersion for GaAs, and all four types of phonons can be observed here.

"Acoustic phonons" have frequencies that become small at the long wavelengths and correspond to sound waves in the lattice. "Optical phonons" can only arise in crystals that have more than one atom in the smallest unit cell. They are called "optical phonons" because in ionic crystals they are excited

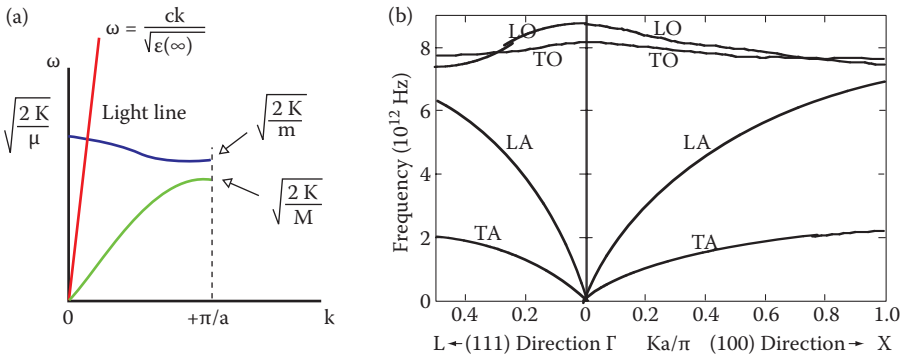


FIGURE 3.158 (a) Dispersion relationship for a diatomic linear lattice. Both optical and acoustic modes come with longitudinal and transverse waves. The straight line represents the "light line" assuming no interaction with phonons $\left(\omega = \frac{ck}{\sqrt{\varepsilon(\infty)}}\right)$ (see Chapter 5, "Summary: Generalized Dispersion Relation" for details). (b) The phonon dispersion for GaAs.

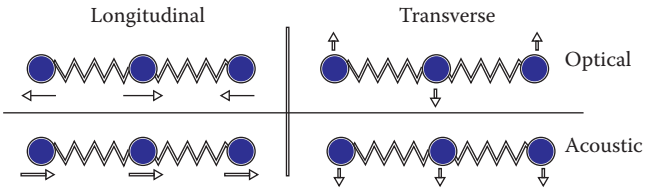


FIGURE 3.159 Four types of phonons: LA, TA, LO, and TO.

very easily by light (IR radiation). The underlying reason is that these vibrations correspond to a mode where positive and negative ions at adjacent lattice sites swing against each other, creating a time-varying electrical dipole moment **P**. Optical phonons that interact in this way with light are called infrared active. In the case of a 3D diatomic lattice there are three optical and three acoustic branches ($6N$ degrees of freedom). The three optical branches can interact with light. A phonon dispersion relation can be measured experimentally by inelastic x-ray or neutron scattering.

Let us investigate the maxima and minima values in the phonon relationship for diatomic 1D crystals. For $\mathbf{k}_{ph} = \pi/a$, where neighboring cells are out of phase, the two solutions are:

$$\omega^2 = \frac{2K}{M}(\text{acoustic mode}) \text{ and}$$
$$\omega^2 = \frac{2K}{M}(\text{optical mode}) \qquad (3.320)$$

For $\mathbf{k}_{ph}$ tending to zero ($\mathbf{k}_{ph}a \lll 1$) the acoustic branch tends to zero (see above), and the optical branch has the following (constant) solution:

$$\omega^2 \cong 2K\frac{m+M}{mM} = 2K\left(\frac{1}{m}+\frac{1}{M}\right) = \frac{2K}{\mu} \qquad (3.321)$$

where μ is the reduced mass of the unit cell $\left(\mu = \frac{1}{m}+\frac{1}{M}\right)$.

Example 3.8:
For GaAs $\approx \hbar\omega_{OP} \approx$ 36 meV, whereas for GaN $\hbar\omega_{OP}$ $\approx$ 90 meV; this difference stems from the much lighter *N* atom $\left[\omega_{OP}^2 \cong 2K\frac{m+M}{mM} = 2K\left(\frac{1}{m}+\frac{1}{M}\right)\right]$.

Interaction of Radiation with Lattice Modes—Polaritons The two types of atoms (m and M) in the unit cells of diatomic lattices in general come with slightly different charges and therefore experience different forces in different directions when an electromagnetic wave is applied. This leads to a macroscopic polarization **P**. The natural frequency of the oscillation of this polarization **P**, we will see, corresponds to the optical modes we just calculated above. In Figure 3.159a the straight line represents the "light line," assuming no interaction with phonons $\left(\omega = \frac{ck}{\sqrt{\varepsilon(\infty)}}\right)$ (for a detailed explanation see section in Chapter 5, "Summary: Generalized Dispersion Relation"). In polar materials, light does interact with the optical phonon modes and ε_r, the relative dielectric constant strongly varies, and ω versus k is not a straight line anymore. The wavelength of light is usually much greater than the lattice spacing ($\lambda \gg a$), especially for the infrared region, where we find that the frequency of the EM is resonant with the optical modes, so the incoming optical wave interacts strongly with very long wavelength phonons (compared with the lattice distance a).

At very long wavelengths, $\mathbf{k}_{ph}$ is close to 0 (see Figure 3.158), and we can assume that the applied field is spatially uniform [$E(t) = E_0e^{-i\omega t}$], which is true over many thousands of atomic spacings. The excursion of the m and M atoms driven by that uniform field is then given as:

$$u = u_0e^{-i\omega t}(\text{for M}) \text{ and } v = v_0e^{-i\omega t}(\text{for m}) \qquad (3.322)$$

Substituting these expressions in the equation of motion leads to:

$$-2Ku_0 + (2K - m\omega^2)v_0 = -qE_0 \qquad (3.323)$$

For this, the solutions for amplitudes u_0 and v_0 are:

$$u_0 = \frac{+qE_0/M}{\omega_T^2 - \omega^2} \text{ and } v_0 = \frac{-qE_0/m}{\omega_T^2 - \omega^2} \qquad (3.324)$$

where ω_T is the resonant frequency, also called the "restrahl" frequency, of the optical modes at $\mathbf{k} = 0$:

$$\omega_T = \sqrt{\frac{2K}{\mu}} \qquad (3.325)$$

This is the so-called transverse optical (TO) frequency, and μ is the reduced mass; this frequency

depends on the harmonic forces in the crystal and not the applied field. The large mass M results in a small amplitude u_0 and a large amplitude for the small mass m in v_0. The opposite signs of these amplitudes refer to the opposite motions that m and M make.

The excitation near ω_T (natural frequency of the optical $\mathbf{k} = 0$ branch) features a large amplitude, and this coupled excitation of a transverse optical phonon to electromagnetic radiation is called a polariton. In the longitudinal direction light and the lattice vibrations do not couple. There are six modes for a polariton because there are three optical phonons and three electromagnetic modes.

The polarization introduced by electromagnetic radiation in the lattice of a polar compound can then be written as:

$$\mathbf{P}_{\text{lattice}} = \text{Ne}(u_0 - v_0) \tag{3.326}$$

where e is the effective charge on the atoms in the lattice, and u_0 and v_0 are the amplitudes calculated in Equation 3.324. The total polarization of a material is given by the sum of this lattice contribution ($\mathbf{P}_{\text{lattice}}$) at low infrared frequencies, a bound electronic contribution ($\mathbf{P}_B$), and a free electron contribution ($\mathbf{P}_E$):

$$\mathbf{P} = \mathbf{P}_{\text{lattice}} + \mathbf{P}_B + \mathbf{P}_E \tag{3.327}$$

In Chapter 5 we connect the polarization $\mathbf{P}$ of a wide range of materials from insulators to metals to their susceptibility, their dielectric constant, refractive index, and absorption coefficient.

Einstein's Heat Capacity Equation Einstein, by introducing quantized noncoupled oscillators, obtained the following expression for the specific heat capacity contribution of the lattice:

$$C_v = 3R\left(\frac{h\nu}{kT}\right)^2 \left\{\frac{\exp\left(\frac{h\nu}{kT}\right)}{\left[\exp\left(\frac{h\nu}{kT}\right)-1\right]^2}\right\}$$

$$= 3R\left(\frac{\theta_E}{T}\right)^2 \frac{\exp\left(\frac{\theta_E}{T}\right)}{\left[\exp\left(\frac{\theta_E}{T}\right)-1\right]^2} = 3RF_E\left(\frac{\theta_E}{T}\right) \tag{3.328}$$

with $h\nu/k = \theta_E$ the Einstein temperature and F_E the Einstein function. At high temperatures ($h\nu \ll kT$) this equation simplifies to the Dulong-Petit form, i.e., $C_v = 3R$, whereas at low temperatures ($h\nu \gg kT$) it reduces to:

$$C_v = 3R\left(\frac{h\nu}{kT}\right)^2 \exp\left(\frac{-h\nu}{kT}\right) \tag{3.329}$$

At low temperature the exponential term becomes dominant and controls the temperature variation of C_v. An exponential decay of C_v agrees with experimental data, except at very low temperatures, where C_v is proportional to T^3.

Debye's Heat Capacity Equation Debye retained the quantization of the atom oscillator energies introduced by Einstein but regarded them as coupled, where $3N$ modes now belong to the whole system rather than the N single atoms. In this model, because of the interconnections between atoms in a solid, the displacement of one or more atoms from their equilibrium positions gives rise to a set of mechanical waves propagating through the material. The amplitude of these "solid" waves is given by the displacements of the atoms from their equilibrium positions. The general Debye expression is:

$$C_v = 9R\left(\frac{T}{\theta_D}\right)^3 \int_0^{\frac{\theta_D}{T}} \frac{e^4 x^4}{(e^x - 1)^2} dx = 9RF_D\left(\frac{\theta_D}{T}\right) \tag{3.330}$$

where $x = h\nu/kT$ and $x_{max} = h\nu_D/kT$ and $k\theta_D = h\nu_D$ with θ_D the Debye temperature. At high temperatures this expression again leads to Dulong-Petit, i.e., $C_v = 3R$, and at low temperatures ($T \ll \theta_D$) we get:

$$C_v = \frac{12}{5}\pi^4 R\left(\frac{T}{\theta_D}\right)^3 \tag{3.331}$$

The model correctly predicts a T^3 temperature dependence for the specific heat of a 3D structure at very low temperatures, and for the internal energy U it predicts a T^4 dependency.

The Einstein and Debye models were compared with experiments in Figure 3.16. Because phonons

follow Bose-Einstein statistics, their equilibrium distribution as a function of temperature is given as:

$$\left\langle n_{k_{ph,p}} \right\rangle = \frac{1}{\exp\left(\frac{\hbar\omega}{k_B T}\right) - 1} \tag{3.332}$$

The lattice energy is then equal to:

$$U = U_0 + \sum_p \sum_{k_{ph}} \left[\left\langle n(\omega_{k_{ph,p}}) \right\rangle + \frac{1}{2} \right] \hbar\omega_{ph,p} \tag{3.333}$$

where p stands for all polarizations of the phonons (LA, TA, LO, and TO). The second term on the right represents the energy $\hbar\omega_{k_{ph,p}}$ multiplied by the probability to have a state occupied at this energy and comes with a summation over all wave vectors k_{ph} and polarization states p. The summation over all wavevectors k_{ph} can be transformed into an integral $\int d\omega G(E) dE$. The energy density is then given by:

$$U' = \frac{U_0}{V} + \sum_p \int \left[\left\langle n(\omega_{k_{ph,p}}) \right\rangle + \frac{1}{2} \right] \hbar\omega_{k_{ph,p}} G(E) dE \tag{3.334}$$

where $G(E)$ is the density of states between ω and $\omega + d\omega$. This number of vibrational states can be derived from the dispersion relation $\mathbf{k}_{ph}(\omega)$ because:

$$G(E) = \frac{\mathbf{k}^2(\omega)}{2\pi^2} \frac{d\mathbf{k}}{d\omega} \text{ in 3D} \tag{3.335}$$

Or the specific heat is:

$$C_{lattice} = \frac{dU}{dT} = \sum_p \int \frac{d\langle n \rangle}{dT} \hbar\omega_{k_{ph,p}} G(E) dE \tag{3.336}$$

Debye assumed that $\omega = v_s \mathbf{k}_{ph}$, where v_s is the sound velocity so that the Debye density of states became:

$$G(E) = \frac{\mathbf{k}^2(\omega)}{2\pi^2} \frac{d\mathbf{k}}{d\omega} = \frac{\omega^2}{2\pi^2 v_s^3} \tag{3.337}$$

Putting this expression in Equation 3.336 leads to the Debye expression as given in Equation 3.330. Because $G(E)$ depends on the dimensionality, the specific heat of the lattice will also depend on the dimensionality of the structure (3D, 2D, 1D, or 0D). In 3D materials, phonons possess three polarizations: two transverse and one longitudinal, whereas in a perfect 1D material (a nanowire), only longitudinal polarizations are possible. In real 1D materials the surface gives rise to additional surface phonon modes. The resulting change in dispersion relation modifies the group velocity and the density of states. Generalizing for different dimensions D, one derives:

$$U \propto T^{D+1} \text{ and } C_{lattice} \propto T^D \tag{3.338}$$

with D = 1, 2, 3 the dimensions of the sample.

For an insulator, heat is carried entirely by phonons, whereas in metals, both electrons and phonons may transport heat. In most metals the conduction by electrons greatly exceeds that of the phonons; typically the phonon contribution at room temperature is only 1% of the electron contribution. As a metal is heated, the only way an electron can take up energy is by moving into a somewhat higher allowed energy level. But a typical electron is buried deep inside the Fermi sea, and there are no empty levels to move to because a pair of electrons of opposing spin already occupies each level above the electron. Only the relatively few electrons at the top of the distribution can find empty levels to move into. They are the only electrons that contribute to the thermal conductivity. The thermal heat conductivity was derived earlier as:

$$\kappa = \frac{1}{3} v_{dx} C_{v,el} \lambda \tag{3.59}$$

In those thermal conductivity calculations we used classical statistics to estimate the electron mean free path, λ, and the drift velocity, v_{dx}. These terms need to be replaced now by λ_F and v_F from quantum mechanics, and using the expression for $C_{v,el}$ we launched in Equation 3.308, we derive:

$$\kappa = \frac{\pi^2}{6} \left(\frac{n k_B^2 T}{E_F} \lambda_F v_F \right) = \frac{\pi^2}{6} \left(n k_B^2 T \lambda_F v_F \right) \frac{2}{m_e^* v_F^2}$$

and with $\tau_F = \frac{\lambda_F}{v_F}$

$$\kappa = \left(\frac{\pi^2}{3} \right) \left(\frac{n k^2 T \tau_F}{m_e^*} \right) \tag{3.339}$$

$C_{v,el}$ is expressed here per unit volume rather than per kilocalorie. The expression gives the thermal conductivity in terms of the electronic

properties of the material. A typical number, say for copper, is 387 $Wm^{-1}K^{-1}$. Using the quantum mechanical expression for electrical conductivity $\left[\sigma = \left(\frac{ne^2\tau_F}{m_e^*}\right)\right.$ from Equation 3.7$\left.\right]$ we obtain the ratio of thermal conductivity to electrical conductivity, or the Wiedemann-Franz law, as:

$$\frac{\kappa}{\sigma} = \frac{\pi^2}{3}\left(\frac{k_B}{e}\right)^2 T \tag{3.340}$$

(compare this with the classical result in Equation 3.60). The Lorenz number, L, defined as $\frac{\kappa}{\sigma T} \equiv L$ (see Equation 3.61), now becomes:

$$L = \frac{\pi^2}{3}\left(\frac{k}{e}\right)^2 \tag{3.341}$$

or 2.45×10^{-8} $W\Omega K^{-2}$. This is the correct answer; no Drude hanky-panky is required. The Lorenz number only depends on the universal constants k and e and therefore is the same for all metals.

Quantized Thermal Transport

Introduction Research in nanoscale thermal properties is recent compared with nanoscale electronic properties, which started in the 1960s. An important reason for this is the fact that measuring and modeling micro- and nanoscale thermal properties accurately is very challenging. An added difficulty is the fact that phonons are quasiparticles (eigenmodes) that can be annihilated or created at any time. Only since the mid-1980s and early 1990s have techniques become available that allow measuring temperature and heat flow at the nanoscale. Small thermocouples at the tip of atomic force microscopes in scanning thermal microscopy are successfully used to measure temperatures with <100-nm resolution. Scanning Seebeck voltage measurements have reached 2–3 nm spatial resolution with the use of sharp heated metallic tips in ultrahigh vacuum environments, and femtosecond laser technology has permitted the measurement of heat transfer and acoustic velocities in thin-film submicrometer structures. Thermoreflectance imaging using visible wavelengths is used to measure the surface temperature of IC chips with submicrometer spatial resolution and 1–100 mK temperature resolution. Very recently a phonon Hall effect under a magnetic field has also been observed. This could also provide a useful source of information about phonons and their transport in solids.

Macroscopic thermal transport theory, with Fourier's law (Equation 3.53) and the diffusion equation (Equation 3.54), breaks down for applications involving thermal transport in small length/time scales, e.g., nanoelectronics, nanostructures, NEMS, ultrafast laser materials processing, and so on. As we mentioned in the beginning of this chapter, this occurs when t is shorter than the scattering mean free time of the energy carriers (τ) and when L is shorter than the scattering mean free path of the energy carriers (λ).

Although classically temperature can in principle be defined for a single atom, the quantum mechanical definition can only be applied on the length scale of the phonon mean free path. Because most of the heat is carried by phonons with large wave vectors, the relevant mean free paths are on the order of 1–100 nm. Consequently, the notion of quantum mechanical temperature is difficult to define in nanostructures, thus limiting the applicability of Fourier's law of heat conduction and the diffusion equation because both rely on the definition of local temperature. Moreover, the diffusion equation treats electrons and phonons as classical particles without associated waves. As a consequence, when a structure's dimensions become comparable to the mean free path and wavelength of heat carriers, classical laws are no longer valid and new approaches must be taken to predict heat transfer at the nanoscale.

Research in heat conduction and phonon transport in MEMS and NEMS is a growing research field. Experimental results have shown, for example, that phonon surface and interface scattering can lower thermal conductivity of silicon thin films and nanowires in the sub-100-nm range by two orders of magnitude.[25] So whereas for bulk material interface effects only dominate at low temperatures, where the phonon mean free path is large, in nanometer-scale thin films, the room temperature thermal conductivity can be significantly modified.

It has also been shown that electrons in low-dimensional semiconductors such as quantum wells and wires have an improved thermoelectric power

factor that may lead to more efficient Peltier coolers. Applications of nanoheat transport will be covered in Volume III, Chapter 8 on actuators.

Quantized Thermal Conductance Thermal conductance, like electrical conductance, becomes quantized at small length scales. The specific electronic heat for a 3D case introduced above as $C_{v,el} = \frac{\pi^2}{2}\left(\frac{k_B T}{E_F}\right) k n_{3D}$ (Equation 3.307) must be modified for different dimensionalities because U depends on the density of states $G(E)_{\#D}$ and $C_{v,el} = \frac{dU}{dT}$. The thermal electronic conductivity expression, based on Equation 3.59, with the Fermi velocity v_{dx} replacing v_d, results in:

$$\kappa_e = \frac{1}{3} v_F C_{v,el} \lambda_e \tag{3.342}$$

The latter expression has been further adapted with an effective mean free path λ_e to reflect the various possible electron scattering mechanisms of electrons in nanodevices [defect scattering, phonon scattering, and boundary scattering (grain boundary or device boundary)]. This effective mean free path λ_e and the scattering constant τ_e that goes with it are calculated according to the so-called Matthiessen rule as:

$$\begin{aligned} \frac{1}{\tau_e} &= \frac{1}{\tau_{defect}} + \frac{1}{\tau_{boundary}} + \frac{1}{\tau_{phonon}} \\ \frac{1}{\lambda_e} &= \frac{1}{\lambda_{defect}} + \frac{1}{\lambda_{boundary}} + \frac{1}{\lambda_{phonon}} \end{aligned} \tag{3.343}$$

In Figure 3.160 we show a plot of λ_e as a function of temperature. At lower temperatures defect scattering dominates, whereas at higher temperatures phonon scattering takes over. From Figure 3.160 one can also surmise that for nanodevices the device boundary ($\lambda_{boundary}$) might start taking over from defect control. For nanodevices and thin films the plateau at low temperature is decreased compared with a bulk solid. In Figure 3.161 the thermal conductivity κ_e as a function of temperature for copper and aluminum is shown; for metals, electrons dominate the thermal conductivity. Scattering of heat flow at the interfaces significantly reduces thermal conductivity. As expected from Equation 3.308, we observe a linear dependence of electronic specific heat with temperature until phonon scattering sets in.

To adapt Equation 3.59 for phonon thermal conductance κ_l, we introduce the effective mean free path λ_l for phonons and replace v_{dx} with the sound velocity in the solid at hand and use $C_{v,lat}$ for the lattice contribution to the total heat capacity C_v:

$$\kappa_l = \frac{1}{3} v_s C_{v,lat} \lambda_l \tag{3.344}$$

The effective mean free path λ_l now reflects the various possible phonon scattering mechanisms and is again calculated using Matthiessen's rule [defect and dislocation scattering, phonon-phonon scattering, and boundary scattering (grain boundary or device boundary)]. This is illustrated in Figure 3.162, where we plot λ_l as a function of T/θ_D.

Device boundary scattering effects dominate at low temperatures but are taken over by defect scattering, and finally phonon-phonon scattering causes a fast decay as T approaches θ_D.

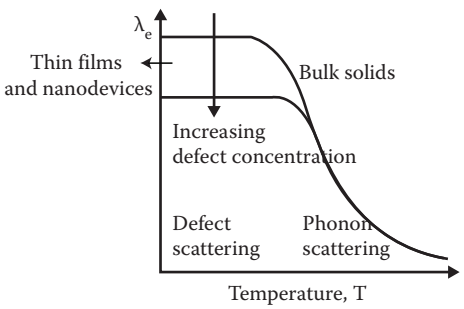


FIGURE 3.160 A plot of λ_e as a function of temperature. At lower temperatures defect scattering dominates, whereas at higher temperatures phonon scattering takes over.

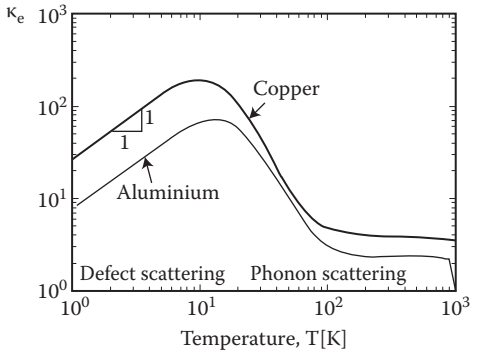


FIGURE 3.161 The thermal conductivity κ_e as a function of temperature for copper and aluminum shows a linear dependence with temperature until phonon scattering sets in.

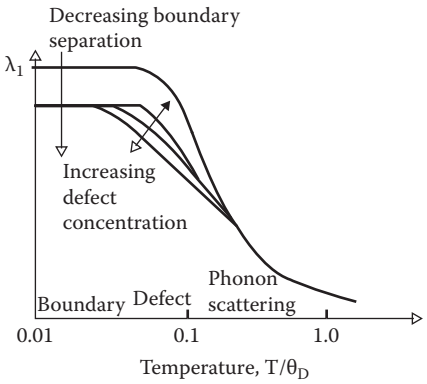


FIGURE 3.162 Plot of phonon effective mean free path λ_l as a function of T/θ_D.

Calculating the phonon contribution to the specific heat of a 3D insulating structure, we found earlier a constant value at $T \gg \theta_D$, and for $T \ll \theta_D$ (where we replaced C_v with $C_{v,lat}$ and $C_{v,el} = 0$) we calculated: $C_v = \frac{12}{5}\pi^4 R\left(\frac{T}{\theta_D}\right)^3$ (Equation 3.331), with a T^3 temperature dependency for the specific heat and a T^4 dependency for the internal energy U. Generalizing for different dimensions D we introduced earlier, $U \propto T^{D+1}$ and $C_{v,lat} \propto T^D$ (Equation 3.338), with D = 1, 2, 3 for the dimensions of the sample. Substituting this expression in Equation 3.344, we obtain for the thermal conductivity κ_l as a function of T/θ_D the relationship shown in Figure 3.163. At $T \gg \theta_D$, $C_{v,lat}$ is a constant, and at $T \ll \theta_D$, $C_{v,lat} \propto T_D$, whereas scattering at low temperatures is static (λ_{static}: boundaries and defects) and has a constant value ($\lambda_{static} = C$), and at high temperatures dynamic Umklapp or phonon-phonon scattering takes over ($\lambda_{Umklapp} = e_{\theta/T}$); the combined effect is reflected in Figure 3.163. The peak in the curve occurs when $\lambda_{static} = \lambda_{Umklapp}$.

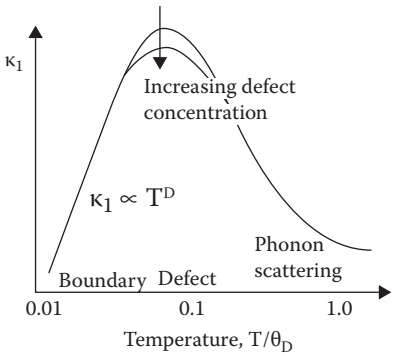


FIGURE 3.163 Phonon thermal conductivity κ_l as a function of T/θ_D.

Comparing Electrical and Thermal Conductance Quantization Quantized electrical conductance in a quantum wire (1D) was dramatically illustrated in Figure 3.138, where we illustrated experiments involving quantum wires in a $GaAs/Al_xGa_{1-x}As$ heterostructure at very low temperatures (close to 0 K). The quantum of electrical conductance we found is $2e^2/h$. In electrical experiments, the chemical potential (μ) and temperature T can be varied independently, so that at low temperatures the sharp edge of the Fermi-Dirac distribution can be swept through 1D modes. With phonons, only the temperature can be swept, and the broader Bose-Einstein distribution function smears out all features except the lowest lying modes at the lowest temperatures.

The quantum thermal conductance is proportional to T because energy is being transported as opposed to quantum electrical conductance, where charge is being transported. In ballistic thermal transport in a 1D structure, in the limit of $kT \ll \hbar\omega$, only the four lowest phonon modes contribute to the thermal conductance. For ideal coupling of phonon waveguide modes [transmission coefficient $T_m(\omega) = 1$] to a thermal reservoir, the following fundamental relation holds for each mode:

$$G_{th} = g_0 = \frac{\pi^2 k_B^2 T}{3h} \qquad (3.345)$$

This quantum represents the maximum possible value of energy transported per phonon mode. This fundamental quantum limit of thermal conductance has been experimentally confirmed in experiments illustrated in Figure 3.164.[26] In Figure 3.164a, a log-log of $G_{th}/16\ g_0$ versus temperature in mK is presented, and as observed, at the lowest temperatures the curve saturates at a value of 1. For this measurement Schwab et al.[26] used a suspended silicon nitride island connected to the rest of the device by a set of four silicon nitride wires (60 nm thick and less than 200 nm wide). The "island" is suspended over a phonon cavity of 4 × 4 μm (Figure 3.164b). The four wires connect to two thin-film heating and sensing Cr/Au resistors (the two bright "C"-shaped objects in the center) on the suspended Si_3N_4 island. The dark regions in Figure 3.164b are empty space, i.e., the cavity underneath the island and bridges.

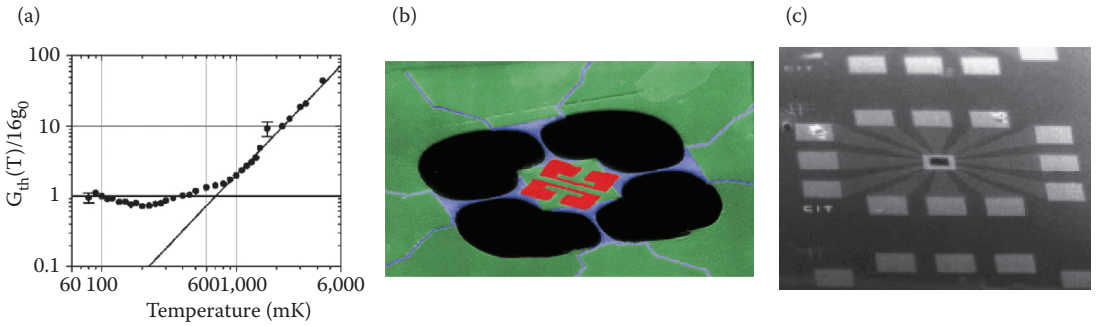


FIGURE 3.164 (a) Log-log of $G_{th}/16\ g_0$ versus temperature in mK. At the lowest temperatures the curve saturates at a value of 1. (b) Set of four 60-nm-thick and less than 200-nm-wide silicon nitride wires suspended over a phonon cavity of 4 × 4 μm. Two thin-film heating and sensing Cr/Au resistors are the bright "C"-shaped objects in the center on a suspended Si_3N_4 island. The dark regions are empty space. (c) The overall device (1.0 × 0.8 mm) has 12 wire-bond pads converging via thin-film niobium leads into the center of the device.

The overall device (1.0 × 0.8 mm) is shown in Figure 3.164c. Twelve wire-bond pads converge via thin-film niobium leads into the center of the device. These leads run atop the "phonon waveguides" to the Cr/Au resistors. Sensing is accomplished with a superconducting quantum interference device (SQUID) thermometry circuit. Because there are four modes per phonon waveguide and there are four of them, one would indeed expect that $G_{th}/16\ g_0 = 1$, as confirmed by the data in Figure 3.164a.

Summarizing quantized thermal and electronic transport:

1. Metal conductivity at room temperature is mostly independent of impurities or lattice defects. The electron concentration is determined by the nature of the metal and is comparable with the atom density. Mobility of electrons in this regime is dominated by lattice vibrations that are caused by sound or simply random thermal motion (phonons). Lattice vibrations (not atoms) scatter the electrons and determine their mean free path in an electric field, and thus for pure metals the resistivity $\rho \propto T$ (at high temperatures).
2. At very low temperatures and for a very pure metal one would expect the resistance to become zero because all lattice vibrations have died out. In reality, near zero the resistivity goes as T^5 (Bloch-Gruneisen law) because this represents the temperature dependence of electrons scattered by phonons with the number of electrons independent of temperature.
3. The film thickness and lattice defects, such as vacancies, grain boundaries, interstitials, dislocations, and impurities, ensure that the resistance remains nonzero even at the lowest temperatures. The combined effect of thermal vibrations, impurities, and defects on the resistivity is given by Matthiessen's rule, where $\rho_{total} = \rho_0 + \rho(T)$ (see expanded Equation 3.23). In this expression ρ_0 is a constant that increases with impurity content, and $\rho(T)$ is the phonon contribution to the resistivity. For pure elements the contribution of defects is on the order of 0.1% of the total, but for heavily cold worked metals it can be as high as 5%.
4. At high temperature the resistivity is limited by lattice thermal motion and at low temperature by lattice defects. The mobility in Equation 3.9 may be extracted from Hall measurements (see Equation 3.65), and when measured at low temperatures it serves as a measure of a material's purity. A typical mobility in single-crystal Si at 300 K is about 1000 cm^2/Vs. In the case of superconductivity (see Chapter 8, Volume III under "Superconductivity") the resistivity disappears above 0 K and below the critical superconducting phase change temperature T_c.

Beyond Schrödinger's Equation

The Other Models

In the quantum mechanics sections above, we relied exclusively on Schrödinger's equation. But by 1926 Schrödinger himself came to accept the serious

limitations of his own theory, particularly its lack of particulate behavior. In 1925 Werner Heisenberg came up with a new approach. Instead of treating the atom as a miniature planetary system, he saw it as a virtual oscillator (remember Planck?), and he set out to find a description that connected the Bohr-Sommerfeld quantum states with the frequencies and intensities of their bright line spectra. He characterized the electron in terms of its linear momentum, **p**, and its displacement from equilibrium, **q**, and attempted to calculate its energy in a quantum state n. Within a few months he had succeeded—with some surprising consequences. In classical theory, the simple act of multiplication is commutative—it does not matter which way round the multiplication is done, so: $2 \times 3 = 6$ and $3 \times 2 = 6$. But in Heisenberg's mathematics, multiplication did not follow this rule: $\mathbf{q} \times \mathbf{p}$ gave a different answer than $\mathbf{p} \times \mathbf{q}$. It was soon appreciated that this noncommutability was not really surprising: Heisenberg's new formalism turned out to be a form of matrix algebra developed almost a century before by William Hamilton. In matrix algebra the following matrices:

$$\begin{vmatrix} 3 & 5 & 4 \\ 1 & 1 & 1 \\ 2 & 3 & 5 \end{vmatrix} \times \begin{vmatrix} 1 & 3 & 5 \\ 2 & 5 & 1 \\ 4 & 3 & 2 \end{vmatrix} = \begin{vmatrix} 29 & 46 & 28 \\ 7 & 11 & 8 \\ 28 & 36 & 23 \end{vmatrix}$$

$$\mathrm{p} \times \mathrm{q} \qquad (3.346)$$

and

$$\begin{vmatrix} 1 & 3 & 5 \\ 2 & 5 & 1 \\ 4 & 3 & 2 \end{vmatrix} \times \begin{vmatrix} 3 & 5 & 4 \\ 1 & 1 & 1 \\ 2 & 3 & 5 \end{vmatrix} = \begin{vmatrix} 16 & 23 & 32 \\ 13 & 18 & 18 \\ 19 & 29 & 29 \end{vmatrix}$$

$$\mathrm{q} \times \mathrm{p}$$

do not lead to the same result. Heisenberg's matrix mechanics worked perfectly if one assumed the noncommutability could be expressed as:

$$\mathbf{pq} - \mathbf{qp} = \frac{\hbar}{\mathrm{i}} \qquad (3.347)$$

where i is the square root of -1, an imaginary number. For solid bodies on a large scale, on the other hand, $\mathbf{pq} - \mathbf{qp} = 0$, as if we were dealing with ordinary numbers. But his theory was purely mathematical—there was no longer a physical or "common-sense" picture with which to visualize the atom. The vital noncommutability was a real problem: what did it mean physically? The closest that Heisenberg could come to the physical significance of noncommutability is that the order in which measurements are made during an experiment is critical to the answer obtained. He showed that a consequence of the noncommutability of pairs of quantities such as momentum, **p**, and position, **q**, of a particle is that **p** cannot be known precisely if **q** is known, and **q** cannot be known precisely if **p** is known. Moreover, if the uncertainty in **p** is $\Delta\mathbf{p}$ and the uncertainty in **q** is $\Delta\mathbf{q}$, then $\Delta\mathbf{p} \times \Delta\mathbf{q}$ must always be equal to or greater than $h/2\pi$. This does not come from our inability to perform a careful experiment—it is absolutely fundamental to the quantum world. In other words, we obtain again the Heisenberg uncertainty principle.

In the meantime, Paul Dirac, working alone in Cambridge, had developed a third quantum theory. It is Dirac's theory that is the most comprehensive and far reaching. Dirac was able to demonstrate that both Schrödinger's and Heisenberg's theories, as well as classical physics, were special cases of his own quantum algebra or quantum mechanics. He was able to incorporate Maxwell's equations of electromagnetic radiation into his theory (the first example of a quantum field theory), as well as relativistic effects (i.e., close to the speed of light). In so doing he predicted *ab initio* that the electron should have a spin of 1/2. Dirac's theory removed the paradox of particle-wave duality: it showed that if a particle was probed in a way that was meant to demonstrate its particle-like properties, it would appear to be a particle, and if it was probed in a way that was meant to demonstrate its wave-like properties, it would appear to be a wave. It seems that it is our own inability to conjure up an appropriate or adequate mental picture of photons, atoms, electrons, and other quantum particles that is at the heart of the particle-wave duality paradox.

With his general relativity theory, Einstein reconciled Newton's laws with special relativity. But, as discussed in Chapter 5, the general relativity theory is inherently classical, and this bothered many physicists as it makes for a difficult marriage between quantum physics and general relativity. Physicists first tried to merge special relativity with quantum

concepts by focusing on how electromagnetic waves interact with matter, which led to quantum electrodynamics (QED). QED is the most precise theory of natural phenomena ever advanced. A major contributor to this fourth model of quantum mechanics was Richard Feynman, who in 1965 received the Nobel Prize for his work in QED. Feynman, one of the greatest theoretical physicists since Einstein, challenged the basic assumption of probability waves associated with electrons. He instead proposed that when an electron moves from here to there it explores all possible paths, and he showed he could assign a number to each of these paths in such a way that their combined average yields exactly the same result for the probability calculated by using the wave function approach. This is known as the "sum-over-paths" or the "sum-over-histories" approach to quantum mechanics. The mathematical formulation of this has an infinite amount of integrals. In practice a large number will suffice. An excellent resource to get better acquainted with the Feynman approach to quantum mechanics can be found in Taylor et al.[27] and in Feynman's own *QED—The Strange Theory of Light and Electrons*.[28]

By the 1970s there were quantum mechanical descriptions for three of the four known forces: strong, weak, and electromagnetic. Only gravity so far has eluded becoming part of a unified field description. String theory is one of most promising candidates for a unified theory of everything.

Antimatter

Dirac's theory alluded to in the preceding section also predicted the existence of antimatter. His equations have solutions that suggest that electron states with negative energy exist—but if this were the case, why did not all the electrons fall into these negative states? He reasoned that all of these states must already be full—what we consider to be empty space is a sea of negative energy electrons. If we give such an electron enough energy, it should break free of the sea and appear as a normal electron, leaving behind a "hole" in the negative energy sea. To get from $-mc^2$ (negative energy electron) to $+mc^2$ (normal electron), one needs $2mc^2$ (Figure 3.165).

Such a hole should appear as a positively charged electron, i.e., a positron, which will recombine with

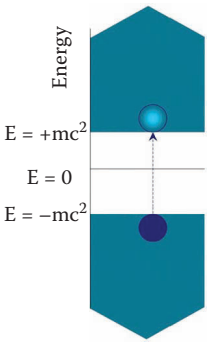


FIGURE 3.165 Positron formation out of a sea of negative energy electrons.

a normal electron and vanish in a burst of energy equal to $2mc^2$. Carl Anderson (1905–1991), an American physicist, discovered positrons in 1932 in cosmic ray showers and received the Nobel Prize in 1936 for his discovery (Figure 3.166). The discipline of particle physics was born, and soon antiparticles of most of the known real particles were found: the antiproton in 1955, the antineutron in 1960, the antideuteron in 1965, and even antihydrogen atoms were created at the FermiLab in 1995. To create an antiparticle all that is needed is sufficient energy—$2mc^2$—to release a real particle from the negative energy sea, leaving the equivalent "hole"—the antiparticle—behind. A particle can annihilate with its antiparticle to form gamma rays. Conversely, gamma rays with sufficiently high energy can turn into a particle-antiparticle pair.

Quantum Reality—Schrödinger's Cat

To understand how dramatic a change of view of reality quantum physics implies, it is good to consider Schrödinger's legendary cat story. A cat is placed in a box, together with a radioactive atom. If the radioactive atom decays and the Geiger counter detects an alpha particle, the hammer hits a flask of prussic acid (HCN), killing the cat. The paradox

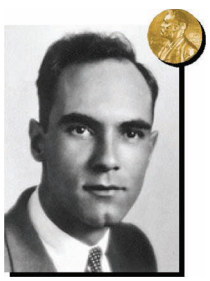


FIGURE 3.166 Carl Anderson (1905–1991).

lies in the clever coupling of quantum and classical domains. Before the observer opens the box, the cat's fate is tied to the wave function of the atom, which is itself a superposition of decayed and undecayed states. Thus, said Schrödinger, the cat must itself be in a superposition of dead and alive states before the observer opens the box, "observes" the cat, and "collapses" its wave function. Schrödinger devised the cat experiment to illustrate just how radically the quantum realm differs from the macroscopic, everyday world that we inhabit (Figure 3.167).

He himself had shown that a particle such as an electron exists in a number of possible states, the probability of each of which is incorporated into an equation known as the wave function. In the case of an atom of radioactive material, for example, the atom has a certain probability of decaying over a given period. Based on our "classical" intuition, we would assume that there are only two possibilities: either the atom has decayed, or it has not. According to quantum physics, however, the atom inhabits both states simultaneously. It is only when an observer tries to determine the state of the atom by measuring it that the wave function "collapses," and the atom assumes just one of its possible states: decayed or undecayed.

During the 1980s, the late theorist John Bell suggested a more palatable version of Schrödinger's experiment, one in which the decay of the atom causes a bottle of milk to spill onto the floor; thus, the superposed cat is hungry or full rather than alive or dead. But either version seems weirdly nonsensical: the outcome seems logical from a quantum physics viewpoint, but common sense tells us that a cat cannot be alive and dead (or hungry and full) at the same time.

The paradox of Schrödinger's cat has provoked a great deal of debate among theoretical physicists and philosophers. Although some thinkers have argued that the cat does exist in two superposed states, most contend that superposition only occurs when a quantum system is isolated from the rest of its environment. Various explanations have been advanced to account for this paradox—including the idea that the cat, or simply the animal's physical environment, can act as an observer. The question is, at what point, or scale, do the probabilistic rules of the quantum realm give way to the deterministic laws that govern the macroscopic world? This question has been brought into vivid relief by the work done by a National Institute of Standards and Technology (NIST) team in 1996, which included Christopher Monroe, Dawn Meekhof, Brian King, and Dave Wineland.[29] This group confined a charged beryllium atom in a tiny electromagnetic cage and then cooled it with a laser to its lowest energy state. In this state, the position of the atom and its "spin" (a quantum property that is only metaphorically analogous to spin in the ordinary sense) could be ascertained to within a very high degree of accuracy, limited by Heisenberg's uncertainty principle. This team of physicists managed to create a "Schrödinger cat-like state of matter" in a single atom. By supercooling a beryllium atom with a laser, then prodding it with a rapid sequence of laser pulses, the physicists managed to get the atom to oscillate in such a way that it exists "in the bizarre state of being in two well-separated positions at once."

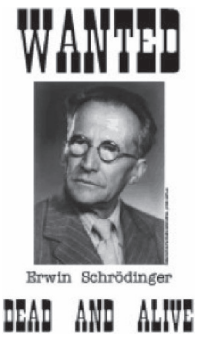


FIGURE 3.167 Dead and alive.

Mathematically, the property of superposition of states is directly related to the special form of linear differential equation, which governs the behavior of a single particle in quantum theory. When this particle interacts with other particles (as in a macroscopic object) or the detecting device or its environment, the equation becomes nonlinear, implying no more superposition and hence the "collapse" of the particle to a definite state.

Bell's Theorem

Bell's theorem says that reality must be nonlocal; the properties of the measured atom are influenced by events in whose details large portions of the universe participate instantly. In local reality, influences cannot travel faster than light. Bell's theorem says that in any reality of this sort, information does

not get around fast enough to explain the quantum facts: reality must be nonlocal. The theorem proves that any model of reality, whether ordinary or contextual, must be connected by influences that do not respect the optical speed limit.

In the summer of 1982, Alain Aspect and his team at the University of Paris-South carried out an experiment to establish the reality below the unreal world of the quantum, and their results dealt a fatal blow to Einstein's "hidden variables" concept. The hidden variable idea was espoused by some physicists like Einstein, who argued that the statistical nature of quantum mechanics suggests that quantum mechanics is "incomplete." There must be some other variables at play. This group of scientists maintained that physical theories must be deterministic to be complete. Bell's theorem proves that local hidden variables are impossible. The experiment concerned the behavior of photons flying off in opposite directions from a source and a detector capable of measuring the polarization of these photons. Because the two photons are emitted together, their polarizations are correlated. According to quantum theory, this polarization does not exist until it is measured, but according to the neorealists, each photon has a "real" polarization from the moment it is created. The results demonstrated that the correlation between the photons is the kind predicted by quantum theory, and, again as predicted by quantum theory, the measurement that is made on one photon has an instantaneous effect on the nature of the other photon. Some interaction links the two photons, although these photons speed away at the speed of light, and relativity theory tells us that no signal can travel faster than light.

Acknowledgments

Special thanks to Nahui Kim (UCI), Arvind Ajoy (IIT Madras), G.S. Jayadeva (IIT Madras), and Dr. Jia Guanyao (UC Irvine).

Questions

3.1: What is the physical meaning of a wave function? In what limit and in what sense does a wave function approach the classical limit?

3.2: What are allowed and forbidden regions in classical mechanics? What is the general behavior of a wave function in quantum mechanical allowed and forbidden regions? What is a turning point?

3.3: Why is the conductivity much larger in quantum theory than in classical theory?

3.4: What is the physical meaning of distribution functions [Maxwell-Boltzmann (classical), Fermi-Dirac, and Bose-Einstein]? How does one calculate the density of states (DOS) in a quantum distribution? How does one relate the Fermi energy to the density of particles?

3.5: A 1-W laser ($\lambda = 530$ nm) shines on a nickel surface that has a reflectivity of 96%. (a) How many photons per second does the laser emit? (b) Do you expect to observe the photoelectric effect? If so, what is the maximum energy of the photoelectrons? (The work function of nickel is 1.9 eV, Planck constant $h = 6.6 \times 10^{-34}$ Js.)

3.6: What is the maximum wavelength of light that can excite hydrogen atoms in the ground state ($n = 1$)? What is the color of hydrogen gas? ($R_H = 1.097 \times 10^7 m^{-1}$)

3.7: (a) What are the de Broglie frequencies and wavelengths of an electron and a proton accelerated to 100 eV? (b) In order to confine an electron to a nucleus, its de Broglie wavelength must be less than 10^{-14} m. What is its kinetic energy in that case? (Planck constant $h = 6.6 \times 10^{-34}$ Js, electron mass 9.1×10^{-31} kg, proton mass 1.7×10^{-27} kg)

3.8: An electron with total energy 2 eV impinges on a finite square potential barrier of height 1 eV. What is the velocity of the electron away from and in the barrier region? (electron mass 9.1×10^{-31} kg)

3.9: A particle is described by the wave function

$$\psi(x) = A\cos\left(\frac{2\pi x}{L}\right) \text{ for } -L/4 \le x \le L/4$$

and $\psi(x) = 0$ otherwise

(a) Determine the normalization constant *A*.

(b) What is the probability that the particle will be found between $x = 0$ and $x = L/8$?

3.10: Consider a particle in an infinitely deep one-dimensional potential well, where V = 0 for $|x| < L$. The wave function is of the form $\psi = A\sin kx + B\cos kx$.
(a) Apply the boundary conditions on the wave function to deduce that $k = n\pi/(2L)$, n = 1,2,3,..., with A = 0 for odd *n*'s, B = 0 for even *n*'s.
(b) Consider the same problem with a shift of origin, so that the well now runs from x = 0 to x = 2L. What is the form of the wave function in this case?
(c) The probability density for finding the particle at a given value of *x* is $\propto |\psi|^2$. Use this fact to normalize the wave function; i.e., find the constant of proportionality *A*.
(d) The time-independent Schrödinger (TISE) equation may be written as $(p^2/2m + V)\psi = E\psi$, where $p = -i\hbar\partial/\partial x$ in 1D. Show that a particle in state *n* has an energy $E_n = \hbar^2\pi^2n^2/(8mL^2)$.

3.11: What is the de Broglie wavelength of a small car moving at highway speeds (60 mph ≈ 100 km/h)? Can the wave nature of the car be detected?

3.12: The wave function e^{-ikx} is the solution of the time-independent Schrödinger (TISE) equation for a free particle [V(x) = 0]. What is the momentum and energy of this particle?

3.13: Through how many volts does a proton have to be accelerated for if it is to reach a million miles per hour? ΔV = ? The charge of a proton = 1.6×10^{-19} C and its mass m = 1.67×10^{-27} kg; $v_i = 0$ m/s; $v_f = 1 \times 10^6$ mph = 4.46×10^5 m/s.

3.14: Do the electrons in atoms and molecules obey Newton's classical laws of motion?

3.15: Suppose we have a solution, $\psi(x)$, of the time-independent Schrödinger equation, with energy, *E*. How do we construct a solution of the time-dependent Schrödinger equation?

3.16: The wave function e^{-ikx} is the solution of the time-independent Schrödinger equation for a free particle (V(x) = 0). What is the momentum and energy of this particle and in which direction does it move?

3.17: What is the boundary condition on the wave function at a point where the potential makes a jump (a) but remains finite and (b) goes to infinity?

3.18: Why is the infinite potential well (particle in a box) an important problem in quantum mechanics?

3.19: What is degeneracy in quantum mechanics?

3.20: What is the degeneracy of the first excited states of the three-dimensional cubic infinite potential well?

3.21: Why is the problem of the harmonic oscillator important in quantum mechanics?

3.22: What are transmission and reflection coefficients of particles and how are they related?

3.23: What are the important steps in solving the Schrödinger equation for a particle penetrating a potential barrier?

3.24: What is tunnelling?

3.25: What is the constraint that leads to the quantization of the principal quantum number, *n*?

3.26: What is spin in quantum mechanics?

3.27: What is the Pauli exclusion principle?

3.28: *Photoelectric effect*: Electromagnetic radiation with a frequency of $1.25 \cdot 10^{15}$ Hz strikes a piece of pure silver. Calculate the maximum speed of the electrons that are emitted. The work function of silver is 4.7 eV. What will be the change in speed be if the intensity of radiation is doubled?
Thanks to Arvind Ajoy, Indian Institute of Technology Madras.

3.29: *Energy levels in a 1D quantum well*: Assume an infinite electronic 1D quantum well of width L_1 = 100 nm. Calculate the ground state energy and the energy gap between the first excited state and the ground state. Repeat the calculations for L_2 = 500 nm. Comment on the result.
Thanks to Arvind Ajoy, Indian Institute of Technology Madras.

3.30: *Degeneracy in a 2D quantum well:* Consider a symmetric 2D quantum well, such that $L_x = L_y$ = 100 nm. Calculate the energy levels of the first two states $\{n_x, n_y\}$. Compare the results with calculations of an asymmetric well,

$L_x = 100$ nm, $L_y = 110$ nm. Thanks to Arvind Ajoy, Indian Institute of Technology Madras.

3.31: *Ionization energy of dopants using the Bohr model*: Consider an *n*-type silicon crystal, doped with phosphorus atoms. Each dopant atom replaces a silicon atom from its lattice site. It hence donates one extra electron to the crystal lattice. Calculate the ionization energy of this extra electron, assuming the other 4 valence electrons, the inner electrons, and the nucleus of the *P* atom as a core of net charge +1. Assume the effective mass of electrons in silicon as $m_e^* = 0.2\ m_0$ and a relative permittivity $\varepsilon_r = 12$.
Thanks to Arvind Ajoy, Indian Institute of Technology Madras

3.32: (a) Plot current density versus voltage for a silver metal wire of length 0.01 m for a voltage varying from 0.1 to 1 V. The resistivity of silver is $\rho = 1.59\ \mu\Omega$-cm at T = 300 K.
(b) Repeat part (a) by assuming Drude's model to find the resistivity. Given a thermal velocity of the electrons of $v_{th} = 0\ 1.57 \cdot 10^5$ m/s, what is the percentage of error in current density at 1 V when compared with part (a)?
(c) Repeat part (a) for silicon assuming $\rho = 3\ \Omega$-m (the resistivity of the semiconductor varies between 0.1 and 60 3 Ω-m depending on the doping level). Comment on current densities of metal and semiconductor.
Thanks to G.S. Jayadeva, Indian Institute of Technology Madras

3.33: (a) Find the energy levels for an electron in a quantum dot made from CdSe with an effective mass $m_e^* = 0.13\ m_0$ and having a de Broglie wavelength $1.23 \cdot 10^{-10}$ m, for $n_1 = 1$, $n_2 = 1, 2$, and $n_3 = 1, 2, 3$. Assume $L_1 = L_2 = L_3 = 1 \cdot 10^{-10}$ m.
(b) Do quantum effects occur in this quantum dot?
(c) What is the kinetic energy of carriers in the quantum dot?
Thanks to G.S. Jayadeva, Indian Institute of Technology Madras

3.34: Consider a 1D bound particle. Show that:

$$\frac{d}{dt}\int_{-\infty}^{\infty} \psi^*(x,t)\psi(x,t)dx = 0$$

Thanks to G.S. Jayadeva, Indian Institute of Technology Madras

3.35: An electron is confined in the ground state in a 1D box of width 10^{-11} m. Its energy is 36 eV. Calculate:
(a) The energy of the electron in the first excited state and the associated wave number and plot the wave function.
(b) The average force on the walls of the box when the electron is in the ground state.
Thanks to G.S. Jayadeva, Indian Institute of Technology Madras

3.36: What is the low temperature limit of the Fermi-Dirac distribution?

3.37: What is the value of the Fermi-Dirac distribution at $E = E_F$?

3.38: How does the number of particle-hole pairs increase with temperature?

3.39: What physical picture was the Drude theory of electrical conduction based on and what correct and incorrect results did it lead to?

3.40: Obtain the plot for Fermi-Dirac function when $\mu = E_F = 0.9$ eV.
Thanks G.S. Jayadeva, Indian Institue of Technology Madras

3.41: What distinguishes fermions and bosons?

Further Reading

Andrade e Silva, J., and G. Lochak. 1969. *Quanta.* London: World University Library.

Ashcroft, N. W., and N. D. Mermin. 1976. *Solid state physics.* Philadelphia: Saunders College.

Atkins, P. W., and J. de Paula. 2001. *Atkins' physical chemistry,* 7th ed. Oxford, UK: Oxford University Press.

Gribbin, J. 1984. *In search of Schrodinger's cat*. New York: Bantam Books.

Griffiths, D. J. 2005. *Introduction to quantum mechanics,* 2nd ed. Upper Saddle River, NJ: Pearson Prentice Hall.

Han, M. Y. 1990. *The secret life of quanta.* Blue Ridge Summit, PA: TAB Books.

Hayward, D. O. 2002. *Quantum mechanics for chemists* (RSC Tutorial Chemistry Texts 14). Royal Society of Chemistry.

Hey, T., and P. Walters. 2003. *The new quantum universe.* New York: Cambridge University Press.

Kittel, C. 2005. *Introduction to solid state physics,* 8th ed. New York: John Wiley.

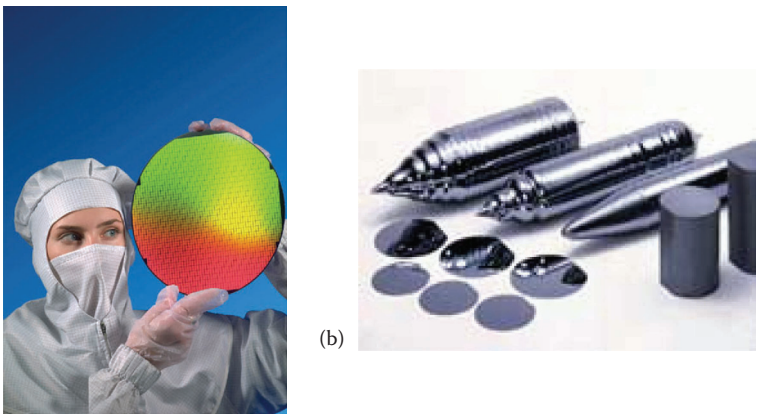
Marshall, I., and D. Zohar. 1997. *Who's afraid of Schrodinger's cat?* New York: Quill William Morrow.
McCarthy, W. 2003. *Hacking matter*. New York: Basic Books.
Moore, W. J. 1967. *Seven solid states.* New York: W. A. Benjamin, Inc.
Sakurai, J. J. 1985. *Modern quantum mechanics.* Redwood City, CA: Addison-Wesley Publishing Company, Inc.
Straus, H. L. 1968. *Quantum mechanics: an introduction*. Englewood Cliffs, NJ: Prentice-Hall.
Turton, R. 1995. *The quantum dot: a journey into the future of microelectronics.* 1995. New York: Oxford University Press.
Young, H. D., and R. A. Freeman, contributing author A. L. Ford. 2004. *Sears and Zemansky's university physics,* vol. 3, 11th ed. San Francisco: Pearson Addison Wesley.

References

1. Hall, E. H. 1879. On a new action of the magnet on electric currents. *Am J Math* 2:287–92.
2. Arndt, M., O. Nairz, J. Voss-Andreae, C. Keller, G. V. D. Zouw, and A. Zeilinger. 1999. Wave-particle duality of C60. *Nature* 401:680–82.
3. Herbert, N. 1987. *Quantum reality: beyond the new physics.* New York: Anchor Books, a Division of Random House, Inc.
4. Harris, R. A. 2006. *Influence of the shape and size of a quantum structure on its energy levels.* PhD diss., University of the Free State, Bloemfontein, South Africa.
5. Wagner, R. S., and W. C. Ells. 1964. Vapour-liquid-solid mechanism of single crystal growth. *Appl Phys Lett* 4:89–90.
6. Ekimov, A. I., A. L. Efros, and A. A. Onushchenko. 1985. Quantum size effect in semiconductor microcrystals. *Solid State Commun* 56:921–24.
7. Efros, A. L., and A. L. Efros. 1982. Interband absorption of light in a semiconductor sphere. *Sov Phys Semicond* 16:772–74.
8. Brus, L. E. 1984. On the development of bulk optical properties in small semiconducting crystallites. *J Luminescence* 31-32:381.
9. Brus, L. E. 1984. Electron-electron and electron-hole interactions in small semiconductor crystallites: the size dependence of the lowest excited electronic state. *J Chem Phys* 80:4403–07.
10. Murray, C. B., C. R. Kagan, and M. G. Bawendi. 1995. Self-organization of CdSe nanocrystallites into three-dimensional qantum dot superlattices. *Science* 270:1335–38.
11. Murray, C. B., D. J. Norris, and M. G. Bawendi. 1993. Synthesis and characterization of nearly monodisperse CdE (E=S, Se, Te) semiconductor nanocrystallites. *J Am Chem Soc* 115:8706–15.
12. Dabboussi, B. O., J. Rodriguez-Viejo, F. V. Mikulec, J. R. Heine, H. Mattoussi, R. Ober, K. F. Jensen, and M. G. Bawendi. 1997. (CdSe)ZnS core/shell quantum dots: synthesis and characterization of a size series of highly luminescent nanocrystallites. *J Phys Chem B* 101:9463–75.
13. Peng, X. G., M. C. Schlamp, A. V. Kadavanich, and A. P. Alivisatos. 1997. Epitaxial growth of highly luminescent CdSe/CdS core/shell nanocrystals with photostability and electronic accessibility. *J Am Chem Soc* 119:7019–29.
14. Hines, M. A., and P. Guyot-Sionnest. 1996. Synthesis of strongly luminescing ZnS-capped CdSe nanocrystals. *J Phys Chem B* 100:468–71.
15. Yu, W. W., E. Chang, R. Drezek, and V. L. Colvin. 2006. Mini review: water soluble quantum dots for biomedical applications. *ScienceDirect* 348:781–86.
16. Bruchez Jr., M., M. Moronne, P. Gin, S. Weiss, and A. P. Alivisatos. 1998. Semiconductor nanocrystals as fluorescent biological labels. *Science* 281:2013–16.
17. Gerion, D., F. Pinaud, S. C. Williams, W. J. Parak, D. Zanchet, S. Weiss, and A. P. Alivisatos. 2001. Synthesis and properties of biocompatible water-soluble silica-coated CdSe/ZnS semiconductor quantum dots. *J Phys Chem B* 105:8861–71.
18. Gur, L., N. A. Fromer, M. L. Geier, and A. P. Alivisatos. 2005. Air-stable all-inorganic nanocrystal solar cells processed from solution. *Science* 310:462–65.
19. O'Regan, B., and M. Gratzel. 1991. A low-cost, high efficiency solar cell based on dye-sensitized colloidal TiO2 films. *Nature* 353:737–40.
20. Wolf, E. L. 1985. *Principles of electron tunneling spectroscopy.* New York: Oxford University Press.
21. Esaki, L., and R. Tsu. 1970. *IBM J Res Develop* 14:61.
22. Dekker, A. J. 1962. *Solid state physics.* Englewood Cliffs, NJ: Prentice-Hall.
23. Pillai, S. O. 2005. *Solid state physics.* New Delhi: New Age International Publisher.
24. Dekker, C. 1999. Carbon nanotubes as molecular wires. *Physics Today* 52:22–28.
25. Li, D., Y. Wu, P. Kim, and A. Majundar. 2003. Thermal conductivity of individual silicon nanowires. *Appl Phys Lett* 83:2934–36.
26. Schwab, K., E. A. Henriksen, J. M. Worlock, and M. L. Roukes. 2000. Measurement of the quantum of thermal conductance. *Nature* 404:974–77.
27. Taylor, E. F., S. Vokos, J. O'Meara, and N. S. Thornber. 1997. Teaching Feynman's sum-over-paths quantum theory. *Comp Physics* 12:190–99.
28. Feynman, R. P. 1985. *The strange theory of light and matter*. Princeton, NJ: Princeton Science Library.
29. Monroe, C., D. M. Meekhof, B. E. King, and D. J. Wineland. 1996. A "Schroedinger Cat" superposition of an atom. *Science* 272:1131.

4

Silicon Single Crystal Is Still King



(a) (b)

(a) A 300-mm Si wafer at Infineon. (b) Silicon boules.

Outline

Introduction

Si Crystallography

Single-Crystal Structure and Conductivity

Single-Crystal Si Growth

Doping of Si

Oxidation of Silicon

Si-Based Electronic Devices

Physicochemical Properties of Si

Appendix 4A: Some Properties of Error Functions and Complementary Error Functions

Questions

Further Reading

References

Introduction

Given the importance of the orientation of silicon single crystal (SSC) for integrated circuit (IC), microelectromechanical system (MEMS), and nanoelectromechanical system (NEMS) applications, we analyze silicon crystallography and band structure in more detail in this chapter and compare them with the GaAs crystallography and band structure. This section will give the reader a better understanding of isotropic and anisotropic etchings of Si, covered in much more detail in Volume II, Chapter 4 on wet chemical etching and wet bulk micromachining—pools as tools. Subsequently, we review the growth of single-crystal Si, the doping of Si, and the oxidation of Si wafers. Although the emphasis in this book is on nonelectronic applications of miniaturized devices, such as sensing and actuating, we do briefly review active electronic devices, such as diodes [including light-emitting diodes (LEDs), photodiodes, solar cells, and Zener diodes] and two types of transistors [bipolar and metal oxide semiconductor field-effect transistor (MOSFET)]. We then analyze in some detail the scaling issues involved in the further miniaturization of Si MOSFETs. In this context, we introduce strained Si based on

silicon-germanium (SiGe) layers, which expands the prospect of continued Si use in the IC industry. In closing this chapter, we review the SCS properties that conspired to make Si so important in electronic, optical, and mechanical devices that one might rightly call the second half of the 20th century the Silicon Age.

Si Crystallography

Diamond-Cubic Structure

Crystalline silicon forms a covalently bonded diamond-cubic structure, which has the same atomic arrangement as carbon in the diamond form and belongs to the more general zinc-blende classification (in which the lattice positions are not all necessarily occupied by the same atom). Silicon, with its four covalent bonds, coordinates itself tetrahedrally, and these tetrahedrons make up the diamond-cubic structure (Figure 4.1). There are only four elements that come in diamond lattices: C, Si, Ge, and α-Sn. Diamond (carbon), silicon, and germanium all have a valence of four, and all have the same crystal structure. However, diamond is an insulator, whereas the others are semiconductors as a result of the energy bandgap difference. Diamond's bandgap is 5.5 eV, silicon's is 1.1 eV, and germanium's is 0.67 eV. The large bandgap of diamond makes it an insulator. The unit cell in a diamond lattice is cubic with atoms at each corner and in the middle of each face [i.e., face-centered cubic (FCC); see Chapter 2]. In the interior there are four additional atoms located along the cube diagonals exactly one-quarter of the way down the diagonal. This structure can also be described as two interpenetrating FCC lattices, one displaced (1/4, 1/4, 1/4)* times the lattice parameter *a* with respect to the other, as shown in Figure 4.2. The lattice parameter *a* for silicon is 5.4309 Å, and silicon's diamond-cubic lattice is surprisingly wide open, with a packing density of 34% compared with 74% for a regular FCC lattice. The {111} planes present the highest packing density, and the atoms are oriented such that three bonds are below the plane. In addition to the diamond-cubic structure, silicon

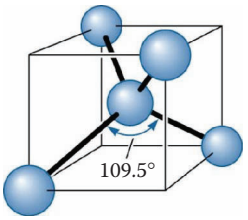


FIGURE 4.1 Covalent bonds are directional. In silicon, a tetrahedral structure is formed, with angles of 109.5° between each covalent bond.

is known to have several stable high-pressure crystalline phases and a stress-induced metastable phase with a wurtzite-like structure [wurtzite is a sulfide mineral of zinc and iron with the composition (Zn,Fe)S], referred to as *diamond-hexagonal* silicon. The latter has been observed after ion implantation and hot indentation.

GaAs has a zinc-blende lattice (zinc blende = ZnS; ZnTe and SiC also have that same structure), which is identical to the diamond lattice, except that the lattice sites are equally partitioned between two different atoms (Ga and As). The lattice parameter *a* for GaAs is 5.65 Å.

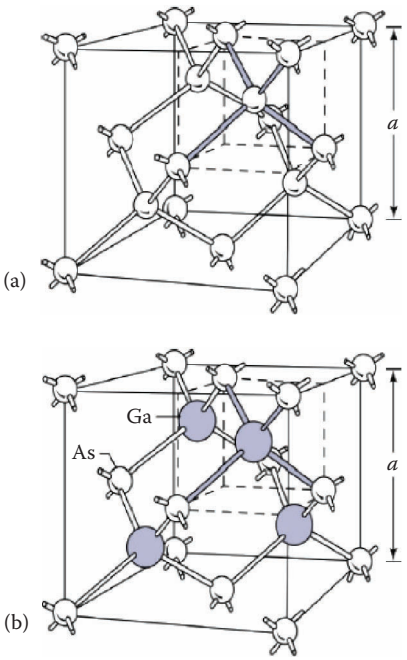


FIGURE 4.2 The "zinc-blende" lattice is a face-centered cubic (FCC) with two atoms in the base at (0, 0, 0) and (1/4, 1/4, 1/4). (a) The diamond-type lattice of Si can be constructed from two interpenetrating face-centered cubic unit cells. (b) GaAs crystal, which is identical to the diamond lattice, except that the lattice sites are equally partitioned between two different atoms (Ga and As).

* For a refresher on Miller indices to mark crystal planes with (), orientations with [], families of planes with {}, and families of orientations with < >, consult Chapter 2 on crystallography.

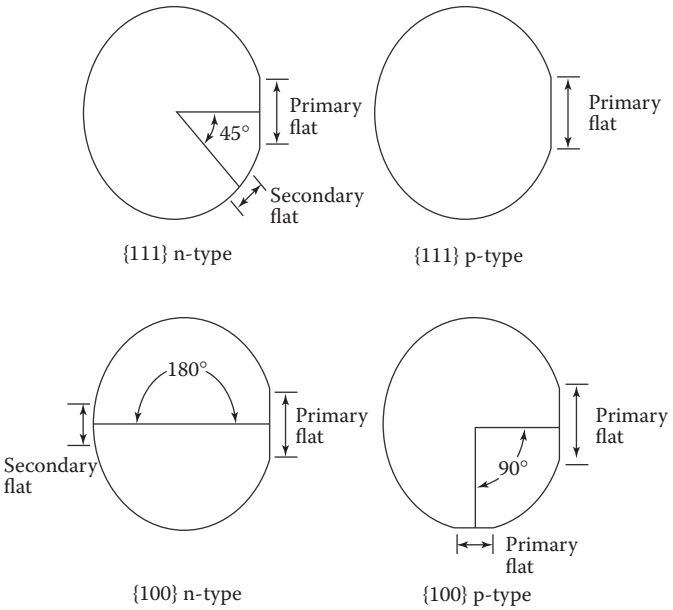


FIGURE 4.3 Primary and secondary wafer flats on silicon wafers.

When ordering silicon wafers, the crystal orientation must be specified. In IC work, [100]-oriented single-crystal Si wafers for metal oxide semiconductor (MOS) transistors and [111] for bipolar transistors are typical, but in MEMS and NEMS, other crystal orientations are used, especially [110]. The [110] wafers break much more cleanly than in other orientations. In fact, it is the only major orientation that can be cleaved with exactly perpendicular edges. The [111] wafers are used less often, as they cannot easily be etched by wet anisotropic etchants except when using special techniques such as laser-assisted etching (see Volume II, Chapter 4).[1] On a [100] wafer, the [110] direction is often made evident by a flat segment, also called an *orientation flat*. The precision on the orientation of the flat is within 3°. The flat's position on [110]-oriented wafers varies from manufacturer to manufacturer but often parallels a [111] direction. Flat areas help determine orientation, placement of slices in cassettes and in fabrication equipment (large primary flat), and help identify orientation and conductivity type (smaller secondary flat) (see also "Single-Crystal Silicon Growth"). Primary and secondary flats on (111) and (100) silicon wafers are indicated in Figure 4.3.

Crystalline silicon substrates are available as circular wafers of 100-mm (4-in) diameter and 525-µm thickness or 150-mm (6-in) diameter with a thickness of 650 µm. Larger 200-mm and 300-mm diameter wafers are currently not economically justified for MEMS or NEMS and are strictly used in the IC industry. Wafers polished on both sides, often used in MEMS, are about 100 µm thinner than substrates of standard thickness.

In Figure 4.4 we show a typical MEMS structure, a piezoresistive pressure sensor, featuring a thin suspended Si membrane, fashioned from bulk single-crystal Si in a wet anisotropic bulk micromachining etch process. Anisotropic wet etching enables the selective thinning of a silicon wafer from a starting thickness of about 425 µm to a typical final thickness of 20 µm or less for the suspended diaphragm. This machining process, which we detail in

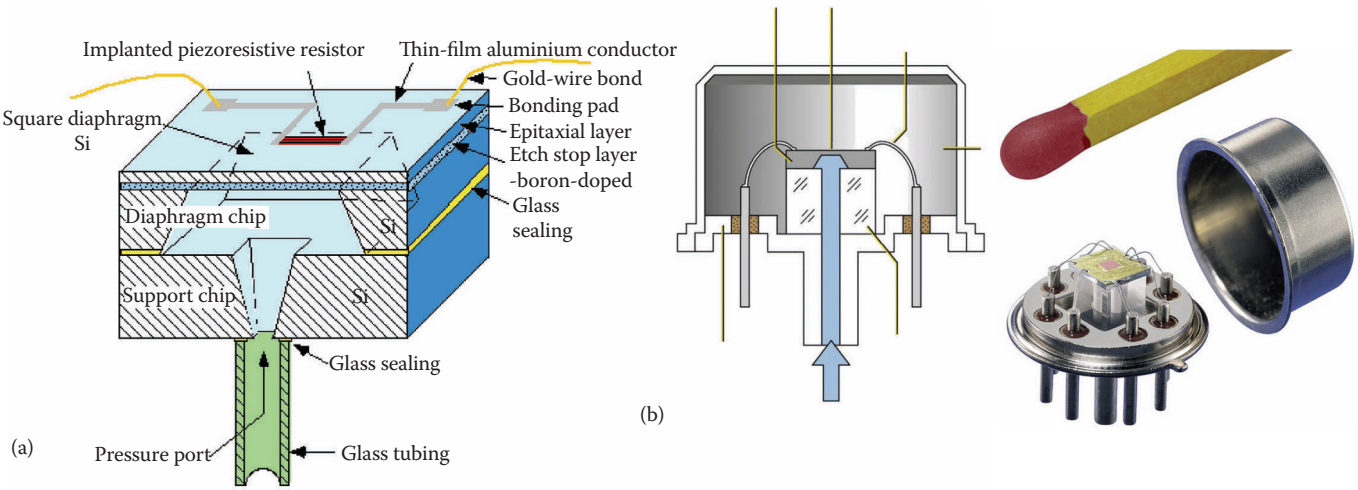


FIGURE 4.4 Piezoresistive pressure sensor. (a) Wet bulk micromachining process was used to craft this thin suspended membrane with implanted piezoresistive elements. (b) Bosch piezoresistive pressure sensor.

Volume II, Chapter 4, features very precise lateral dimensions and tight membrane thickness controls (e.g., in this case, an epitaxial Si layer that comes with a highly boron-doped etch stop layer). The thin silicon membrane suspended in the thick Si frame deforms with pressure, and the piezoresistors implanted in the membrane change resistance with the strain induced by the bending of the membrane. It was this type of Si MEMS structure that demonstrated that batch fabrication of miniature Si components did not need to be limited to integrated circuitry.

Some Important Crystallographic Planes in the Silicon Lattice

Introduction

To appreciate the different three-dimensional (3D) shapes resulting from anisotropically etched single-crystal Si (SCS), as shown in the MEMS pressure sensor in Figure 4.4, some of the more important geometric relationships between different planes within the Si lattice need further clarification. We consider only silicon wafers with a (100) plane or a (110) plane as the top surface. We will accept, for the moment, that in anisotropic alkaline etchants, the {111} family of planes has the highest atom-packing density and is nonetching compared with the other planes. Because the {111} planes are essentially not attacked by the etchant, the sidewalls of an etched pit in SCS will ultimately be bounded by this type of plane, given that the etch time is long enough for features bounded by other planes to be etched away. The types of planes introduced initially depend on the geometry and the orientation of the mask features.

In the sections below, we show how simple vector algebra (using Equation 2.14) enables one to calculate the angles between different planes or directions and thus makes it possible to predict the shapes resulting from anisotropic wet etching of single-crystal Si.

The reasons for this etching anisotropy will be revealed in Volume II, Chapter 4.

[100]-Oriented Silicon

The direction of a set of planes, we saw in Chapter 2, is the normal to that set of planes, so Si wafers with a (100) plane as the bounding surface are [100]-oriented wafers. In Figure 4.5, the unit cell of a silicon lattice is shown together with the correct orientation of a [100]-type wafer relative to this cell.[2] It can be seen that intersections of the nonetching {111} family of planes with the {100} family of planes (e.g., the wafer surface) are mutually perpendicular and lying along the <110> family of directions. Provided that a mask opening (perhaps a rectangle or a square) is accurately aligned with the primary orientation flat (the [110] direction), only {111} planes will be introduced as sidewalls from the very beginning of the

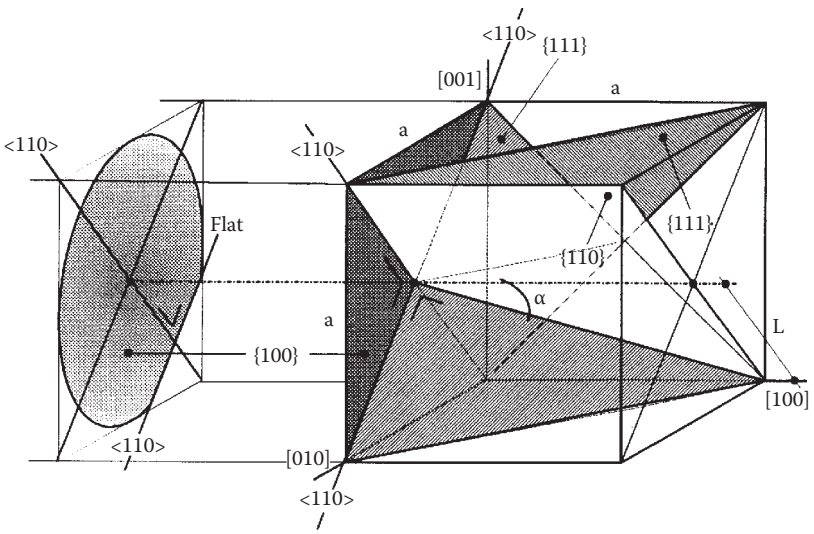


FIGURE 4.5 A (100) silicon wafer with reference to the unity cube and its relevant planes. (From Peeters, E. 1994. *Process development for 3D silicon microstructures, with application to mechanical sensor design.* PhD diss., Katholieke Universiteit Leuven, Leuven, Belgium. With permission.[2])

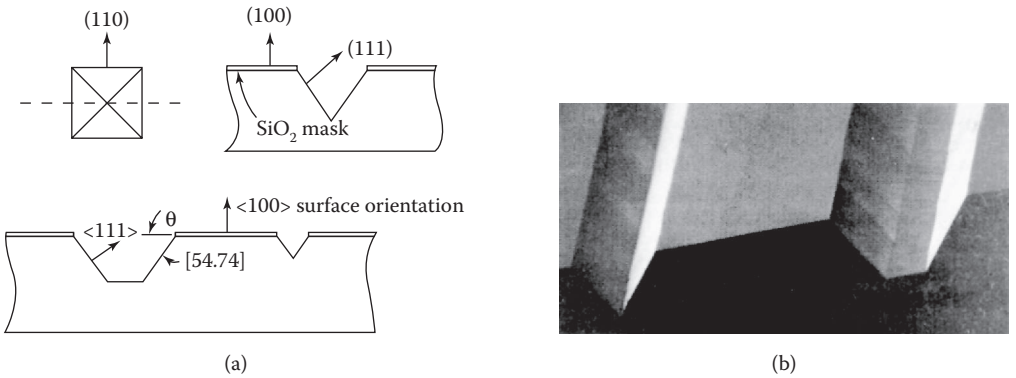


FIGURE 4.6 Anisotropically etched features in a (100) wafer with square mask (schematic) (a) and rectangular mask (scanning electron microscope micrograph of resulting V- and U-grooves) (b).

etch. Because the nonetching character of the {111} planes renders an exceptional degree of predictability to the recess features, this mask arrangement is most often used in commercial applications.

During etching, truncated pyramids (square mask) or truncated V-grooves (rectangular mask) deepen but do not widen (Figure 4.6). The edges in these structures are <110> directions, the ribs are <211> directions, the sidewalls are {111} planes, and the bottom is a (100) plane parallel with the wafer surface. After prolonged etching, the {111} family of planes is exposed down to their common intersection, and the (100) bottom plane disappears, creating a pyramidal pit (square mask) or a V-groove (rectangular mask) (Figure 4.6). As shown in Figure 4.6, no underetching of the etch mask is observed because of the perfect alignment of the concave oxide mask opening with the [110] direction. Misalignment still results in pyramidal pits, but the mask will be undercut. For a mask opening with arbitrary geometry and orientation, a circle, for example, and for sufficiently long etch times, the anisotropically etched recess in a (100) wafer is pyramidal, with a base perfectly circumscribing the circular mask opening.[2] Convex corners (>180°) in a mask opening will always be completely undercut by the etchant after sufficiently long etch times. This can be disadvantageous (e.g., when attempting to create a mesa rather than a pit), or it can be advantageous for undercutting suspended cantilevers or bridges. In the section on corner compensation in Volume II, Chapter 4, the issue of undercutting will be addressed in more detail. In corner compensation, the convex corner undercutting is compensated by clever mask layout schemes.

The slope of the {111} sidewalls in a cross-section perpendicular to the wafer surface (100) and parallel to the wafer flat is marked by the angle $\theta_{(100),(111)}$ in Figures 4.6a and 4.7. This angle, $\theta_{(100),(111)}$, was calculated in Chapter 2, using Equation 2.14, as:

$$\begin{aligned}\theta_{(100),(111)} &= \cos^{-1}\frac{\left(h_1h_2+k_1k_2+l_1l_2\right)}{\sqrt{h_1^2+k_1^2+l_1^2}\sqrt{h_2^2+k_2^2+l_2^2}}\\ &= \cos^{-1}\frac{\left(1+0+0\right)}{\sqrt{1}\sqrt{3}}\\ &= 54.74^\circ\end{aligned} \qquad (2.14)$$

There are six equivalent (100) planes [the {100} family] and eight equivalent (111) planes [the {111} family] for a total of 48 combinations on how they can intersect, but these intersecting angles are only either one of two, and 54.74° is one of them. The same angle of 54.74°, derived from Equation 2.14, can also be calculated from the lattice parameter *a* and the length *L*, both marked in Figure 4.5. The off-normal angle α, i.e., the intersection of a (111) sidewall and a (110) cross-secting plane, is calculated as:

$$\tan\alpha = \frac{L}{a}$$
$$\text{with } L = a\frac{\sqrt{2}}{2} \text{ or } \alpha = \arctan\frac{\sqrt{2}}{2} = 35.26^\circ \qquad (4.1)$$
$$\text{and } 54.74^\circ \text{ for the complementary angle}$$

The tolerance on this angle is determined by the alignment accuracy of the wafer surface with respect to the (100) plane. Wafer manufacturers typically specify this misalignment to 1° (0.5° in the best cases).

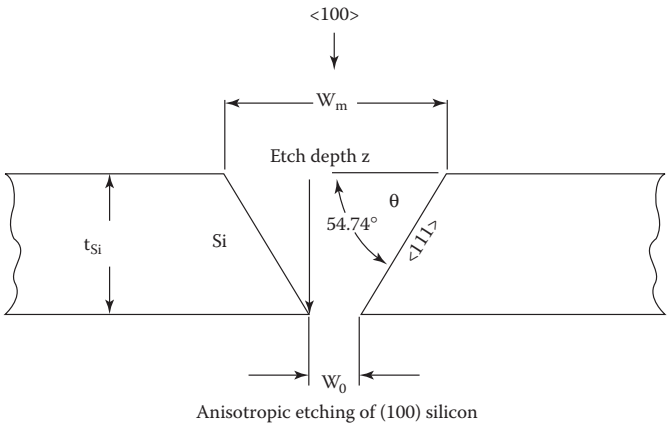


FIGURE 4.7 Relation of bottom cavity plane width W_0 with mask opening width W_m.

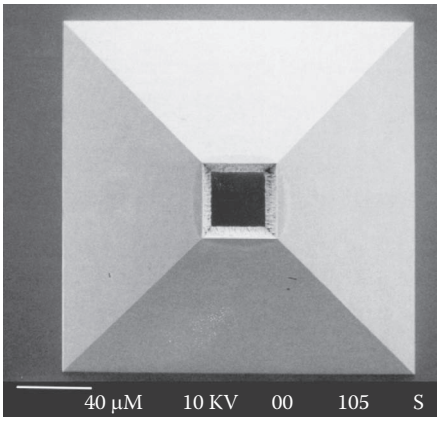


FIGURE 4.8 Orifice or via through a Si wafer made by anisotropic etching.

The 8 equivalent (111) planes [the {111} family] and 12 equivalent (110) planes [the {110} family] form 96 combinations, but they can intersect each other only at 35.26°, 90°, or 144.74°.

The width of the rectangular or square cavity bottom plane, W_0, in Figure 4.7, aligned with the <110> directions, is completely defined by the etch depth, z, the mask opening, W_m, and the above-calculated sidewall slope:

$$\begin{aligned} W_0 &= W_m - 2\,z\cot 54.74^\circ \\ &= W_m - \sqrt{2}\,z \end{aligned} \tag{4.2}$$

The larger the opening in the mask, the deeper the point at which the {111} sidewalls of the pit intersect. The etch stop at the {111} sidewalls' intersection occurs when the depth is about 0.7 times the mask opening. If the oxide opening is wide enough, $W_m > 849$ µm (for a typical 6 in wafer with thickness $t_{si} = z = 600$ µm), the {111} planes do not intersect within the wafer. The etched pit in this particular case extends all the way through the wafer, creating a small orifice or via (see Figure 4.8). If a high density of such vias through the Si is required, the wafer must be made very thin; otherwise, too much Si real estate is wasted.

Corners in an anisotropically etched recess are defined by the intersection of crystallographic planes, and the resulting corner radius is essentially zero. This implies that the size of a silicon diaphragm is very well defined, but it also introduces a considerable stress concentration factor. The influence of the zero corner radius on the yield load of diaphragms can be studied with finite element analysis (FEA).

One way to obtain vertical Si sidewalls instead of 54.7° sidewalls using a [100]-oriented Si wafer may be understood from Figure 4.9. It can be seen that there are {100} planes perpendicular to the wafer surface and that their intersections with the wafer surface are <100> directions. These <100> directions enclose a 45° angle with the wafer flat (i.e., the <110> directions). By aligning the mask opening with these <100> orientations, {100} facets are introduced as sidewalls. The {110} planes etch faster than the {100} planes and are not introduced. Because the bottom and sidewall planes are all from the same {100} family, lateral underetch equals the vertical etch rate, and rectangular channels, bounded by slower etching {100} planes, result (Figure 4.10).

Because the top of the etched channels are exposed to the etchant longer than the bottom, one might expect the channels in Figure 4.10 to be wider at the top than at the bottom. We use Peeters' derivation to show why the sidewalls remain vertical.[2] Assume the width of the mask opening to be W_m. At a given depth, z, into the wafer, the underlying Si is no longer masked by W_m, but rather by the intersection of the previously formed {100} facets with the bottom surface. The width of this new mask is larger than the lithography mask W_m by the amount the latter is being undercut. Let us call the new mask width W_z, the effective mask width at a depth z. The relation between W_m and W_z at any moment is given by the lateral etch rate of a {100} facet and the time that facet was exposed to the etchant at depth z:

$$W_z = W_m + 2R_{xy}\Delta t_z \tag{4.3}$$

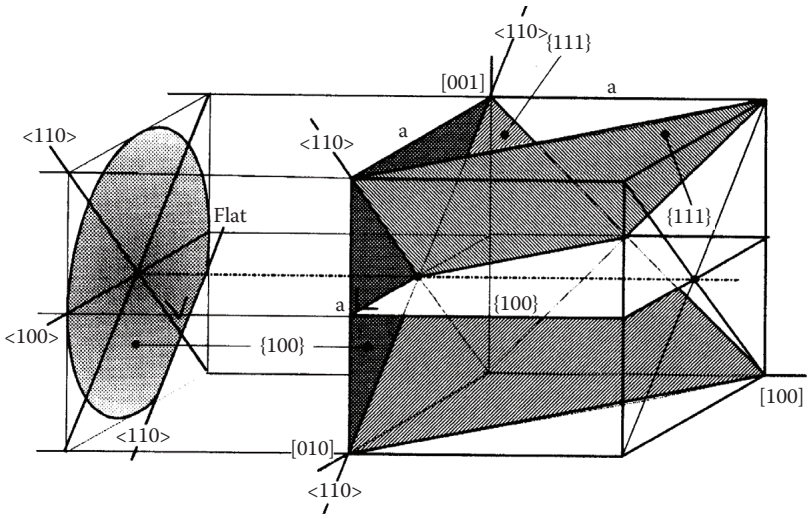


FIGURE 4.9 A (100) silicon wafer with [100] mask-aligned features introduces vertical sidewalls. (From Peeters, E. 1994. *Process development for 3D silicon microstructures, with application to mechanical sensor design.* PhD diss., Katholieke Universiteit Leuven, Leuven, Belgium. With permission.[2])

where R_{xy} = lateral underetch rate (i.e., the etch rate in the *x-y* plane)

Δt_z = etch time at depth z

The underetching, U_{xy}, of the effective mask opening W_z is given by:

$$U_{x,y} = TR_{x,y} - R_{x,y}\Delta t_z \tag{4.4}$$

where T is the total etch time so far. The width of the etched pit, W_{tot}, at depth z is further given by the sum of W_z and twice the underetching for that depth:

$$W_{tot} = W_z + 2U_{x,y} = W_m + 2TR_{x,y} \tag{4.5}$$

Or, because T can also be written as the measured total etch depth z divided by the vertical etch rate R_z, Equation 4.5 can be rewritten as:

$$W_{tot} = W_m + 2z\left(\frac{1}{R_z}\right)R_{x,y} \text{ or since } R_{xy} = R_z$$
$$W_{tot} = W_m + 2z \tag{4.6}$$

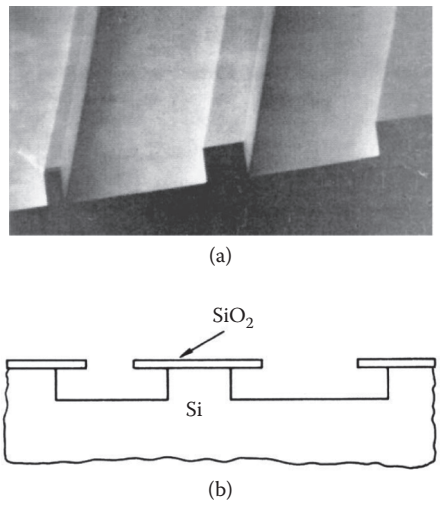


FIGURE 4.10 Vertical sidewalls in a (100) wafer obtained using anisotropic etching. (a) Anisotropically etched features in a (100) wafer with rectangular mask (scanning electron microscope micrograph). (b) Undercutting of the rectangular mask.

Therefore, the width of the etched recess is equal to the photolithographic mask width plus twice the etch depth. In other words, the walls remain vertical independent of the depth z because W_m does not depend on z.

For long etch times, {111} facets take over from the vertical {100} facets. These inward sloping {111} facets are first introduced at the corners of a rectangular mask and grow larger at the expense of the vertical sidewalls until the latter ultimately disappear altogether. Therefore, alignment of mask features with the <100> family of directions to obtain vertical sidewalls in [100]-oriented wafers is not very useful for the fabrication of diaphragms. However, it can be very effective for anticipating the undercutting of convex corners on [100] wafers. This

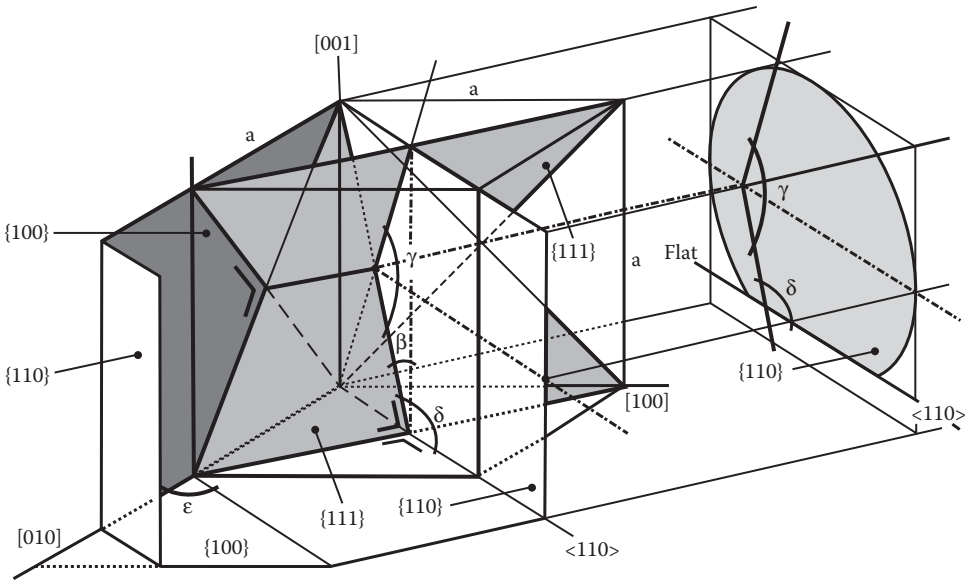


FIGURE 4.11 A (110) silicon wafer with reference to the unity cube and its relevant planes. The wafer flat is in the [110] direction. (From Peeters, E. 1994. *Process development for 3D silicon microstructures, with application to mechanical sensor design.* PhD diss., Katholieke Universiteit Leuven, Leuven, Belgium. With permission.[2])

useful aspect will be revisited when discussing corner compensation.

[110]-Oriented Silicon

Silicon wafers with a (110) plane as the bounding surface are [110]-oriented wafers (see Chapter 2). In Figure 4.11, we show a unit cell of Si properly aligned with the surface of a [110] Si wafer. This drawing will enable us to predict the shape of an anisotropically etched recess on the basis of elementary geometric crystallography. Earlier, we determined that there are 8 equivalent (111) planes [{111} family] and 12 equivalent (110) planes [{110} family] for 96 combinations that can intersect each other at only three different angles: 35.26°, 90°, and 144.74° (calculated using Equation 2.14). Four of the eight equivalent (111) planes are perpendicular to the (110) wafer surface. The remaining four are slanted at 35.26° or 144.74° with respect to the surface. Whereas the intersections of the four vertical (111) planes with the (100) wafer surface are mutually perpendicular, they enclose an angle δ or δ + γ with the wafer flat (the [110] direction). It follows that a mask opening that will not be undercut [i.e., oriented such that resulting feature sidewalls are exclusively made up by vertical and slanted {111} family planes] cannot be a rectangle aligned with the wafer flat, but must be a parallelogram with a large inside angle γ of 109.47° and a small angle of 180° – 109.46° = 70.5° (see Figure 4.12).

The angles δ and γ can also be calculated from the lattice parameter *a* as marked in Figure 4.11. We start by calculating the angle β (Figure 4.11)[2]:

$$\tan\beta = \frac{\frac{1}{2}a\frac{\sqrt{2}}{2}}{\frac{a}{2}} = \frac{\sqrt{2}}{2} \tag{4.7}$$

and from β we can deduce the other pertinent angles:

$$\gamma = 180^\circ - 2\beta = 180^\circ - 2\arctan\left(\frac{\sqrt{2}}{2}\right) = 109.47^\circ \tag{4.8}$$

$$\delta = 90^\circ - \beta = 90^\circ + \arctan\left(\frac{\sqrt{2}}{2}\right) = 125.26^\circ \tag{4.9}$$

$$\varphi = 270^\circ - \delta = 144.74^\circ \tag{4.10}$$

The large inside angle of the parallelogram γ is 109.47° (Equation 4.8), and the small angle is thus 70.53°. A groove etched in [110]-oriented wafers has the appearance of a complex polygon delineated by six (111) planes, four vertical and two slanted (Figure 4.13). The bottom of the etch pit shown in

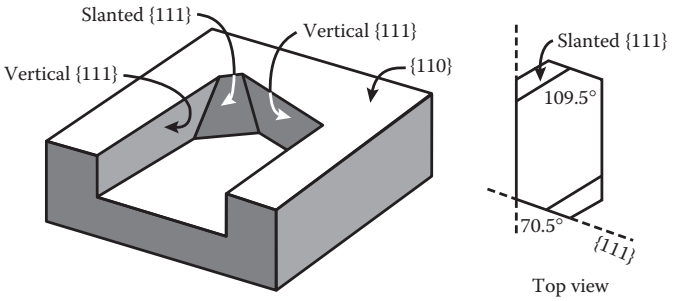


FIGURE 4.12 Illustration of anisotropic etching in [110]-oriented silicon. Etched structures are delineated by four vertical (111) planes and two slanted (111) planes from the {111} family. The vertical {111} planes intersect each other at angles of 109.5° and 70.5°.

Figure 4.13 is bounded at first by planes from the {110} and/or {100} families, depending on the etch time. At short etch times, one mainly sees a flat (110) bottom (Figure 4.12). As the {110} family of planes are etching slightly faster than the {100} family of planes, the flat (110) bottom gets progressively smaller, and a V-shaped bottom bounded by a {100} family of planes results. The angle ε as shown in Figure 4.13 equals 45°, being the angle enclosed by

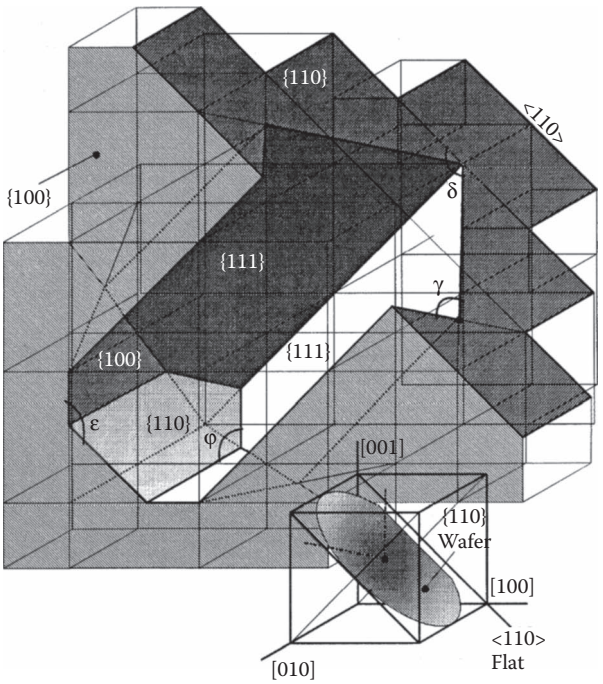


FIGURE 4.13 A (110) silicon wafer with anisotropically etched recess inscribed in the Si lattice. γ = 109.47°, δ = 125.26°, φ = 144.74°, and ε = 45°. (From Peeters, E. 1994. *Process development for 3D silicon microstructures, with application to mechanical sensor design.* PhD diss., Katholieke Universiteit Leuven, Leuven, Belgium. With permission.[2])

the intersections of {100} and {110} bottom planes. Even longer etch times result in shallow {111} planes forming at the bottom, and these eventually stop the etching process.

For [110] etching, an arbitrary window opening is circumscribed by a parallelogram with the previously derived orientation and skewness for sufficiently long etching times. Another difference between [100]- and [110]-oriented silicon wafers is that on the [110] wafers, it is possible to etch under microbridges crossing at a 90° angle a shallow V-groove [formed by (111) planes]. To undercut a bridge on a (100) plane, the bridge cannot be perpendicular to the V-groove; it must be oriented slightly off normal.[3]

Single-Crystal Structure and Conductivity

Intrinsic semiconductors are undoped semiconductors with electrical properties native to the material, whereas extrinsic semiconductors are doped semiconductors with electrical properties controlled by added impurities. At low temperature, pure silicon is an insulator because all of its valence electrons participate in bonding. Pure Si lacks electrons that can wander around freely in the crystal, and as a consequence, at absolute 0 K (T = 0 K), the conductivity (ρ) is very small. Thermal agitation at room temperature (kT ~ 0.025 eV) provides energy to electrons, and there is a finite probability that some of them will be able to break away to become conduction electrons. With intrinsic Si in thermal equilibrium, there is a balance between continuous bond breaking and bond formation. In this process, electrons and holes live fleeting existences with a recombination time on the order of microseconds.

The bandgap for a semiconductor in **k**-space (momentum space; see Chapters 2 and 3) is often represented in a simplified format as shown in Figure 4.14. In semiconductors and insulators, the Fermi level, E_F, falls within the forbidden energy gap.

The density of states (DOS) function $G(E)$ was introduced in Chapter 3 as the measure of volume that is available to hold charge carriers in a crystal. The number of states per energy interval is a complicated function, but the distribution of states near energy minima or maxima is proportional to $\sqrt{E}$.

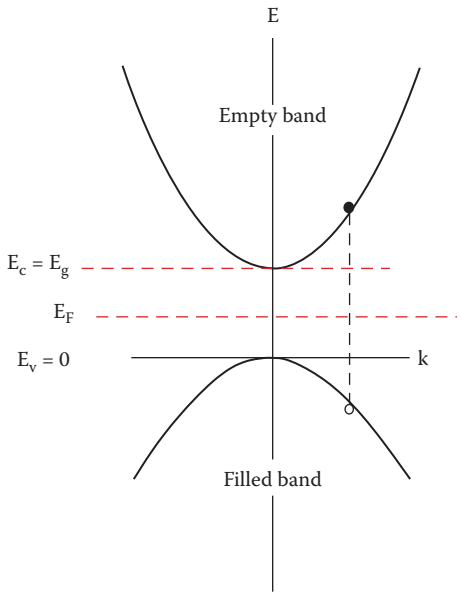


FIGURE 4.14 Energy band diagram in **k**-space for an intrinsic semiconductor. The open dot represents a hole in the valence band, and the filled dot represents an electron in the conduction band.

The DOS per unit volume for the parabolic bands in Figure 4.14 is given as:

$$G(E) = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}} \qquad (3.229)$$

With reference to Figure 4.14, where we have marked the bottom of the conduction band as E_c (= E_g) and the top of the valence band as E_v, we rewrite the DOS available for electrons at the bottom of the conduction band as:

$$G_n(E) = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} (E - E_c)^{\frac{1}{2}} \qquad (4.11)$$

with m_e^* representing the effective electron mass. The DOS available for holes at the top of the valence band is:

$$G_h(E) = \frac{1}{2\pi^2}\left(\frac{2m_h^*}{\hbar^2}\right)^{\frac{3}{2}} (E_v - E)^{\frac{1}{2}} \qquad (4.12)$$

with m_h^* representing the effective hole mass. To calculate the carrier density in the states available at the bottom of the conduction band and the top of the valance band (a large 3D model of semiconductor is assumed), we use:

$$n(E, T)dE = G(E)f(E)dE \qquad (3.230)$$

where we multiply the DOS with the Fermi-Dirac distribution function. This distribution takes into account that fermions (electrons, holes) cannot occupy the same quantum state. Figure 3.107 illustrates this kind of particle distribution, and we recognize that a sharp interface exists between filled and unfilled states; the function changes rapidly in a small region of a few kT around E_F. Assuming $E - E_F > 3kT$ (= 75 meV), the Fermi-Dirac equation reduces to the Boltzmann expression, and integrating from the bottom of the conduction band E_c to ∞ (or top of the conduction band) gives us the following equilibrium electron concentration $n(E, T)$:

$$n(E,T) = \int_{E_c}^{\infty} G(E)f(E)dE = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} \times \int_{E_c}^{\infty} (E - E_c)^{1/2} e^{-(E-E_F)/kT} dE \qquad (4.13)$$

and

$$n(E,T) = 2\left(\frac{m_e^* kT}{2\pi\hbar^2}\right)^{\frac{3}{2}} e^{-\left(\frac{E_C - E_F}{kT}\right)} = N_c e^{-\left(\frac{E_C - E_F}{kT}\right)} \qquad (4.14)$$

where $N_c = 2\left(\frac{m_e^* kT}{2\pi\hbar^2}\right)^{\frac{3}{2}}$ is called the *effective DOS in the conduction band*.

For the concentration of holes, p, in the valence band, again assuming a Boltzmann distribution ($E - E_F > 75$ meV) and integrating $p(E, T)dE = G(E)[1 - f(E)]dE$ from $-\infty$ (or bottom of the valence band) to E_v, we similarly obtain:

$$p(E,T) = 2\left(\frac{m_h^* kT}{2\pi\hbar^2}\right)^{\frac{3}{2}} e^{-\left(\frac{E_F - E_V}{kT}\right)} = N_v e^{-\left(\frac{E_F - E_V}{kT}\right)} \qquad (4.15)$$

where $N_v = 2\left(\frac{m_h^* kT}{2\pi\hbar^2}\right)^{\frac{3}{2}}$ is called the *effective DOS in the valence band*.

The values for N_c and N_v are calculated from first principles based on the detailed band structures: for Si, one obtains $N_c = 2.8 \times 10^{19}$ cm^{-3} and $N_v = 1.04 \times 10^{19}$ cm^{-3} (at 300 K). These relations for $n(E, T)$ and $p(E, T)$ are valid for any semiconductor, with or without impurities. There is a one-to-one relation

between Fermi energy and local carrier density; if one adds electrons, E_F will increase, and if one removes electrons, E_F will decrease. In pure "intrinsic" semiconductors, the concentration of electron (n) and holes (p) arises only from the balance of electron and hole formation; therefore, $n = p = n_i$, where n_i is the material's intrinsic carrier concentration, and the law of mass action prescribes:

$$n_i^2 = n \times p \tag{4.16}$$

This equation remains valid when dealing with a nonintrinsic semiconductor (doped) in thermal equilibrium. In an ideal crystal, at equilibrium, the concentration of the intrinsic carriers thus is given as a product of Equations 4.14 and 4.15, or:

$$n_i = \sqrt{N_c N_v}\exp\left(-\frac{E_g}{2kT}\right) \tag{4.17}$$

a result that depends on the width of the bandgap ($E_g = E_c - E_v$) and temperature but is independent of the Fermi level E_F. The concentration of intrinsic free electrons/holes, n_i, generated this way can be expressed as:

$$n_i = 3.9 \times 10^{16}\ T^{\frac{3}{2}} e^{-0.56\ eV/kT}\ cm^{-3} \tag{4.18}$$

with $n_i = 1.45 \times 10^{10}$ cm^{-3} at 300 K. This translates into a room temperature resistivity for pure Si of $2.3 \times 10^5\ \Omega$cm. To get a feel for this number of intrinsic carriers generated in Si, realize that at room temperature there is only one pair of thermally produced carriers in a volume of Si containing 10,000 billion silicon atoms. The intrinsic carrier density as a function of temperature is shown for intrinsic Si and GaAs in Figure 4.15 (the intrinsic carrier concentrations at room temperature are marked by bullets).*

From Equations 4.14 and 4.15, equating n and p, we derive the following expression for E_F:

$$E_F = \frac{1}{2}E_g + \frac{kT}{2}\ln\left(\frac{m_h^*}{m_e^*}\right)^{\frac{3}{2}} \tag{4.19}$$

* The difference between calculated intrinsic carrier densities and experimental values may be off by a factor of 2.

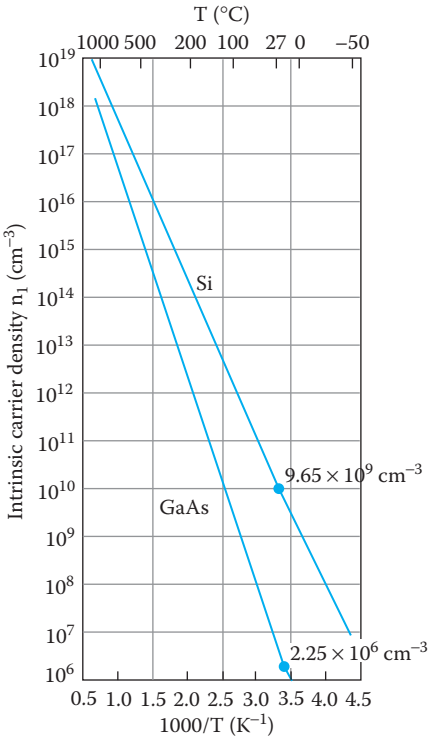


FIGURE 4.15 Intrinsic carrier concentration as a function of temperature for Si and GaAs.

In other words, the chemical potential μ, or Fermi level E_F, has some dependence on T, but if m_h^* and m_e^* are similar, that dependence becomes very small. In the latter case, the Fermi level is situated very nearly in the middle of the bandgap, as sketched in Figure 4.14. Semiconductors and insulators have qualitatively similar band structures, with the quantitative distinction that the bandgap $E_g > 3.0$ eV in insulators.

Single-Crystal Si Growth

Silicon, the second-most abundant element on the Earth after oxygen, is the principal ingredient of beach sand. To make wafers from sand, the silicon must be refined and purified. The refined silicon is melted, and trace amounts of impurities are added for doping the silicon n- or p-type and is then crystallized to form boules. Silicon boules are sliced into wafers and polished.

Silicon crystal growth comprises the primary process toward the construction of ICs and most of the existing micro- and nanomachines. In the Czochralski (CZ) crystal pulling method (Figure

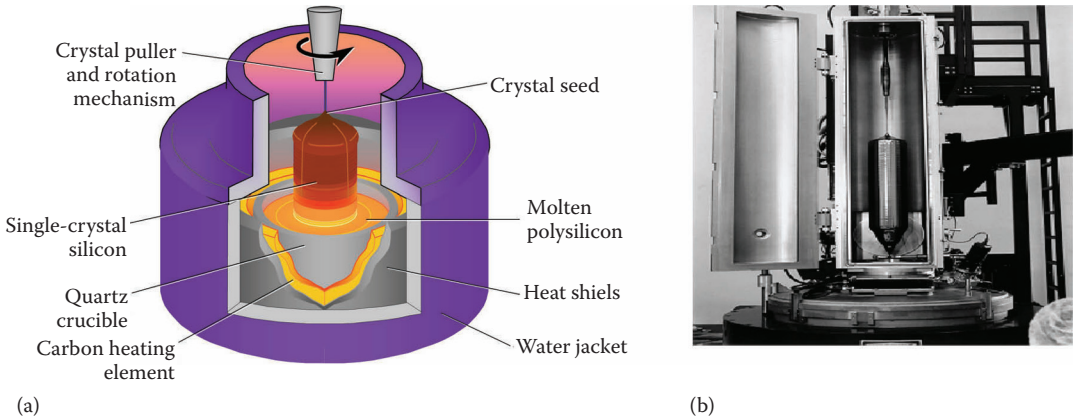


FIGURE 4.16 Schematic (a) and actual picture (b) of a Czochralski crystal puller.

4.16), a silicon crystal seed is grown into a single-crystal Si by pulling it slowly upward, at about 2 to 5 cm/h, from a molten and ultrapure (99.999999999%) polysilicon melt (Si melts at 1414°C). The molten silicon slowly rotates, nested in a silica crucible while pulling the growing crystal. Fused silica reacts with hot silicon and releases oxygen into the melt. Hence, CZ-Si has an indigenous oxygen concentration of approximately 10^{18} atoms/cm^3. The melt must not be stirred; otherwise, more oxygen is transported from the SiO_2-Si (crucible/liquid) interface to the Si (liquid)/Si (solid) interface.

Float zone crystal growth (Figure 4.17) is used to avoid all contact with crucibles to eliminate impurities such as oxygen. This method requires a slow rotating polycrystalline silicon rod, locally melted and recrystallized by a scanning RF heating coil, resulting in very pure silicon. The latter type of silicon is better suited for devices requiring a low doping level to produce, for example, low leakage diodes as required for detectors and power devices. In the float zone process, the regrowing silicon crystal rejects dopants and other impurities. Impurities tend to stay in the liquid, and refining can be accomplished with multiple melting/recrystallization passes. The latter process is called zone refining. The best purity achieved by chemical means is typically one impurity in every 1 million atoms. With zone refining this can be reduced to less than one impurity in every 10 billion atoms.

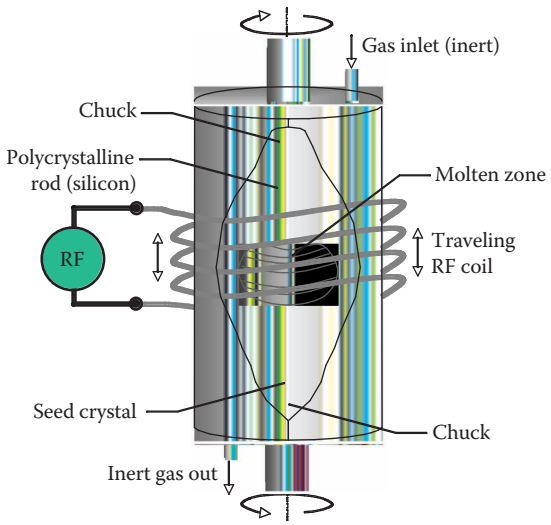


FIGURE 4.17 Float zone crystal growth.

Modern silicon boules can reach a diameter of more than 300 mm and a length of 1–2 m. They are first characterized for resistivity and crystal perfection. Resistivity is measured with a four-point probe (see Volume III, Chapter 6 on metrology), whereas x-ray or electron beam diffraction helps evaluate the crystal perfection (Chapter 2). To prepare wafers from the grown Si boule, both ends (the seed and the tail) are cut off, and the diameter of the boule is trimmed by grinding on a lathe. Based on the x-ray crystal orientation, one or more flats are grinded along the length of the ingot (Figure 4.18). The accuracy of this process is ±0.5° (in the very best case). Wafer flats (illustrated in Figure 4.3) help orientation determination and placement of slices in cassettes and fabrication equipment (large primary flat) and help identify orientation and conductivity type (smaller secondary flat). On a {100} wafer, the largest flat is oriented in the <110> direction. The drawback of these flats is that the usable area on the wafer is significantly decreased.

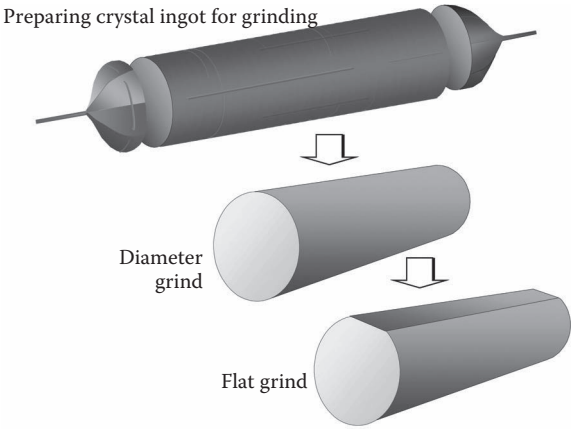


FIGURE 4.18 Ingot cap (seed and tail) cutting, and diameter and flat grind.

For some 200-mm and 300-mm diameter wafers, only a small notch is cut from the wafer to enable lithographic alignment, and no dopant type or crystal orientation information is conveyed.

The ingot/boule is then sliced into wafers, typically using inner diameter (ID) sawing, guided by argon laser markings. Mechanical grinding rounds the wafer edges. Wafer-edge rounding cannot be taken lightly because it removes microcracks at the wafer edge, makes the wafers more fracture resistant, increases the mask life, and minimizes resist beading. The edge-rounding process also makes for safer quartz boat loading by increasing the number of transfers possible of those wafers without chipping. The slices are then lapped with a mixture of hard Al_2O_3 powder in glycerin to produce a wafer with flatness uniform to within 2 µm. After a damage-removing etch, the wafers are polished, cleaned, and inspected. The slicing, lapping, edge treatment, damage removal, and polishing result in a material loss of up to 50%.

Typical specifications for 125-mm to 300-mm diameter Si wafers are summarized in Table 4.1.

Notice that the thickness control on Si wafers is limited, which has important consequences for fabricating thin Si membranes.

The Si wafer preparation process is summarized in Figure 4.19.

Doping of Si

Extrinsic versus Intrinsic Semiconductors

Intrinsic semiconductors, we saw earlier, are undoped semiconductors with electrical properties native to the material. Extrinsic semiconductors are doped semiconductors with electrical properties controlled by added impurities. Doping is the intentional addition of impurities to a material through ion implantation or diffusion. Doping by ion implantation and diffusion is different from other semiconductor processes because these processes do not create a new layer on the wafer. Instead, ion implantation and diffusion change the electrical characteristics of precise areas within an existing layer on the wafer.

On incorporation into the crystal lattice, the dopant either gives up (donor) or receives (acceptor) an electron from the crystal. In the former case, an extra electron is available (e^-), and in the latter case a hole (h^+) is created. Adding either electrons (e^-) or holes (h^+) varies the conductivity (σ) of the material. Adding more electrons makes the material more n-type, and adding more holes makes the material more p-type. Elements having one more or one less valence electron than Si (group IV) are used as substitutional donors (to make n-type) or acceptors (to make p-type), respectively. The first type of dopant is pentavalent, and the second type is trivalent. Boron (III), phosphorus (V), arsenic (V), and antimony (V) represent the more commonly

TABLE 4.1 Typical Specifications for Monocrystalline Silicon Wafers

Parameter	125 mm	150 mm	200 mm	300 mm
Diameter (mm)	125 ± 1	150 ± 1	200 ± 1	300 ± 1
Thickness (mm)	0.6–0.65	0.65–0.7	0.715–0.735	0.755–0.775
Bow (µm)	70	60	30	<30
Total thickness variation (µm)	65	50	10	<10
Surface orientation	±1°	±1°	±1°	±1°

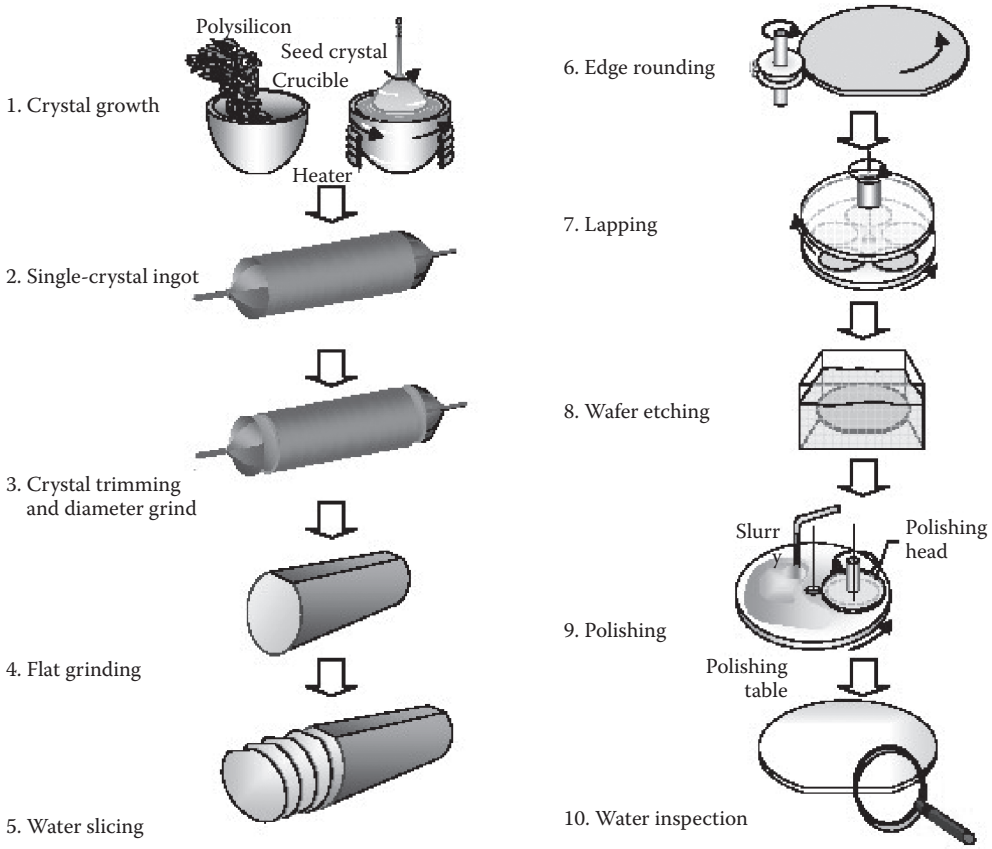


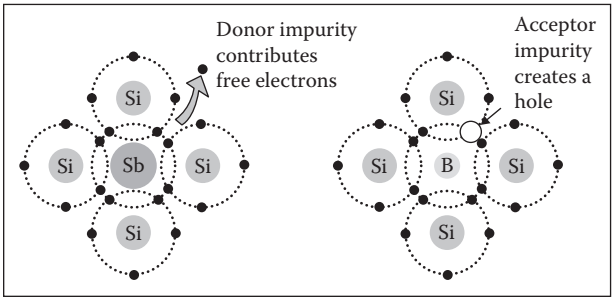
FIGURE 4.19 Silicon wafer preparation.

used dopant elements for single-crystal silicon (see Table 4.2). The two principal means of doping Si in wafer form are diffusion and ion implantation (as illustrated in Figure 4.20).

Dopants can be incorporated in the ingots during crystal growth or in sliced wafers in a planar process. Fabrication of circuit elements and micromachines also requires a method for selective n- or p-type doping of the silicon substrate. For diffusion, a uniformly doped ingot is sliced into wafers; an oxide film is grown on the wafers; and the oxide film is patterned and etched using photolithography, exposing specific sections of the silicon to be doped. The wafers are then coated with a solid doping source by spin

TABLE 4.2 Doping of Silicon with Pentavalent and Trivalent Atoms

		Concentration (atoms/cm^3)			
Dopant	Material Type	$<10^{14}$ (very lightly doped)	10^{14}–10^{16} (lightly doped)	10^{16}–10^{19} (doped)	$>10^{19}$ (heavily doped)
Pentavalent	n	n^{--}	n^{-}	n	n^{+}
Trivalent	p	p^{--}	p^{-}	p	p^{+}



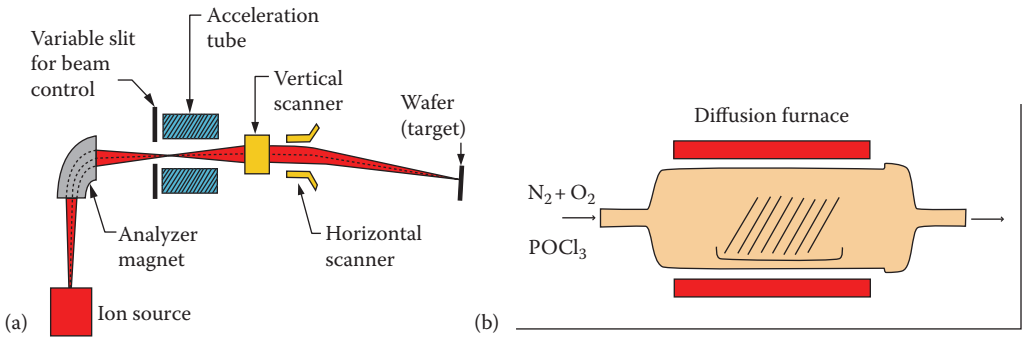


FIGURE 4.20 Ion implantation (a) and diffusion (b) to dope a silicon wafer.

coating and are heated in a furnace at 800–1250°C to drive the doping atoms into the exposed silicon areas. In ion implantation, a particle accelerator is used to accelerate doping atoms so that they can penetrate a silicon wafer to a depth of several micrometers. The process creates some lattice damage, which may be repaired by annealing the wafer at a moderate temperature for a few minutes.

Si Doping by Diffusion

Doping of silicon by diffusion is a chemical process carried out in a tube furnace to introduce a minute amount of impurities and to uniformly spread the incorporated impurities (Figure 4.20b). Dopant atoms may move into lattice vacancies, defining substitutional diffusion, or they may displace silicon atoms, moving them to interstitial positions, in an interstitial model. Finally, dopants may simply go between the Si atoms without displacing them (Figure 4.21).

Dopants are typically diffused thermally into a Si substrate at temperatures between 950°C and 1250°C. Silicon dioxide is used for a mask. Typical dopant sources used are BBr_3, $POCl_3$ (liquids), Sb_2O_3, As_2O_3, P_2O_5, B_2O_3 (solids), PH_3, AsH_3, and B_2H_6 (gases). To a first approximation, Fick's first and second laws describe the diffusion of dopants in a substrate. The dopant flux J is proportional to the concentration gradient as given by Fick's first law:

$$J = -D\frac{\delta N(x,t)}{\delta x} \tag{4.20}$$

where D is the diffusion coefficient (units of cm^2/s), and $\delta N(x,t)/\delta x$ is the concentration gradient of the dopant. The concentration N is a function of position x and time t, and in this expression, D is assumed constant, which is a reasonable assumption when the concentration of the dopant is low. Equation 4.20 has a negative sign, and $\delta N(x,t)/\delta x$ is negative for a decrease in the concentration with depth; consequently, the flux into the sample has a positive value.

The flux gradient ($\delta J/\delta x$) is proportional to the change of the concentration with time:

$$\frac{\delta J}{\delta x} = -\frac{\delta N(x,t)}{\delta t} = -D\frac{\delta^2 N(x,t)}{\delta x^2} \tag{4.21}$$

This is also called the *continuity equation* or Fick's second law. Solutions for Equation 4.21 have been obtained for various simple conditions, including diffusion with a constant surface concentration (N_s) and diffusion with constant total dopant amount (Q). In the first scenario, impurity atoms may be transported from a vapor source onto the semiconductor surface and diffuse into the semiconductor wafer. In the second scenario, a fixed amount of dopant is already in the semiconductor, and a heat treatment distributes it further into the material.

In gaseous doping, the initial condition at t = 0 is N(x,0) = 0, which states that the dopant concentration in the host semiconductor is initially zero. The boundary conditions are that the dopant

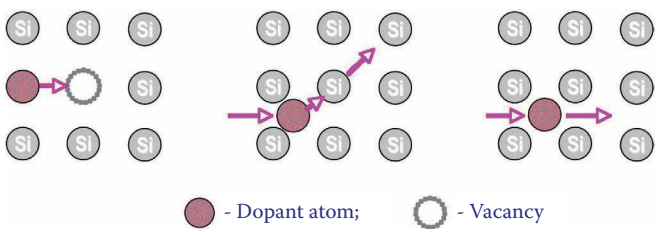


FIGURE 4.21 Diffusion mechanisms of dopant atoms in Si.

concentration at the Si surface is constant [N(0,t) = N_s], whereas in the bulk, the concentration remains zero [N(∞,t) = 0] during the whole doping process, resulting in the following solution for the differential Equation 4.21:

$$N(x,t) = N_s \text{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \tag{4.22}$$

where erfc = complementary error function (see Figure 4.22a and Appendix 4A)
$\sqrt{Dt}$ = diffusion length
N_s = surface concentration of the dopant (in cm^{-3})

Generally speaking, this same boundary condition also applies for the predeposition of dopant atoms onto semiconductor substrates. The temperature of the substrates during predeposition is usually kept between 800°C and 1200°C.

The total amount of dopant atoms per unit area of the semiconductor, or the dose, $Q(t)$, is given by integration of $N(x,t)$ from x = 0 to x = ∞:

$$Q(t) = \int_0^\infty N_s \text{erfc}\left(\frac{x}{2\sqrt{Dt}}\right)dx = \frac{2\sqrt{Dt}}{\sqrt{\pi}} N_s \cong 1.13 N_s \sqrt{Dt} \tag{4.23}$$

The gradient of the diffusion profile, $\delta N(x,t)/\delta x$, can be obtained by differentiating Equation 4.22:

$$\frac{\delta N(x,t)}{\delta x} = -\frac{N_s}{\sqrt{\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right) \tag{4.24}$$

Another boundary condition applies when the dopant concentration at the silicon surface is limited; in other words, the diffusing amount Q is constant and C(∞,t) = 0. In this boundary condition, the surface dopant concentration decreases as the material supply is depleted. The initial condition at t = 0 is again N(x,0) = 0, and the boundary conditions are mathematically expressed as:

$$\int_0^\infty N(x,t)dx = Q \text{ and } N(\infty,t) = 0 \tag{4.25}$$

where Q is the total amount of dopant or the dose. The solution of the diffusion equation satisfying the above conditions is:

$$N(x,t) = \frac{Q}{\sqrt{\pi Dt}} \exp\left(\frac{-x^2}{4Dt}\right) \tag{4.26}$$

In this case, the resulting dopant concentration profile is Gaussian instead of an erfc function (see Figure 4.22b). At the surface x = 0, and Equation 4.26 at t = 1 becomes:

$$N_s = N_{01} = \frac{Q}{\sqrt{\pi Dt}} \tag{4.27}$$

or the surface concentration changes as $(Dt)^{-1/2}$. Therefore, the dopant surface concentration decreases

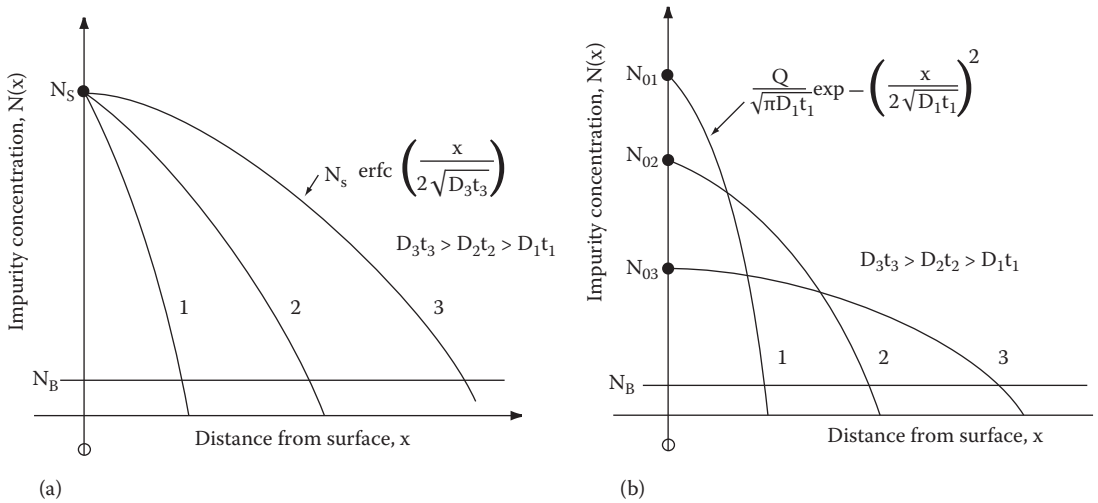


FIGURE 4.22 Diffusion profiles. (a) Normalized complementary error function (erfc) versus distance for successive diffusion times. N_s is the surface concentration, and N_B is the bulk concentration. (b) Normalized Gaussian function versus distance for successive diffusion times.

with time because the dopant moves into the semiconductor as time increases (see Figure 4.22b from N_{01} to N_{03}, with $t = 1 < t = 2 < t = 3$). The gradient of the diffusion profile is obtained by differentiating Equation 4.26:

$$\frac{\delta N(x,t)}{\delta x} = -\frac{x}{2Dt} N(x,t) \qquad (4.28)$$

The gradient is zero at $x = 0$ and at $x = \infty$, and the maximum gradient is obtained at $x = \sqrt{2Dt}$. The complementary error function and the Gaussian distribution are both functions of a normalized distance, $x/2(Dt)^{1/2}$. Hence, by normalizing the dopant concentration with the surface concentration, each distribution can be represented by a single curve valid for all diffusion times, as shown in Figure 4.23.

This type of boundary condition is typical for drive-in diffusion. Drive-in diffusion is a high temperature (>800°C) operation performed on semiconductor wafers in an inert ambient. The heat treatment causes motion of dopant atoms in the direction of the concentration gradient (diffusion) and is used to drive dopant atoms deeper into the semiconductor. The dopants introduced in the substrate by a predeposition step are redistributed deeper into the substrate by such a drive-in step. A two-step diffusion sequence is commonly used to form an n-p junction; during predeposition, a diffusion layer is formed under a constant surface concentration condition, which is followed by a drive-in diffusion, or redistribution, under a constant total dopant condition. For most practical cases, the diffusion length $(Dt)^{1/2}$ for the predeposition diffusion is much smaller than that for the drive-in condition. Hence the predeposition profile can be treated as a delta function at the surface. This two-step junction formation is illustrated in Figure 4.24 with a predeposition from a constant source (erfc) and a limited source diffusion (Gaussian).

The D in the above equations is the dopant diffusivity and is given by:

$$D = D_0(T_0)\exp\left(-\frac{E_a}{kT}\right) \qquad (4.29)$$

where E_a is the activation energy, whose value depends on the transport mechanism of the dopant

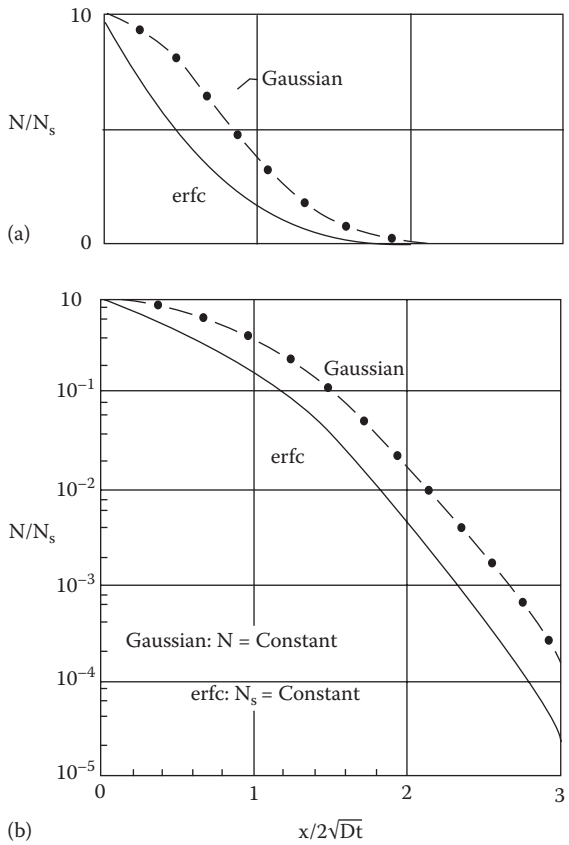


FIGURE 4.23 Normalized concentration versus normalized distance for the erfc and Gaussian functions. (a) Linear plot. (b) Semilog plot.

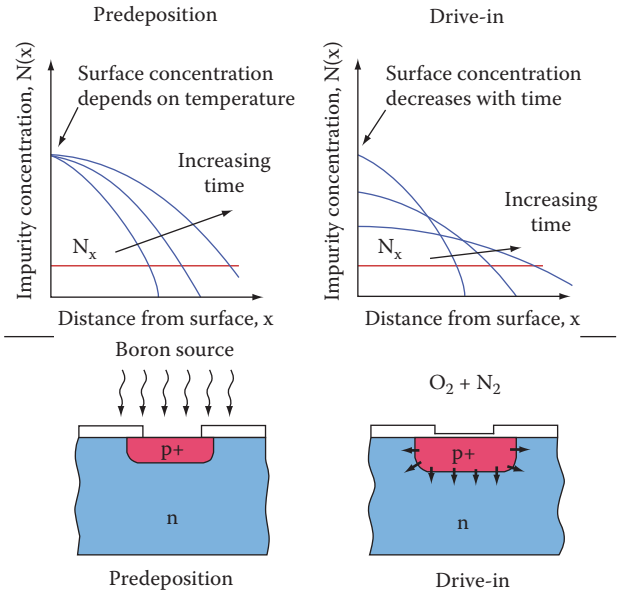


FIGURE 4.24 Two-step junction formation with a predeposition from a constant source and a limited source diffusion.

TABLE 4.3 Typical Values for D_0 and E_a for Commonly Used Dopant Atoms

Element	D_0 (cm²/s)	E_a (eV)
B	10.5	3.69
Al	8.00	3.47
Ga	3.60	3.51
In	16.5	3.90
P	10.5	3.69
As	0.32	3.56
Sb	5.60	3.95

atoms. The diffusion coefficient, D_0, is only a constant for a given reaction and concentration.

In Table 4.3, typical values for D_0 and E_a are given for commonly used dopant atoms. Dopant atoms take on several different ionization states and diffuse by several different mechanisms. The effective diffusion coefficient is the sum of the diffusivities of each species, which are in turn a function of temperature and dopant concentration. Hence, actual dopant concentration profiles deviate from the simple first-order functions given above.

The dopant diffusivities of phosphorus, arsenic, and boron in silicon at 1100°C are typically 2×10^{-13}, 1×10^{-13}, and 1×10^{-13} cm²/s, respectively. Below the intrinsic carrier concentration (~5×10^{18} cm^{-3} for Si at 1000°C), D is independent of the carrier concentration. Above the intrinsic concentration, the diffusion increases with carrier concentration and generally follows a power law as a result of the diffusion mechanism becoming point defect (vacancies and interstitials) assisted.

The point at which the diffused impurity profile intersects the background concentration is the metallurgical junction depth, x_j. At that depth, the net impurity concentration is zero. In the case of a complementary error function, the depth is defined as:

$$x_j = 2\sqrt{Dt\,\mathrm{erfc}^{-1}\left(\frac{N_B}{N_S}\right)} \tag{4.30}$$

When dealing with a Gaussian distribution, one obtains:

$$x_j = 2\sqrt{Dt\ln\left(\frac{N_s}{N_B}\right)} \tag{4.31}$$

The latter case is illustrated in Figure 4.25, where we show the diffusion of an n-type impurity in a p-type Si background substrate.

It is important to remember that the concentration profile of the dopant in the semiconductor never changes abruptly (this would imply a zero or infinite flux), so diffusion always spreads impurities gradually around the edges of any mask. Lateral diffusion is ~85% of the junction depth. This lateral diffusion continues every time the wafer is heated, resulting in an expanding die area. Besides lateral diffusion, diffusion doping also suffers from surface contamination interference, poor doping control, dislocation generation, and difficulty in creating ultrathin junction depths. Next, we will see that for each generation of IC technology, the trend has been to reduce the junction depth, which favors ion implantation (see comparison Table 4.4).

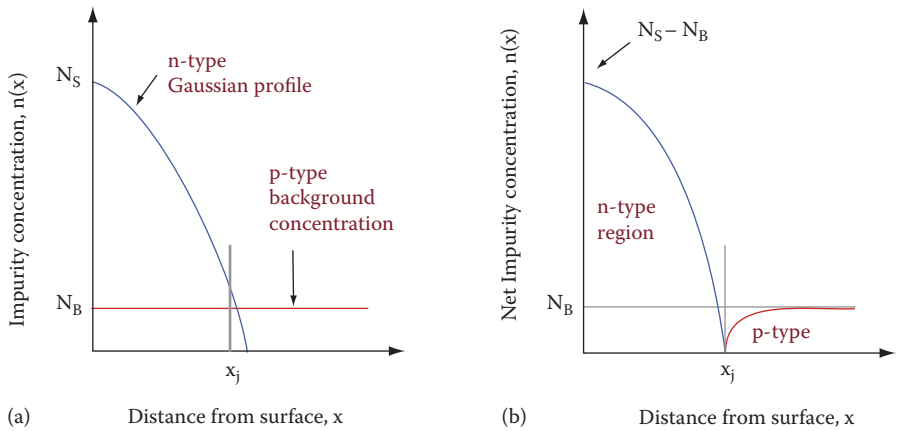


FIGURE 4.25 Diffusion of n-type impurity into a p-type Si background substrate. Determination of x_j in case of a Gaussian dopant distribution. Two representations are shown. In (a), the *y*-axis represents the impurity concentration; in (b), the *y*-axis represents the net impurity concentration (i.e., corrected for the background impurity concentration N_B).

Table 4.4 Comparison of Thermal Diffusion and Ion Implantation

Ion Implantation and Annealing	Solid-/Gas-Phase Diffusion
Advantages	
Room temperature mask	No damage created by doping
Precise dose control (10^{11}–10^{16}/cm^2)	Batch fabrication
	Cheaper
Accurate depth control	
Disadvantages	
Implant damage enhances diffusion	Usually limited to solid solubility
Dislocations caused by damage may cause junction leakage	Low surface concentration hard to reach without a long drive-in
Implant channeling may affect profile	Low dose predeposition very difficult to achieve
Very expensive	

Si Doping by Ion Implantation

Ion implantation is the introduction of selected impurities (dopants) by means of high-voltage ion bombardment to achieve desired electronic properties in defined areas. Photolithography is used to mask the regions of the device where doping is not desired. For masking materials, most films used in the IC process can be used in the ion-implantation process—photoresists, SiO_2, Si_3N_4, metal film, etc.—in contrast with diffusion, where choices are much more limited. High energy ion beam implantation is a low-temperature physical process and offers the advantage of being able to place ions at various depths in the sample, independent of the thermodynamics of diffusion and problems with solid solubility and precipitation. Ion implantation for doping was invented by Shockley in 1954. The technique is now very commonly used, with penetration depths in silicon of As, P, and B typically being 0.5, 1, and 2 µm at 1000 keV. A large range of extremely accurate doses is accessible: from 10^{11} to 10^{16}/cm^2, and even buried (or retrograde) dopant profiles are possible. Unfortunately, however, ion beams produce crystal damage because the implant energy—between 30 and 100 keV—is thousands of times the silicon bond energy. Crystal damage can reduce electrical conductivity, but fortunately most of the damage can be eliminated by annealing at 700–1000°C, which is below the diffusion temperature. The anneal temperature is high enough for Si interstitial atoms to recombine with vacant lattice sites and for the dopant atoms to move to substitutional lattice sites where they are electrically active. The combination of excellent spatial and dose control, as well as ease of manufacture, has led to widespread use of ion implantation. For the newest and future generation of complementary metal oxide silicon (CMOS) devices, which require highly doped, abrupt, and very shallow profiles (junction depths less than 30 nm!), the dopant diffusion step during annealing must be minimized.

A beam of energetic ions implants charged dopant atoms into the silicon substrate. The acceleration energy and the beam current control the implantation depth and dopant concentration. The implanted dopant atoms enter unmasked regions in the silicon lattice at high velocity. The stopping mechanism of the ions involves nuclear collisions at low energy and electronic interactions at high energy.

The acronyms associated with ion implantation include:

- Projected range, R_p: average distance traveled by ions parallel to the beam
- Projected straggle, ΔR_p: standard deviation or fluctuation in the projected range
- Lateral straggle, $\Delta R//$: standard deviation or fluctuation in the final rest position, perpendicular to the beam (also called *transverse straggle*)
- Peak concentration, N_p: concentration of implanted ions at R_p

In the absence of crystal orientation effects, the concentration profile, to a first-order approximation, is Gaussian:

$$N_i(x) = N_p \exp\left[-\frac{1}{2}\left(\frac{x - R_p}{\Delta R_p}\right)^2\right] \qquad (4.32)$$

The range, R_p, is determined by the acceleration energy, the ion mass, and the stopping power of the material. Heavy ions do not travel as far into the crystal as light ions. They have a more narrow distribution (ΔR_p) than light ions. The crystalline symmetry of the silicon lattice can have a dramatic impact on the implant profile. Once an ion enters a lattice channel, the small angle scattering events on the channel walls translate to ion penetration for a long distance before electron drag or a sharp collision stops the ion. Orientation of the substrate surface away from the plane perpendicular to the beam (tilt of 3–7°) prevents channeling, which can occur along crystal planes, ensuring better reproducibility of R_p. Ion channeling leads to an exponential tail in the concentration versus depth profile. This tail results from the crystal lattice and is not observed in an amorphous material. Channeling may also be prevented by creating a damaged layer on the Si wafer surface or by depositing a surface-blocking amorphous layer such as an oxide. The three methods to limit ion channeling are compared in Figure 4.26.

The total implanted dose (ions/cm^2), Q_i, can be either calculated by integrating the charges per unit time (current i) arriving at the substrate over time or also by integrating Equation 4.32 over all of x:

$$Q_i = \int i\,dt = \int_0^{+\infty} N_i(x)\,dx = \sqrt{2\pi}\,\Delta R_p N_p \qquad (4.33)$$

Technologies studied to make the very shallow junctions needed in the latest CMOS devices are ultralow-temperature processing and rapid thermal annealing. For deep diffusions (>1 μm), implantation is used to create a dose of dopants, and thermal diffusion is used to drive in the dopant.

Diffusion is a cheaper and more simplistic method than ion implantation but can only be performed from the surface of the wafers. Dopants also diffuse unevenly and interact with each other, altering the diffusion rate. Ion implantation is more expensive and complex. Ion implantation, however, except for annealing, does not require high temperatures and allows for greater control of dopant concentration and profile. It is an anisotropic process and therefore does not spread the dopant implant as much as diffusion. This aids in the manufacture of self-aligned structures, which greatly improve the performance of MOS transistors. Ion beams can implant enough material to form new materials, for example, oxides and nitrides, some of which show improved wear and strength characteristics. Ion implanters may be purchased from Varian Semiconductor (http://www.vsea.com) and Eaton Corporation (http://www.eaton.com). This equipment is very expensive, and usually service providers are used instead.

In Table 4.4 we compare thermal diffusion and ion implantation.

Band Diagram of an Extrinsic Semiconductor

In this section we establish the band diagram and carrier density for extrinsic semiconductors. The conductivity σ (Ωm)$^{-1}$ is an intrinsic characteristic of a material and is independent of geometry. In extrinsic semiconductors, negative and positive charge carriers may contribute to the conductivity. In Chapter 3 we saw that the conductivity is

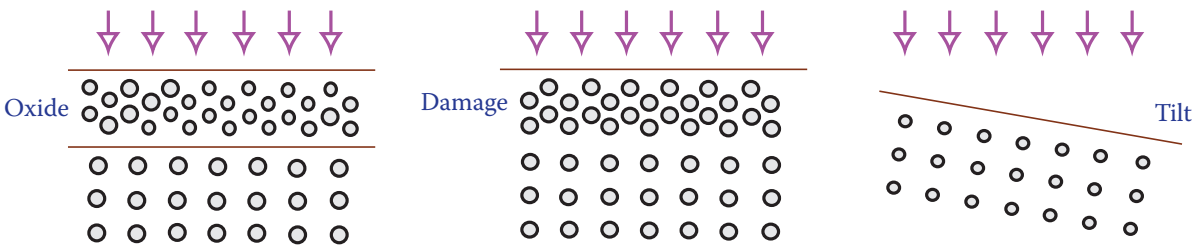


FIGURE 4.26 How to prevent ion channeling during implantation.

the product of the number of free charges and their mobility [for electrons (e^-) the mobility in $cm^2V^{-1}s^{-1}$ is symbolized as μ_e (see Equation 3.10)]:

$$\sigma = \mu_e ne \tag{3.10}$$

The positive charge carriers (h^+) are called *holes* as they correspond to an absence of electrons, and μ_h is their mobility. When both the concentration of holes (p) and the concentration of electrons (n) are considered, the conductivity σ of a solid is given as:

$$\sigma = 1/\rho = e(\mu_e n + \mu_p p) \tag{4.34}$$

where e is the electron charge, n is the electron concentration, and p is the hole concentration. In the case of a metal, we saw that resistance increases with temperature because the number of carriers remains the same but scatters more at higher temperatures. In a semiconductor, the resistance decreases as the temperature increases because more free carriers are generated (n and p increases). A semiconductor is called n-type when $e\mu_n n$ dominates the conductivity, in which case the majority of carriers are electrons, and the semiconductor is p-type if $e\mu_p p$ dominates, in which case the majority carriers are holes.

By adding a donor atom (energy E_D) to an intrinsic semiconductor at T = 0 K, an extra electron is bound to the positive charge on the donor atom. When the donor, at higher temperatures, ionizes, the neutral donor concentration N_D converts to positively charged N_D^+. The electron bound to the positive ion is in an energy state, $E_D = E_g - \Delta E$, where ΔE is the binding energy. An electron that moves onto an acceptor atom has energy E_A and coverts a neutral acceptor concentration N_A into a negatively charged N_A^-. In Figure 4.27, we show donor ions (E_D) energetically close to the bottom of the conduction band (E_C) and acceptor ions (E_A) close to the top of the valence band (E_V).

To estimate the Bohr's radius and the binding energy (ΔE) involved in the donor holding onto its electron, we use the simple Bohr model for the hydrogen atom that was introduced in Chapter 3 but adapt the pertinent parameters to describe an impurity atom in a crystal rather than a hydrogen atom in vacuum. In Chapter 3 we calculated the Bohr radius of the hydrogen atom using Equation 3.89 as:

$$a_0 = \frac{\hbar^2 4\pi\varepsilon_0}{m_e e^2} = 0.53\ \text{Å} \tag{3.89}$$

For the binding energy involved, we use Equation 3.91:

$$\Delta E = E_n = -\frac{m_e e^4}{32\hbar^2\pi^2\varepsilon_0^2 n^2} = -\frac{E_0}{n^2} \tag{3.91}$$

With the Bohr radius equal to 0.53×10^{-10} m and n = 1, we get a value of 13.6 eV = 21.8×10^{-19} J for

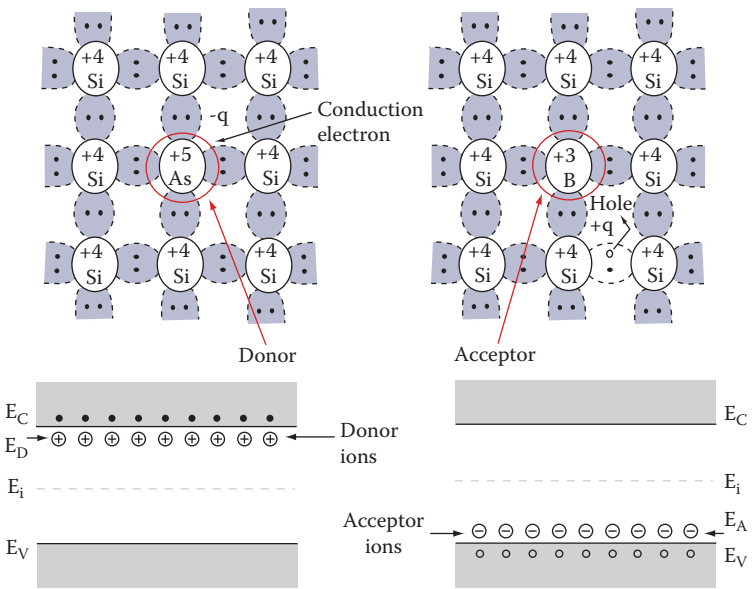


FIGURE 4.27 Donors and acceptors in the Si lattice and their position in an energy band diagram. Here E_i stands for the Fermi level of the intrinsic material.

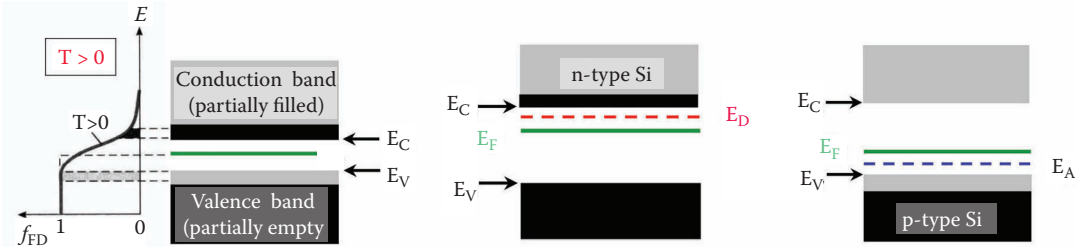


FIGURE 4.28 Band diagram for intrinsic and extrinsic Si compared. Black is filled with electrons, and gray is empty.

the energy required to remove an electron from a hydrogen atom (see also Figure 3.36). For a typical solid, we saw in Chapter 3 one must replace m_e with the electron effective mass $m_e^* \sim 0.15\ m_e$ and ε_r of 1 in the case of hydrogen with $\varepsilon_r \sim 15$ for a typical semiconductor so that a 53-Å Bohr radius is obtained from:

$$a_0 = \frac{0.53 \times \varepsilon_r}{m_e^*/m_e} = \frac{0.53 \times 15}{0.15} = 53\ \text{Å} \qquad (4.35)$$

and 9 meV is calculated for the binding energy of the electron to the donor atom from:

$$\Delta E = \frac{13.6}{\varepsilon_r^2}\frac{m_e^*}{m_e}\ \text{eV} = \frac{13.6}{(15)^2} \times 0.15 = 9\ \text{meV} \qquad (4.36)$$

The ionization energy of the donor tells us how far below the conduction band the donor energy level is situated ($\Delta E = E_C - E_D$). A similar calculation provides us with the distance above the valence band edge of the acceptor levels in a p-type material. This Bohr-like doping model assumes no interaction between dopant sites, i.e., a nondegenerate semiconductor.

At T = 0 K, all electrons are in the lowest available energy states, and no electrons are excited from donor states (N_D) into the conduction band (n = 0), but with a low binding energy of 9 meV at room temperature ($kT \sim 25$ meV $> \Delta E = E_C - E_D$) all of the donors will be ionized (N_D^+), and the number of free electrons in the conduction band n is given by:

$$n = N_D^+ \sim N_D \qquad (4.37)$$

This is the relevant regime for most electronics; note that the density of electrons is independent of temperature. Electrons do freeze out (i.e., return to the valence band) if kT becomes too small to overcome the binding energy. From Equation 4.14 we calculate that:

$$E_F = E_c - k_B T \ln\left(\frac{N_C}{n}\right) \qquad (4.38)$$

where $N_C = 2.8 \times 10^{19}\ \text{cm}^{-3}$ and $n \sim N_D \sim 10^{17}\ \text{cm}^{-3}$; we can see that the Fermi level is about 150 meV below the conduction band (~6 kT). This situation is sketched in the middle panel of Figure 4.28, where we compare intrinsic silicon with n-type and p-type Si. Analogously, from Equation 4.15 we derive:

$$E_F = E_v - k_B T \ln\left(\frac{N_v}{p}\right) \qquad (4.39)$$

Some typical values for energy levels of impurities in Si are listed in Table 4.5.

The Bohr-like doping model assumes no interaction between dopant sites; if adjacent dopant atoms are within 2 Bohr radii, then orbits overlap, and we are dealing with a degenerate semiconductor. This happens when N_D (~n) ~ N_c or when N_A (~p) ~ N_v. The degenerate semiconductor is then defined by E_F ~/> E_C or E_F ~/< E_v. In these cases, the Fermi level is situated either inside the conduction or the valence band.

TABLE 4.5 Some Typical Impurity Levels in Si

Phosphorous, P	$E_C - E_D = 44$ meV
Arsenic, As	$E_C - E_D = 49$ meV
Boron, B	$E_A - E_V = 45$ meV
Aluminum, Al	$E_A - E_V = 57$ meV
Gallium, Ga	$E_A - E_V = 65$ meV
Gold, Au	$E_A - E_V = 584$ meV
	$E_C - E_D = 774$ meV

Oxidation of Silicon

Oxidation Kinetics of Si-Silica Is One of the Keys to Silicon's Success

Silicon owes a lot of its success to its stable oxide. By contrast, germanium's oxide is soluble in water, and GaAs produces only leaky oxides. In IC design, silicon oxides, like other dielectrics, function as insulation between conducting layers, as diffusion and ion masks, as capping layers of doped oxides to prevent the loss of dopants, for gettering impurities, and for passivation to protect devices from impurities, moisture, and scratches. Doped oxides also act as a solid-state source of ions that can be diffused into a substrate at high temperatures. In micromachining, silicon oxides serve the same purposes but also very often act as sacrificial material. The stability and ease of formation of SiO_2 on a Si surface was one of the reasons that Si replaced Ge as the semiconductor of choice in the late 1960s. Silica is the distinguishing feature of silicon versus other important semiconductor materials (see Table 4.6). Silica is inert to most chemicals, protects the Si surface, and forms a reasonably good diffusion barrier (although Si_3N_4 is much better in this respect). It is thermally stable up to 1600°C (T_m is 1830°C), easy to grow, and easy to pattern and etch selectively. In wet etching, HF preferentially etches SiO_2, not Si, and in dry etching it can easily be plasma etched in CHF_3. The material also makes for an excellent dielectric (insulator) with a 9-eV bandgap, a dielectric strength of 10^7 V/cm, and a resistivity ρ of $10^{15}\Omega$cm.

Silicon dioxide growth involves the heating of a silicon wafer in a stream of steam or in wet or dry oxygen/nitrogen mixtures (Figure 4.29) at elevated temperatures (between 600°C and 1250°C) in a horizontal or vertical quartz furnace. Dry oxide grows more slowly, but the oxide layers are more uniform, and there are also relatively few defects at the oxide-silicon interface. Defects interfere with the proper operation of semiconductor devices, and because a dry oxide has especially low surface state charges, it forms an ideal thin dielectric for MOS transistors (30–100 Å; see below). In wet oxides, hydrogen atoms, liberated by the decomposition of the water molecules during oxidation, produce imperfections that degrade the oxide quality somewhat. A major advantage of wet oxides includes their faster growth rate, making them useful for thick layers of field oxides (>3000 Å; see below). Silicon readily oxidizes, even at room temperature, forming a thin native oxide approximately 20 Å thick. The high temperature aids diffusion of oxidant through the growing surface oxide layer to the interface to form thick oxides quickly:

$$Si + 2H_2O\ (gas) \rightarrow SiO_2 + H_2\ (wet\ oxidation)$$

Reaction 4.1

$$Si + O_2\ (gas) \rightarrow SiO_2\ (dry\ oxidation)$$

Reaction 4.2

Another gas-phase oxidation method of Si is the pyrogenic method (Figure 4.29). Besides oxygen and nitrogen, the gas also contains hydrogen, which generates ultrapure steam. In the oxidizing gas stream, one also may add 2% HCl or Cl_2; this helps to remove (getter) impurities by volatizing any metal impurities ($AuCl_3$ and $CuCl_2$).

TABLE 4.6 The Question of Good-Quality Native Oxides on Semiconductors

Semiconductor (Bandgap)	Native Oxide
Si (1.1 eV)	Yes
Ge (0.7 eV)	No
GaAs (1.4 eV)	No
SiC (~2.2 eV)	No
GaN (~3 eV)	No
Diamond (5 eV)	No

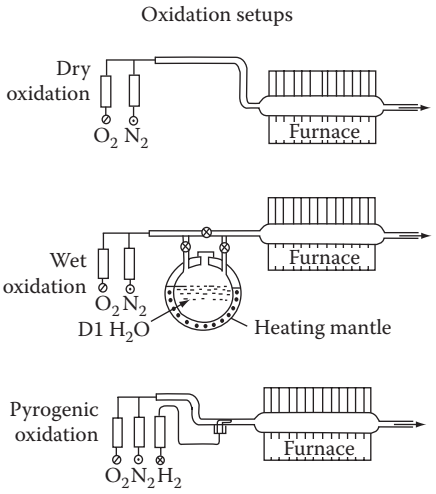


FIGURE 4.29 Oxidation setups.

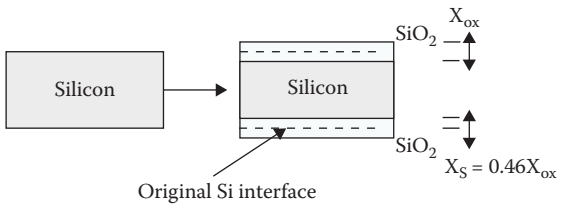


FIGURE 4.30 Silicon thickness converted, X_S. Oxidation by inward diffusion of oxidant through the growing oxide to the interface.

The ratio of silicon thickness converted (Figure 4.30), X_S, to resulting oxide thickness, X_{ox}, is proportional to their respective densities:

$$X_S = 0.46X_{ox} \tag{4.40}$$

When 10,000 Å of oxide has grown, 4,600 Å of Si will be consumed; in other words, the amount of silicon consumed is 46% of the final oxide thickness. This relationship is important for calculating step heights that form in silicon microstructures. Very good adhesion to Si leads to volume expansion upward, resulting in compressive stress in the SiO_2 film.

All gas-phase oxidation processes involve three steps in series: gas-phase transport of oxidant to the surface (F_1), diffusion through the existing oxide (F_2), and the oxidation reaction itself (F_3). The fluxes are:

$$F_1\ (\text{gas}) = h(C^* - C_{ox})$$
$$F_2\ (\text{oxide}) = D(C_{ox} - C_i)/x_{ox} \tag{4.41}$$
$$F_3\ (\text{reaction}) = k_s C_i\ (\text{see Figure 4.31})$$

At equilibrium $F_1 = F_2 = F_3$, and this yields the concentration of oxidant at the interface oxide/silicon, C_i, as:

$$C_i(X_{ox}, k_s, D, P_G) \tag{4.42}$$

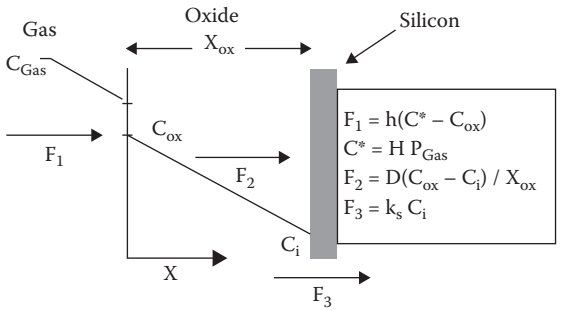


FIGURE 4.31 Fluxes in the silicon oxidation reaction.

where X_{ox} = oxide thickness
k_S = Si oxidation rate constant, which is a function of temperature, oxidant, crystal orientation, and doping
D = oxidant diffusivity, which is a function of temperature and oxidant
P_G = partial pressure of the oxidant in the gas phase
C* = equilibrium oxidant concentration in the oxide = HP_G, with *H* being Henry's law constant

The oxide growth rate is given by:

$$\frac{dX_{ox}}{dt} = \frac{C_i k_s}{N} \tag{4.43}$$

where *N* stands for the number of molecules of oxidant per unit volume of oxide (2.2×10^{22} cm^{-3} for dry oxygen and 4.6×10^{22} cm^{-3} for wet oxidation). Solution of this differential equation in the Bruce Deal-Andy Grove model for oxidation (1965) yields:

$$X_{ox}(t) = \frac{A}{2}\left[\left(\sqrt{1 + \frac{t + \tau}{A^2/4B}}\right) - 1\right] \tag{4.44}$$

where *A* and *B* are rate constants:
$A = 2D(1/k_s + 1/h)$ (units of μm)
$B = 2DC^*/N$ (units of μm^2/h—parabolic rate constant)
$\tau = (X_i^2 + AX_i)/B$
h = gas-phase mass transport coefficient in terms of concentration in the solid
C* = equilibrium oxidant concentration in the oxide = HP_G, with *H* being Henry's law constant
X_i = initial oxide thickness

The reason that wet oxide growth is much faster than dry oxidation can be found in the value for *C** in the rate constant *B*. In dry oxygen at 1000°C, $C^*(O_2) = 5.2 \times 10^{16}$ cm^{-3}, whereas at the same temperature in wet oxygen $C^*(H_2O) = 3.0 \times 10^{19}$ cm^{-3} (three orders of magnitude larger). This more than compensates for the fact that the diffusion coefficient of oxygen in SiO_2 is larger than that of water [$D(O_2) > D(H_2O)$]. This general relationship for silicon oxidation is illustrated in Figure 4.32. From this figure it can be seen that for thin oxides [short

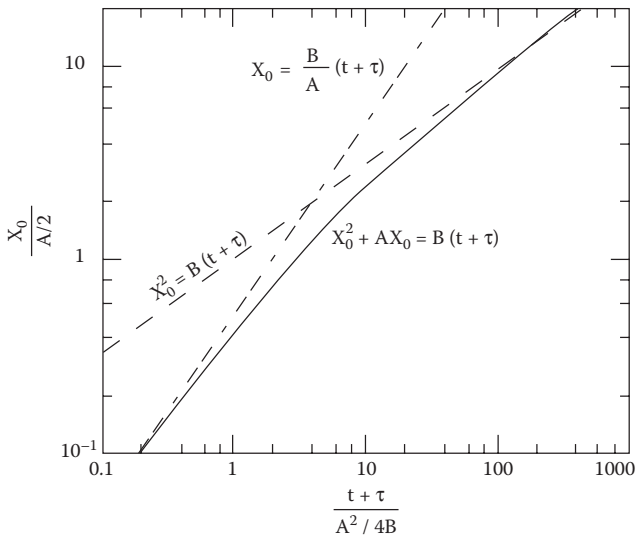


FIGURE 4.32 Silicon dioxide thickness versus growth time.

oxidation times with $(t + \tau) \ll A^2/4B$], the process is reaction rate limited:

$$X_{ox}(t) = \frac{B}{A}(t+\tau) \qquad (4.45)$$

where B is the parabolic rate constant, and B/A is the linear rate constant (in μm/h). For thick oxides (long oxidation times with $t \gg \tau$ and $t \gg A^2/4B$), we deduce from Figure 4.32 that the process is diffusion limited, and a parabolic law results:

$$X_{ox}(t) = \sqrt{B(t+\tau)} \qquad (4.46)$$

where B is the parabolic rate constant. From Figure 4.32 we also conclude that oxide thickness versus growth time is initially linear when the oxidation reaction at the surface dominates but parabolic when the diffusion to the oxide/silicon interface dominates.

A, B, and τ are strong functions of temperature and the gas ambient (wet vs. dry) and weak functions of orientation and doping (both type and level). The oxidation rate is also influenced by the presence of halogen impurities or hydrogen in the oxidizing gas, the oxidizing gas pressure, and the use of a plasma or a photon flux during oxide growth.

For $X_0 < 200$ Å, the oxidation is always faster than that predicted by the Deal-Grove model, and empirical models are used to fit the data. Even the substrate cleaning method influences the oxide thickness in this regime. The Deal-Grove model is also limited to the 700–1300°C temperature range and to the 0.1–25 atm pressure range.

Short of using an ellipsometer to determine the oxide thickness (see Volume III, Chapter 6 on metrology), the silicon oxide thickness color table (Table 4.7) comes in handy. This table also lists some typical applications for oxides of various thicknesses. Charts showing the final oxide thickness as a function of temperature, oxidizing environment, and time are widely available.

TABLE 4.7 Oxide Color Table and Applications of Thermal Oxides: Observed Perpendicularly under Daylight Fluorescent Lighting

Color	Thickness (Å)				Application
Gray	Up to 100				Tunneling oxides in EEPROM (flash memories) and gate oxides (around 30 Å in 180-nm CMOS)
Tan	500				
Brown	700				Pad oxides
Blue	800				Masking oxides during implant, diffusion, etc.
Dark violet-red violet	1000	2800	4600	6500	Field oxide for isolation (e.g., LOCOS)
Light blue-metallic blue	1500	3000	4900	6800	Field oxide for isolation (e.g., LOCOS)
Green	1800	3300	5200	7200	
Yellow	2100	3700	5600	7500	
Gold with slight yellow orange	2200	4000	6000		
Orange-melon	2500	4400	6200		

Some new types of oxidation under investigation include the growth of thick oxides at high pressures and the growth of ultrathin oxides (<30 Å) in rapid thermal oxidation (RTO). High-pressure oxidation of up to 25 atm of O_2 allows for a reduced temperature in preparing thick oxide fabrication. In the case of ultrathin oxides, conventional thermal oxidation must be performed <800°C to reduce the interfacial roughness between Si and SiO_2, which leads to reduced transistor channel mobility. The solution is RTO around 1050°C; high quality gate oxides have been demonstrated this way, and present development is aimed at wet RTO with in situ steam generation (ISSG). This method is said to be yet five times faster than wet oxidation with better quality than dry oxidation.

With 20 million devices per chip today, electrically isolating them from each other—to prevent cross-talk—is becoming more of a difficult task. Historically, there have been three main choices: 1) field oxidation, 2) local oxidation of silicon (LOCOS), and 3) the latest, shallow trench isolation (STI). To isolate features using field oxides, thick field oxides (>10,000 Å) are grown and patterned over electrically inactive regions of the wafer and etched. In LOCOS, a silicon nitride mask allows for the selective thermal oxide growth in CMOS and bipolar (Bi)-CMOS processes. In STI, also known as "box isolation technique," high-aspect-ratio trenches (5 µm deep and 2 µm wide) are etched in the substrate, and then through repetitive, relatively short oxidations, the trench is refilled.

In Figure 4.33, we compare field oxidation (A) with the LOCOS process (B). It is the latter process we currently want to detail a bit further. The LOCOS process was for a time so important for the isolation of MOS devices that it became almost synonymous with the term "MOS," until recently, when LOCOS was supplanted by the "box isolation technique."

Localized Oxidation of Silicon—LOCOS

In LOCOS, as the name implies, the goal is to oxidize Si only locally, wherever a field oxide is needed.

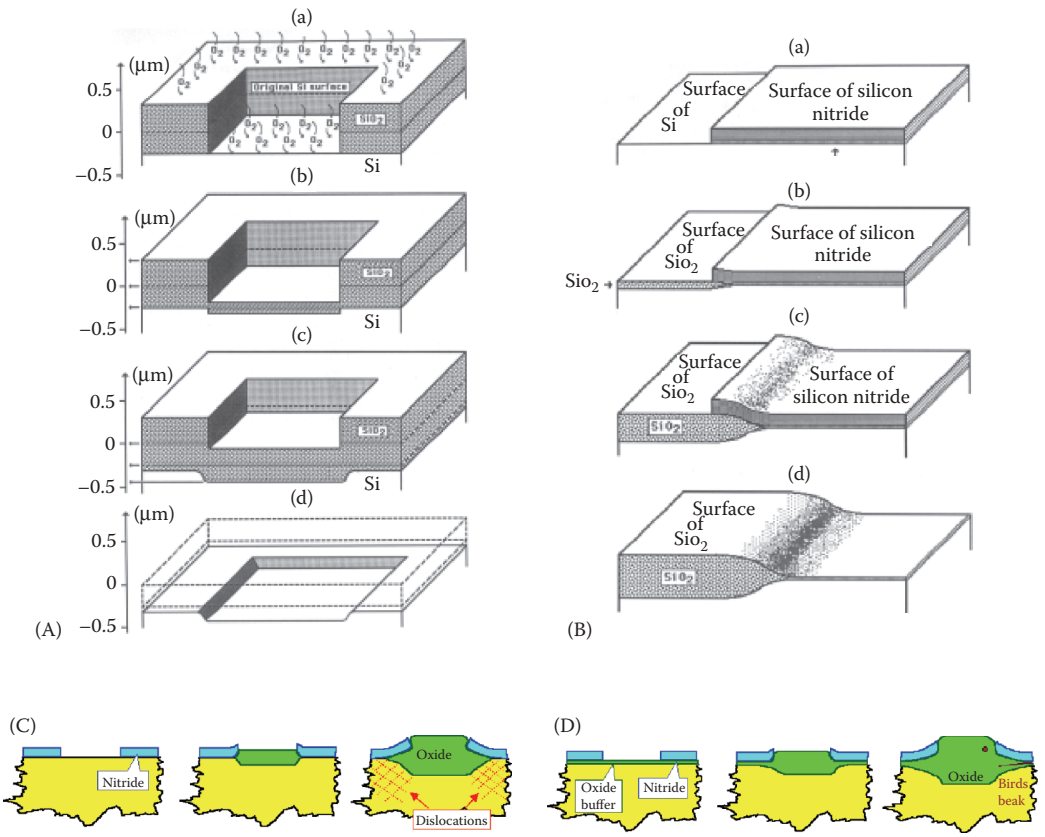


FIGURE 4.33 Device isolation through field oxide isolation (A) and local oxidation of Si (LOCOS) (B). (C) Silicon nitride is deposited directly on Si. (D) Silicon nitride is deposited on a buffer layer of oxide. See also the bird's beak.

For a local oxidation, a material that does not allow oxygen diffusion at the typical oxidation temperatures of 1000–1100°C must protect the areas of the Si not to be oxidized. The only material that is used for this purpose is silicon nitride, Si_3N_4. It can be deposited and structured easily, and it is compatible with Si IC processing. However, Si_3N_4 introduces two major problems when deposited directly on top of Si. Using a Si_3N_4 layer for protecting parts of the Si from thermal oxidation does not prevent oxygen diffusion through the oxide already formed from oxidizing the Si underneath the Si_3N_4 (lateral oxidation). Another major problem is the volume expansion that occurs in Si oxidation, which expands a given volume of Si so that the nitride mask is then pressed upward at the edges as illustrated in Figure 4.33C. With increasing oxidation time and oxide thickness, pressure under the nitride mask increases, and at some point, the critical yield strength of Si is exceeded. Plastic deformation then starts, and dislocations are generated and move into the Si. Below the edges of the local oxide a high density of dislocations may destroy the electronic function under construction. The oxidation step is not even necessary to produce the dislocations because Si_3N_4 layers are always under large stresses even at room temperature and exert shear stresses on the Si. The important message is that one cannot deposit Si_3N_4 directly on Si.

To save the LOCOS concept, one must introduce a buffer oxide. A buffer oxide between the Si_3N_4 mask and the Si substrate relieves the stress building up during oxidation. The SiO_2 buffer layer acts as a "grease" material: the hard Si_3N_4 is now pressing down on something "soft," and the stress felt by the Si does not reach the yield stress. This way one avoids dislocations, but a comparatively large lateral oxidation instead results, leading to a configuration known as "bird's beak" for the obvious reason shown in Figure 4.33D. The lateral extension of the field oxide via the bird's beak is comparable with its thickness and limits the minimum feature size. Although this was not a serious problem in the early days of IC technology, it could not be tolerated beyond the middle of the 1980s. One solution was the use of a poly-Si layer as a sacrificial layer. It was put on top of the buffer oxide below the nitride mask and structured with the same mask. It provided sacrificial Si for the "bird's beak," and the total dimension of the field oxide could be reduced. But even this is not sufficient for feature sizes around and below 1 µm.

The LOCOS process eventually became a very complicated process in its own right, and with feature sizes shrinking ever more, LOCOS reached the end of its useful lifespan in the 1990s and had to be replaced by "box isolation." The idea in box isolation is clear: etch a hole (with vertical sidewalls) in the Si wherever you want an oxide, and next simply "fill" it with oxide.

Orientation and Dopant Dependence of Oxidation Kinetics

The thermal oxidation rate, as mentioned above, is slightly influenced by the orientation of the Si substrate. The effect involves the linear oxidation rate constant used in the Deal-Grove model in the regime where the surface reaction is rate limiting. This constant was given as B/A (in µm/h) in Equation 4.45. The ratio of this constant, for a (111) Si plane to that for a (100) Si plane, is given by:

$$\frac{\frac{B}{A}(111)}{\frac{B}{A}(100)} = \frac{C_1(111)\exp\left(-\frac{2.0\ \text{eV}}{kT}\right)}{C_1(100)\exp\left(-\frac{2.0\ \text{eV}}{kT}\right)} = 1.68 \qquad (4.47)$$

Notice that the activation energy for both dry and wet oxidation in the surface reaction limited regime is 2 eV. Thus, a (100) surface oxidizes about 1.7 times more slowly than a (111) surface. The lower oxidation rate of (100) surfaces might be because of the fewer silicon bonds on the surface with which oxygen can react. The available bond density at a (111) plane equals 11.76×10^{14} cm^{-2} and 6.77×10^{14} cm^{-2} for the (100) plane. The linear oxidation rate for Si follows the sequence (110) > (111) > (311) > (511) > (100), corresponding to an increasing activation energy that incorporates a term for the bond density in the plane, as well as one for the bond orientation. As might be expected, steric hindrance results in higher activation energy.[4]

In anisotropic wet etching, as we will discover in Volume II, Chapter 4, the sequence of the etch rates

is mostly reversed compared with that of oxidation. For example, a (100) plane etches up to 100 times faster than a (111) plane. It has been postulated that the slow etching of a (111) Si plane relative to a (100) plane may be because of its more efficient oxidation, protecting it better against etching than a (100) plane.[5] In this interpretation—etching being an electrochemical process—one assumes that anodic oxidation, just like thermal oxidation, will occur faster at a (111) Si surface. To our knowledge, no evidence to support this theory has been presented as of yet. Moreover, Seidel et al.[6] pointed out that the huge anisotropy in wet etching of Si could not be credited to the number of bonds available on different Si planes, as that number barely differs by a factor of 2 (see Volume II, Chapter 4). From the numbers observed, such an explanation could be valid for the anisotropy in thermal oxidation. There is no agreement, however, about the exact mechanism, and an understanding of the orientation dependence of the oxidation rate is still lacking. It has been attributed at various times to the number of Si-Si bonds available for reaction and the orientation of these bonds, the presence of surface steps, mechanical effects such as stress in the oxide film, and the attainment of maximum coherence across the Si/SiO_2 interface.[4] It should also be noted that the oxidation rate sequence depends on temperature and oxide thickness—one more reason to take any model with a grain of salt.[4]

In Si, n-type dopants (As and P) at high concentrations (n^{++}) lead to increases in oxidation rate, particularly at low temperatures. It is found that B/A is affected but that B remains unaffected. Ho-Plummer explained this phenomenon by postulating that point defects (vacancies) are being consumed during the volume expansion that accompanies Si oxidation.[7,8] The oxidation rate is changed much less in case of highly boron-doped (p^{++}) wafers. Whereas boron atoms are taken up by the oxide, P and As pile up at the Si/SiO_2 interface.[9]

Oxidation is also influenced by stress. Etched substrates exhibiting shaped structures oxidize at different rates than planar wafers. Oxide layers formed on silicon are under compressive stress, even in the planar case, and these stresses can be much larger on curved surfaces because the volume expansion is dimensionally confined. For example, oxidation occurring in a confined corner, in which volume expansion is more difficult, is different from oxidation of a flat surface. Retardation of oxidation is very strong for sharp corners, i.e., for r small or $1/r$ large (e.g., a factor of two retardation in some cases). Retardation is also more pronounced for lower temperatures, and there are virtually no corner effects observed for 1200°C. Interior (concave) corners show more pronounced retardation effects than exterior (convex) corners.

It is important to include stress when simulating oxide growth in the LOCOS process (see above). Si oxidation modeling using ATHENA software predicts the oxide shape much better when stress is included as a parameter (http://www.silvaco.com/tech_lib/simulationstandard/1997/aug/a3/a3.html). In Volume III, Chapter 10 we show how oxidation sharpening is used to make sharp Si needle tips for scanning tunneling microscopes (STMs) and atomic force microscopes (AFMs; Volume III, Figure 10.12).

Nonthermal Methods for Depositing SiO_2 Films

Thermal oxidation involving the consumption of a thin layer of the Si surface results in excellent adhesion and good electrical and mechanical properties. It is an excellent technique when available. However, the thermal oxidation described in the previous section is by no means the only method available to implement a SiO_2 film. Two nonthermal SiO_2 deposition methods include electrochemical oxidation of Si and chemical vapor deposition. Electrochemical oxidation of Si has been studied (see, e.g., Schmidt[10] and Lewerenz[11]) but has not led to commercial usage. Even though these anodic oxides are pinhole free, they exhibit many interface states. Madou et al.[12] investigated mechanisms for anodic oxidation of Si and the introduction of oxide "dopants" into these anodically formed oxides. Such oxide dopants can be used as a source for doping of the underlying Si at elevated temperatures. Further study material on silicon oxidation can be found in Fair[13] and Katz.[4]

Silicon dioxide is frequently used as an insulator between two layers of metallization. In such cases, some form of deposited oxide must be used rather

than grown oxides. Deposited oxides can be produced by various reactions between gaseous silicon compounds and gaseous oxidizers. Deposited oxides tend to possess low densities and large numbers of defect sites, meaning they are not suitable for use as gate dielectrics for MOS transistors but are acceptable for use as insulating layers between multiple conductor layers or as protective overcoats.

For deposition of oxides at lower temperatures, we will learn in much more detail in Volume II, Chapter 7 that one may use low-pressure chemical vapor deposition (LPCVD), such as in the following reaction:

$$SiH_4 + O_2 \rightarrow SiO_2 + 2\ H_2\text{: } 450°C \qquad \text{Reaction 4.3}$$

Another advantage, besides the lower temperature, is the fact that one can dope an LPCVD silicon oxide film easily to create, for example, a phosphosilicate glass (PSG):

$$SiH_4 + 7/2\ O_2 + 2\ PH_3 \rightarrow SiO_2\text{: } P + 5\ H_2O\text{: } 700°C \qquad \text{Reaction 4.4}$$

Phosphorus-doped glass, also called *P-glass* or *PSG*, and borophosphosilicate glasses (BPSGs) soften and flow at lower temperatures, enabling the smoothing of topography. They etch much faster than SiO_2, which benefits their application as sacrificial material in surface micromachining (see Volume II, Chapter 7).

Properties of Thermally Grown SiO_2

A film of thermal SiO_2 performs excellently as a mask against diffusion of the common dopants in Si. The diffusion coefficient, D, of boron at 900°C in SiO_2 is 2.2×10^{-19} cm^2/s compared with 4.4×10^{-16} in Si. For phosphorus, D equals 9.3×10^{-19} cm^2/s compared with 7.7×10^{-15} in Si (from the common Si dopants, only Ga diffuses fast through the oxide with $D = 1.3 \times 10^{-13}$). For some other elements, SiO_2 forms a poor diffusion barrier. The amorphous oxide has a more open structure than crystalline quartz—only 43% of the space is occupied (see Figure 4.34). Consequently, a wide variety of impurities (especially alkali ions such as Na^+ and K^+) can readily diffuse through amorphous SiO_2. Diffusion through the open SiO_2 structure happens especially fast when the oxide is hydrated. One

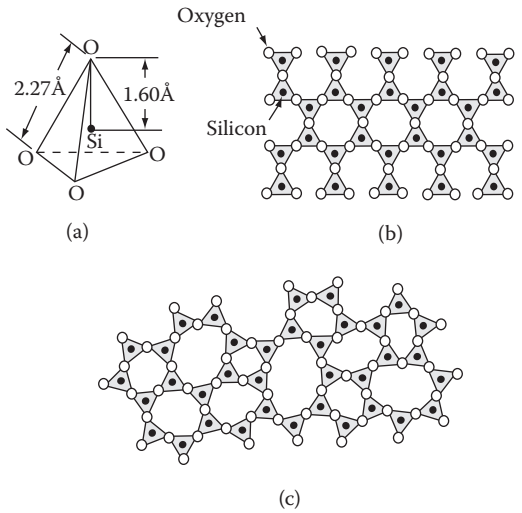


FIGURE 4.34 (a) Basic structural unit of silicon dioxide. (b) Two-dimensional representation of a quartz crystal lattice. (c) Two-dimensional representation of the amorphous structure of silicon dioxide.

of the reasons for the poor performance of silicon dioxide-based ion-sensitive field-effect transistors (ISFETs) in aqueous solutions can be traced back to the simple observation that SiO_2 in water almost behaves like a sponge for ions.[14] This is why, in such devices, SiO_2 is often topped off with the ionic barrier material Si_3N_4. The use of SiO_2 as a diffusion mask often stems from convenience; SiO_2 is easy to grow, whereas Si_3N_4 cannot be put directly onto Si without problems.

The quality of silicon dioxide depends heavily on its growth method. Dry oxidation at high temperatures (between 900 and 1150°C) in pure oxygen produces a better quality oxide than steam oxidation. Such a thermal oxide is stoichiometric, has a high density, and is basically pinhole free. Wet oxidation in steam occurs much faster but produces a lesser quality oxide, with water causing a loosening effect on the SiO_2, making it more prone to impurity diffusion. Both types of oxidation are carried out in a quartz tube. Oxide thicknesses of a few tenths of a micrometer are used most frequently, with 1–2 μm being the upper limit for conventional thermal oxides. Addition of chlorine-containing chemicals during oxidation increases the dielectric breakdown strength and the rate of oxidation while also improving the threshold voltage of many electronic devices.[15] However, too high concentrations of halogens at high temperatures can pit the silicon surface. The

influence of chlorine, hydrogen, and other gases on the SiO_2 growth rate and the resulting interface and oxide quality has been a fertile research field, a short summary of which can be found in Fair.[13] It should be noted that for the IC industry, the electronic quality of the Si/SiO_2 interface is expressed through low concentrations of interface trap states, low fixed oxide surface charges, low bulk oxide trapped charges, and mobile charges (impurity ions such as Li^+, Na^+, K^+) (Figure 4.35);[16] the interface quality is of little consequence in micromachining. In micromachining, the oxide is used as a structural element, a sacrificial layer, or a dielectric in a passive device, applications where the interface quality rarely matters. One notable exception is the aforementioned micromachined ISFET, where the electronic properties of the gate oxide are key to its functioning; a perfect oxide and oxide/semiconductor interface hold as much importance in this case as it does in the IC industry.

Because of molecular volume mismatch and thermal expansion differences, a thermal silicon oxide is compressed. The compressive stress depends on the total thickness of the silicon dioxide layer and can reach hundreds of MPa. Oxide films thicker than 1 μm can cause bowing of the Si wafer. Wet oxidation relaxes the compressive stress while speeding up the SiO_2 growth.

In Table 4.8, we review some of the more significant properties of thermally grown SiO_2. The oxide is characterized by a variety of means reviewed in Volume III, Chapter 6 on metrology. Electrical tests are conducted to establish the capacitance (CV) at high and low frequencies to calculate the interface density, to measure dielectric breakdown, and for electrical stressing (with temperature ramp). Optical tests are carried out for thickness and optical constant determination using color (see Table 4.7), interferometry (needs calibrated substrate), and ellipsometry. Physical tests for thickness, surface quality, etc. involve profilometry, AFM, and cross-sections by scanning electron microscopy (SEM)/transmission electron microscopy (TEM).

Oxidation furnaces may be purchased from Thermcraft (http://www.thermcraftinc.com/profile.html) and Tystar (http://www.tystar.com) (which makes Tytan Horizontal Furnace Systems; http://www.tystar.com/furnace.htm).

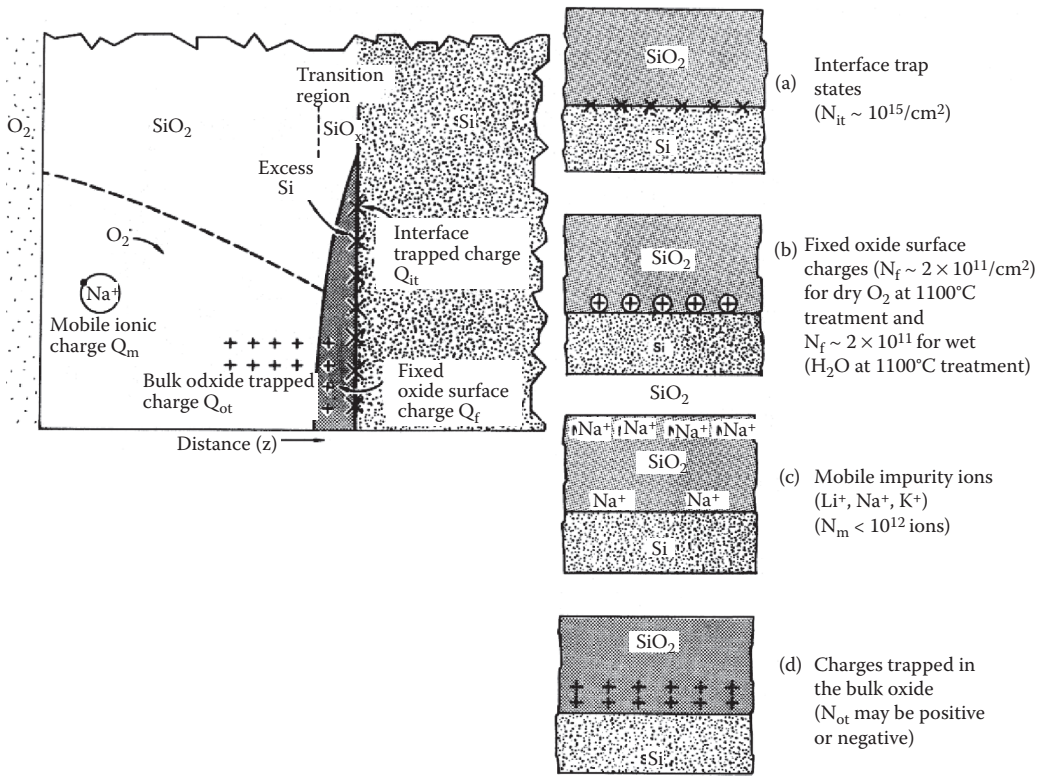


FIGURE 4.35 Location of oxide charges in thermally oxidized silicon structures as defined by Bruce Deal. Q_f and Q_{it} are much lower for (100) silicon than for (111).

TABLE 4.8 Properties of Thermal SiO_2

Property	Value
Density (g/cm³)	2.24–2.27 (dry) and 2.18–2.20 (wet)
Dielectric constant	3.8–3.9
Dielectric strength (V/cm)	2×10^6 (dry) and 3×10^6 (wet)
Energy gap (eV)	~8
Etch rate in buffered HF (Å/min)	1000
Infrared absorption peak (μm)	9.3
Melting point (°C)	~1700
Molecular weight	60.08
Molecules/cm³	2.3×10^{22}
Refractive index	1.46
Resistivity at 25°C (Ωcm)	3×10^{15} (dry) and $3–5 \times 10^{15}$ (wet)
Specific heat (J/g°C)	1.0
Stress in film on Si (dyne/cm²)	$2–4 \times 10^9$ (compressive)
Thermal conductivity (W/cm°C)	0.014
Thermal linear expansion coefficient (°C⁻¹)	5×10^{-7} (0.5 ppm/°C)

The End of the Line for SiO_2?

In the following sections we learn about the gate oxide thickness crisis. The projected gate oxide thickness for 70 nm CMOS of 10–15 Å results in large tunneling currents (see Chapter 3) through the oxide. It is a challenge to find a replacement for thermal silicon dioxide with comparable dielectric properties. Oxides with high dielectric constant $\varepsilon_r >> 3.9$ are most desirable. Possible SiO_2 replacements developed at present are silicon oxynitride (SiO_xN_y), Al_2O_3, Ta_2O_5, and Hf and Zr silicates.

Si-Based Electronic Devices

Introduction

The following active electronic devices will be reviewed in this section: diodes, special-purpose diodes [including light-emitting diodes (LEDs), photodiodes, solar cells, Zener diodes, avalanche diodes, and tunneling (Esaki) diodes], and two types of transistors (bipolar and MOSFET). Si and Ge are the two most common single elements that are used to make diodes. A compound semiconductor that is commonly used is GaAs, especially in the case of LEDs because of its large direct bandgap. We analyze in some detail the scaling issues involved in the further miniaturization of Si MOSFETs, and in this context we introduce strained Si, which expands the prospect of continued Si use. Fundamental limits to MOSFET downscaling include not only electromagnetic, quantum mechanical, and thermodynamic effects but also technical and economic (practical) considerations. In the current volume Chapters 3 and 5 on quantum mechanics, and photonics, respectively and in Volume III, Chapter 3 on nanotechnology, alternative device operating principles that exploit the quantization effects, rather than being hampered by them, are investigated. These approaches include quantum wells, quantum wires, quantum dots, resonant tunneling diodes and transistors, superlattices, and molecular electronics.

Diodes

Many semiconductor devices are based on the properties of the junction between p-type and n-type semiconductors (Figure 4.36). Russell Kohl at Bell Labs

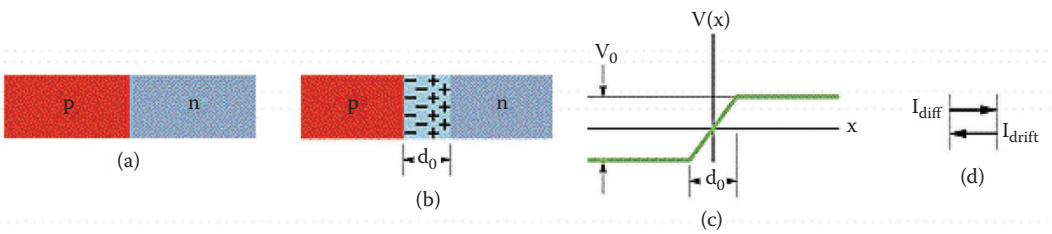


FIGURE 4.36 (a) p and n semiconductor before majority carriers cross the junction. (b) Motions of majority charge carriers across the junction uncover two space-charge layers associated with uncompensated donor ions (to the right of the plane) and acceptors ions (to the left). (c) Associated with the space charge is a contact potential difference V_0, which acts to limit the flow of majority carriers. (d) The diffusion of majority carriers across the junction produces a diffusion current, I_{diff}. The small concentration of minority carriers on either side of the junction moves in the opposite direction of the majority carriers, giving rise to a drift current, I_{drift}. In an isolated junction, the diffusion current is compensated for by the drift current, with the result that the net current through the junction is zero.

invented the solid-state p-n junction diode in 1940. This device, when exposed to light, produced a voltage of 0.5 V across the junction and represented a stepping stone to the invention of the transistor (see Chapter 1). In a diode p-n junction, electrons from the n-type diffuse into the p-type and holes diffuse the other way until the charge transfer builds up a big enough potential difference to halt the diffusion. As typical for diffusion, carriers flow from where there are many to where there are few. This motion of the majority charge carriers across the junction causes two space-charge layers to form associated with the uncompensated donor ions (to the right of the junction) and acceptors ions (to the left of the junction). The interface where the p- and n-type materials meet is called the metallurgical junction (Figure 4.36a). The two space-charge regions form a depletion zone that is free of mobile charge carriers, with the ionized donor and acceptor ions fixed in their lattice sites. This double layer of opposite net electric charges is also called a *dipole layer* or a *sandwich*. A depletion layer, d_0, typically extends a few micrometers on either side of the junction (Figure 4.36b). Associated with the space charge is a contact potential difference V_0, which acts to limit the flow of majority carriers with a self-induced electric field (10^3–10^5 V/cm), opposing further diffusion (Figure 4.36c). The diffusion of majority carriers across the junction produces the diffusion current, I_{diff}. A small concentration of minority carriers on either side of the junction moves in the opposite direction of the majority carriers and gives rise to a drift current, I_{drift} (Figure 4.36d). The electrical field is driving this current, not a concentration gradient. In a junction in equilibrium, the drift current compensates for the diffusion current, resulting in a net zero current through the junction. Even though there is no net current, electrons flow across the junction; it is just that the rates of flow in each direction are equal; such equilibrium is called *dynamic equilibrium*. In a real p-n junction, the boundaries of the depletion zones are not as sharp as shown in Figure 4.36, but the contact potential curve is smooth with no sharp corners.

The contact potential or barrier voltage, according to the Boltzmann relation, is given as:

$$V_0 = \frac{kT}{e} \ln \frac{p_p}{p_n} \tag{4.48}$$

where p_p = hole concentration in the p-type material
p_n = hole concentration in the n-type material

The term kT/e (= V_T) is the thermal voltage (~26 mV at room temperature). With $p_p \approx N_A$ and $p_n \approx n_i^2/N_D$, where n_i is the intrinsic carrier concentration, the barrier voltage can be approximated as:

$$V_0 = \frac{kT}{e} \ln \frac{N_A N_D}{n_i^2} \tag{4.49}$$

where N_A = concentration of immobile acceptors on the p-side of the junction
N_D = immobile donors on the n-side of the junction

For Si at room temperature, V_0 is typically around 0.67 V (with $N_A = 10^{17}$ cm^{-3} and $N_D = 10^{15}$ cm^{-3}). A typical width of the junction with these numbers is about 4 μm, and the electrical field E in this case is about 60 kV/cm.

The majority carriers in the depletion layer disappear through recombination in which electrons and hole pairs annihilate each other. When equilibrium is reached, a single Fermi level characterizes the p-n junction (Figure 4.37).

The above describes a p-n junction in an equilibrium situation. Now, we consider what happens when we apply a bias, as shown in Figure 4.38. When the p-side of a p-n junction is connected to the + terminal of a voltage source, and the n-side to the – terminal, the majority carriers are repelled by their respective terminals toward the junction. This is called *forward bias*, and with such a bias the depletion layer becomes narrower (d_F), and most electrons can go across the junction (right to left flow of electrons makes a positive current left to right because positive current flow is opposite of

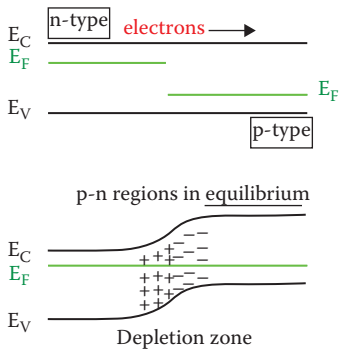


FIGURE 4.37 A single Fermi level characterizes the whole p-n junction in equilibrium.

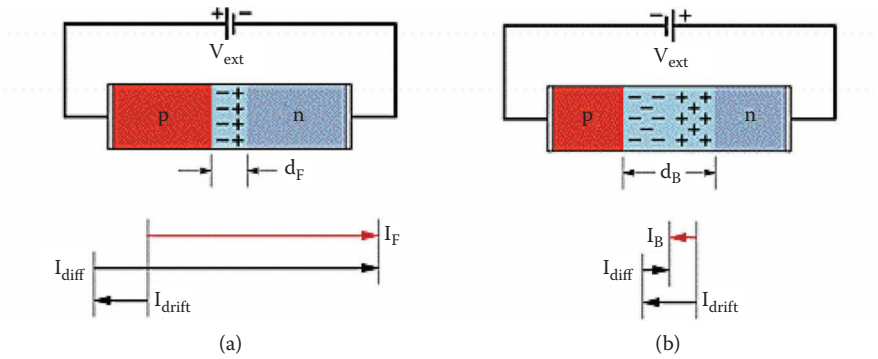


FIGURE 4.38 (a) The forward-biased connection of a p-n junction, showing the narrowed depletion (d_F) and the larger forward current I_F. (b) The reverse (back)-biased connection, showing the widening depletion zone (d_B) and the small reverse (back) current I_B.

the negative electron flow) (Figure 4.38a). With the shrinking of the depletion layer, the energy required for charge carriers to cross the depletion region decreases exponentially. Therefore, as the applied voltage increases, current starts to flow across the junction. The barrier potential of the diode is the voltage at which appreciable current starts to flow through the diode. The barrier potential varies for different materials. When connected in the opposite direction, the barrier gets higher, and practically no current flows. This is called *back bias* or *reverse bias*. The depletion layer is very wide (d_B) when the reverse bias is high (Figure 4.38b). Very few electrons get into the depletion layer from the neutral parts of the diode. Only electrons "produced" inside the depletion layer will move through it. A small leakage current, also saturation current, flows under these reverse-biased conditions.

A typical current-voltage plot (I–V) for a diode is shown in Figure 4.39. Mathematically, the forward current (I_F), with diffusion dominating, is described by an exponential as a function of the voltage. The reverse "leakage" current (I_B), where drift dominates, is almost constant over most of the reverse voltage range (negative potentials in Figure 4.39). Combined we obtain:

$$I = I_0(e^{\frac{eV}{\eta kT}} - 1) \qquad (4.50)$$

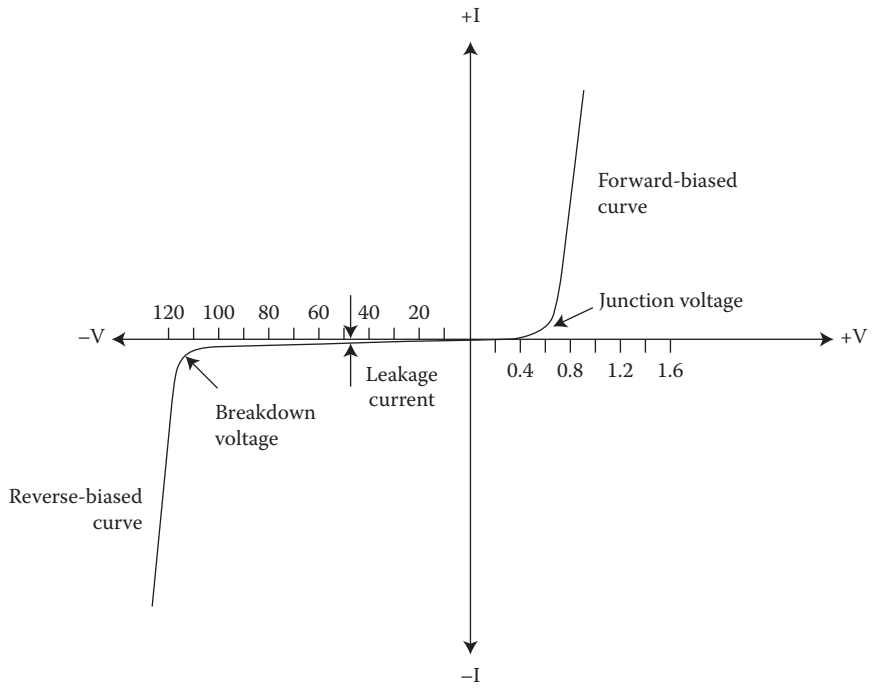


FIGURE 4.39 Typical diode current-voltage curve.

This equation is referred to as the Shockley equation. In it, η is the emission coefficient for the diode, which is determined by the way the diode is constructed and somewhat varies with diode current. For a silicon diode, η is around 2 for low currents and goes down to about 1 at higher currents. The reverse current, I_0, depends on the number of electrons per second that "appear" in the depletion layer, not on the voltage. Only the temperature and the bandgap have an influence on I_0. The simple Shockley equation does not explain the sudden increase of the leakage current at voltages above the breakdown voltage. We will explain this effect further below when discussing Zener and avalanche diodes.

One important use for diodes is to convert alternating current into direct (unidirectional) current in power supply circuits. Diodes are also useful in some logic devices, and we can make some types of digital logic circuits using only diodes and resistors.

Special-Purpose Diodes

Introduction

There are many types of specialized diodes. In this section we briefly introduce light-emitting diodes (LEDs), solar cells, Zener diodes, avalanche diodes, and tunnel or Esaki diodes. In Chapter 5 we discuss double-barrier resonant tunneling diodes (DBRTs) and compare laser diodes with quantum well lasers. The existence of the zero point energy for confined electrons in a quantum well laser (see Chapter 3) makes a quantum well laser much more efficient than a diode laser. In a quantum well, there are no allowed electron states at the very lowest energies. What that means is that the lowest conduction state in a quantum well is broad so that many electrons can be accommodated. Similarly, the top of the valence band has plenty of states available for holes. This means that it is possible for many holes and electrons to combine and produce photons with identical energy for enhanced probability of stimulated emission (lasing) (see Chapter 5). Peltier diodes are described in Volume III, Chapter 8 on actuators.

Light-Emitting Diodes

Electrons and holes can recombine in extrinsic semiconductors that have a substantial number of electrons in the conduction band and many holes in the valence band. On recombination, energy conservation requires that photons and/or phonons be produced. The only place where there are enough electrons and holes to produce bright light is at the p-n junction when a large current is flowing (i.e., in a forward-biased p-n junction). If the bandgap corresponds to the energy of a photon of visible light, then that type of radiation is produced. This is the light-emitting diode (LED), used extensively in electronic devices and cars as indicator lamps, as picture elements in color matrix displays for laptop computers, and so on. Being an indirect bandgap material, Si is currently not used for LEDs, and the most utilized material is direct bandgap GaAs (see Chapter 5). The earliest LEDs produced infrared or visible red light, but LEDs in yellow, green, and blue visible light colors are now available (Figure 4.40). These LEDs convert electrical input to light output. Greater difference in energy levels in the p- and n-sides of the diode produces greater energy change, and consequently higher-frequency, shorter-wavelength light.

A forward-biased p-n junction, showing electrons being injected into the n-type material and holes into the p-type material, is sketched in Figure 4.41. Light is emitted from the narrow depletion zone each time an electron and a hole combine across that zone. The ratio of generated photons to the electrons injected across the junction is called the quantum efficiency, which is a key measure of how well a light emitter is working. For high-performance III–V LEDs, the efficiency is around 10%.

GaAs is a direct energy gap semiconductor with an energy gap of 1.424 eV in the near-infrared portion of the electromagnetic spectrum. This semiconductor is often used in LED construction in combination with AlAs, an indirect energy gap semiconductor with an energy gap of 2.168 eV in the yellow-green portion of the electromagnetic spectrum. As the mole fraction x of $Al_xGa_{1-x}As$ is raised from zero, the energy gap of the resulting compound increases from that of GaAs to that of AlAs. When the value of x is 0.45, the semiconductor switches from a direct energy gap to an indirect energy gap. GaAs and AlAs differ in lattice constants by less than 0.2% at 250°C, which allows for

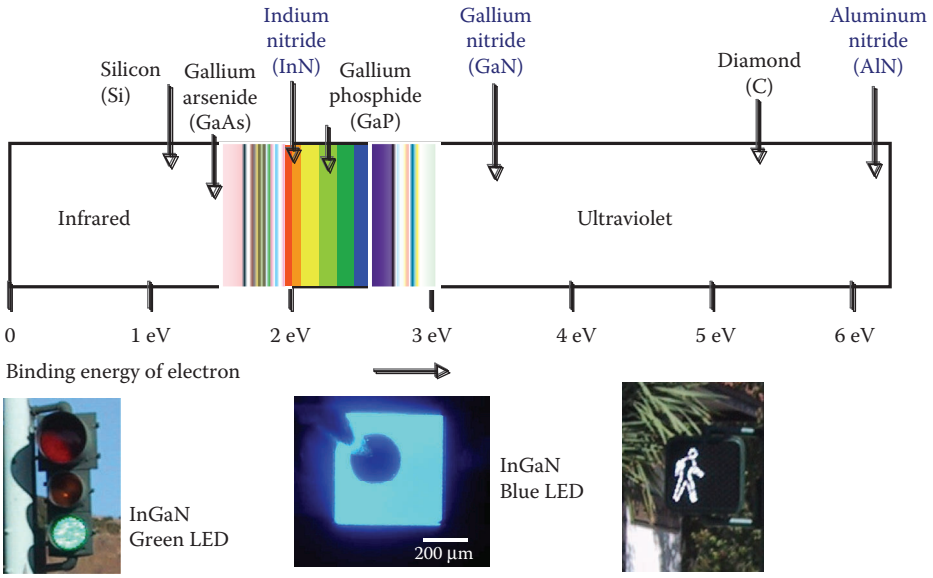


FIGURE 4.40 LEDs of all colors: the larger the difference in energy levels in the p- and n-sides of a diode, the shorter the wavelength of the light.

the growth of very high-quality AlGaAs films on GaAs substrates. Molecular beam epitaxy (MBE) is a crystal-growth technique in which material is deposited by evaporating individually controlled source materials onto a heated substrate in an ultrahigh vacuum chamber. The growth rates in MBE are very slow (0.1–2 mm/h). MBE is capable of producing high-quality AlGaAs heterostructure devices with very uniform thickness and doping characteristics. Crystal growth of LEDs by metalorganic chemical vapor deposition (MOCVD) and liquid-phase epitaxy (LPE) began in the late 1960s. By using MOCVD, it is easier to scale up to larger growth areas than with LPE, and the uniformity and surface morphology of MOCVD grown lasers are superior (MBE, LPE, and MOCVD are compared in Volume II, Chapter 7).

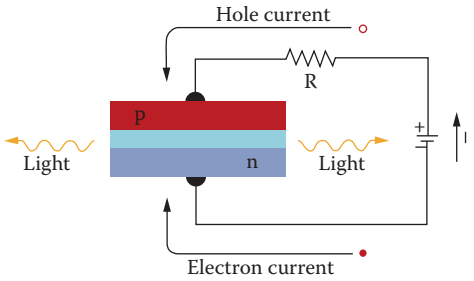


FIGURE 4.41 A light-emitting diode (LED): forward based p-n junction, showing electrons being injected into the n-type material and holes into the p-type material.

The absorption coefficient for direct and indirect bandgaps was calculated using Fermi's golden rule as $\alpha(\omega) \propto G(E) \propto (\hbar\omega - E_g)^{\frac{1}{2}}$ (Equation 3.301) and $\alpha(\omega) \propto G(E) \propto (\hbar\omega - E_g \pm \hbar\nu_p)^2$ (Equation 3.302) (with ν_p the phonon frequency and + for absorption and – for emission), respectively. Emission in an indirect bandgap semiconductor (e.g., Si) is a three-particle interaction (1 electron + 1 photon + 1 phonon); therefore, the transition is unlikely, and such materials are poor light emitters. In silicon, few excited electrons generate photons, and most recombinations result in heat rather than light; thus, the quantum efficiency is poor. In other words, silicon as an indirect bandgap is a lousy light emitter: the probability that a phonon with just the right amount of momentum will meet an electron/hole pair is not very good. The occurrence of such a photon-generating transition in a III–V material is thousands of times more likely than in silicon. As a consequence, until recently, most people in the semiconductor industry would assure you that silicon is pretty much worthless at turning electricity into light, and as we just saw, LEDs and lasers are made of the more exotic III–V semiconductors. But these materials, including gallium arsenide, gallium nitride, and indium phosphide, are anywhere from 30 to 200 times more expensive than silicon. Moreover, the manufacturing infrastructure and know-how about

making ICs from silicon is useless for making chips from III–V semiconductors. To connect the manufacturing of microprocessors with that of LEDs and lasers, one would like to develop cheap, integrated optoelectronics with all components made from silicon.

In Chapter 3, we saw that one approach to getting more light out of silicon involves quantum confinement, with electron/hole pairs physically restricted to a small volume, typically less than 30 nm^3, or 300 times the size of a typical atom. This can be accomplished, for example, by embedding nanocrystals of silicon in an insulating silicon dioxide shell. Within such a nanocrystal, the energy levels of the valence and conduction bands differ significantly from those in bulk crystals, and in general, the smaller the nanocrystal, the larger the bandgap (blue shift). The latter represents an example of tuning a device's optical properties by bandgap engineering, which is achieved in this case by the fine control of the nanocrystal's growth during the manufacturing process. It turns out that quantum confinement also reduces silicon's momentum problem, thus increasing the probability that injected electrons will produce photons. In confined silicon nanostructures, one observes a blue shift for the bandgap absorption onset, and some forbidden transitions become feasible. In other words, through nanotechnology, Si might yet be made into a good candidate for LEDs and lasers after all (another approach to make Si a better photonic material is seeding the silicon with lanthanide rare earth element ions, which give off light when electrically excited).

One important reason to use LEDs is that they last longer, are brighter, and are more efficient than incandescent lights. LED lamps use as little as 10% of the electricity that an incandescent light uses and have lifespans of 100,000 h, which is more than 11 years of continuous use. A laser diode is made by putting an LED-type structure in a resonant cavity formed by polishing the parallel end faces of the semiconductor (see Chapter 5). Photons bounce between the reflective ends, colliding with atoms and stimulating the emission of additional photons that are in phase with the others in the region. In a laser, the concentrated, active region bounded by the reflective ends is known as the resonant cavity. Laser diodes are used in optical storage devices and for high-speed optical communication. As we already alluded to in Chapter 3, they are less efficient than quantum well lasers because more electrons and holes can be accommodated in the latter devices. This means that it is possible for many more holes and electrons to combine and produce photons with identical energy for enhanced probability of stimulated emission (lasing).

Photodiodes and Solar Cells

The photodiode is the inverse of the light-emitting diode (LED). Rather than giving off light it is sensitive to incoming light. Photons of energy E_g or greater create electron/hole pairs in the junction region. The electrons and holes are then swept out of the junction region by the electric field that exists before they can recombine. Photodiodes can be operated in two different modes, namely, the photovoltaic mode and the photoconductive mode. In the photovoltaic mode, like in a solar cell, the illuminated photodiode generates a voltage. However, the dependence of this voltage on the light intensity is rather nonlinear, and the dynamic range is typically small. Although the maximum sensitivity is not achieved, the area of the junction is large, and enough electron/hole pairs are produced so that significant solar energy can be taken from the device. In the photoconductive mode, one applies a reverse voltage to the diode (i.e., a back bias) and measures the resulting photocurrent. The dependence of the photocurrent on the light power can be very linear over six or more orders of magnitude of the light power. This type of diode is very sensitive and is used in CD players.

The electron/hole pairs created by light in a photodiode result in an additional component of current flowing in the diode:

$$I = I_0\left(e^{\frac{eV}{kT}} - 1\right) - I_{optical} \tag{4.51}$$

with the first term the usual diode equation (Equation 4.50), and the second term the photocurrent, $I_{optical}$, which is proportional to the light intensity. Photodiodes are constructed so their p-n junction can be exposed to incoming photons

through a clear window or lens. To increase the volume in which radiation can be absorbed, positive intrinsic negative (PIN) diodes are often used. A PIN diode has a central undoped, or intrinsic, layer, forming a p-type/intrinsic/n-type structure. An internal electric field is maintained over a wider layer with the benefit that the depletion region now exists almost completely within the intrinsic region, which is a constant width (or almost constant) regardless of the reverse bias applied to the diode. This intrinsic region can be made large, increasing the volume where electron/hole pairs can be generated. For these reasons many photodetectors are based on PIN diodes.

In solar cells, photons coming from the sun excite electrons in the valence band and raise them to the conduction band. Illumination of a p-n junction produces hole pairs that diffuse to the depletion region. The minority carriers (electrons in p-type material) are swept across the depletion region and increase the reverse current. In these cells one tries to maximize the minority carrier lifetime, τ, so that long diffusion lengths, $L = (D\tau)^{1/2}$, in excess of 100 µm, can be attained. In a p-type material, electrons are the minority carriers, and the number of electrons generated per unit area per second (generation rate G) in a p-type material within a diffusion length of the depletion layer is given as GL. The larger G and L, the more electrons reach the edge of the depletion layer, where they are collected by sweeping them across the depletion layer, resulting in a current, $I_{optical}$ (see Equation 4.51), which is given by:

$$I_{optical} = eGL \tag{4.52}$$

With a reverse bias applied to the solar cell, Equation 4.52 represents the increase in current with light intensity as illustrated in Figure 4.42b. The current is bias-independent as long as $L >> d$, with d the depletion layer thickness. The depletion layer thickness is the only voltage-dependent parameter.

Under open circuit conditions, no current is measured, and I in Equation 4.50 is equal to zero. On illumination, an open circuit voltage, $V_{optical}$, appears across the n-p junction:

$$V_{optical} = \frac{kT}{e} \ln \frac{I_{optical}}{I_0} \tag{4.53}$$

This open circuit voltage counteracts the flow of optically generated carriers, and the junction potential (V_0; see Equation 4.48) is lowered so that an equal and opposite flow of electrons from n-type to p-type is established.

Research interest in solar cells today, with soaring oil prices, is very high. Solar energy reaching the surface of earth is about 800 W/m^2. The current problem with solar cells is not really the efficiency (20%) but rather the cost of production. Although it has come down by a factor of 100 in 50 years, it still has a factor of 2–5 to go to down to

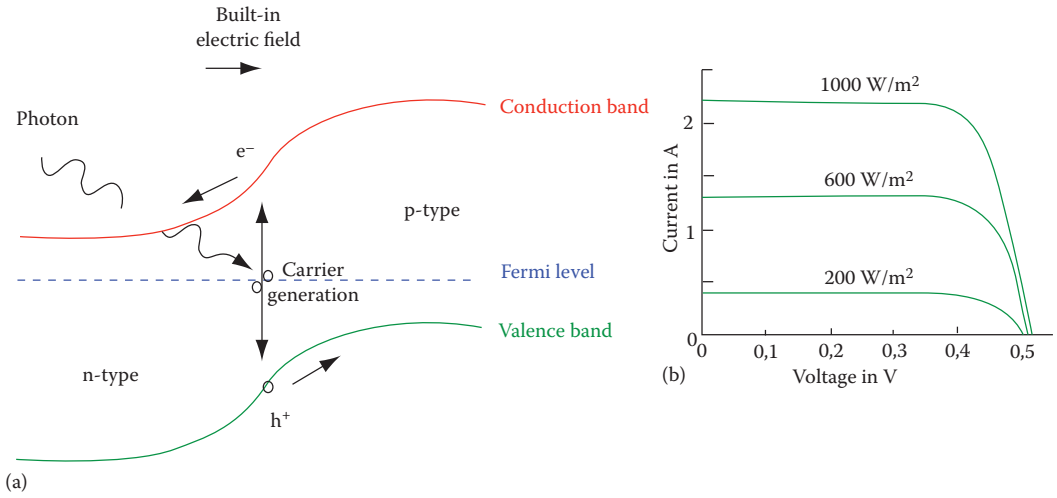


FIGURE 4.42 Illumination of a p-n junction produces hole pairs that diffuse to the depletion region. The minority carriers (electrons in p-type materials) are swept across the depletion region and increase the reverse current (a). The higher the illumination intensity the higher the reverse current (b).

be competitive with fossil plants. Another problem is that no energy can be produced at night, so energy storage is needed to provide continuous service. The highest efficiency solar cells are made of silicon single crystals. These are difficult to manufacture and are expensive. The most popular alternatives are polycrystalline and amorphous Si thin films, as well as GaAs and several of its alloys. The latter are sensitive to different parts of the solar spectrum because they have different bandgaps.

We detail progress in MEMS and NEMS applied to solar cells, including liquid junction-based solar cells (the Gräetzel cell), in Volume III, Chapter 9 on power and brains in miniature devices.

Zener, Avalanche, and Tunnel (Esaki) Diodes

In the breakdown region in the diode current-voltage curve in Figure 4.39, the diode exhibits a sudden increase in current. This is called the Zener breakdown, named for Dr. Clarence Melvin Zener of Southern Illinois University, inventor of the device. The Zener diode works at strong negative bias. If the reverse voltage is high enough to raise the top of the valence band on the p-side above the bottom of the conduction band on the n-side, then valence electrons can directly tunnel into the conduction band, and there is a sudden breakdown of the resistance (Figure 4.43). Zener diodes are used to regulate voltage: they do not allow voltage to go beyond the Zener breakdown point. They are equipped with cooling fins to keep them from melting under high voltage and high current (high power). Zener diodes

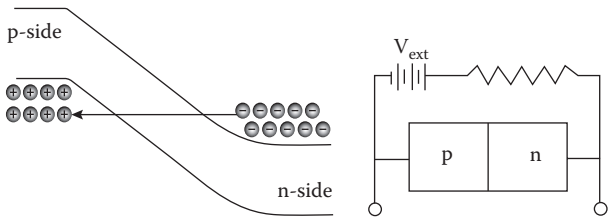


FIGURE 4.43 Band diagram for Zener diode: the top of the valence band on the p-side is above the bottom of the conduction band on the n-side, and valence electrons can directly tunnel into the conduction band, causing a sudden breakdown of the resistance. Also shown is the use of a reverse-biased p-n diode for a voltage clamping circuit. The current saturates to a value determined by the external circuit resistor, while the output voltage is clamped at the diode breakdown voltage.

are mostly used to produce an accurate reference voltage for measurement devices or analog-digital converters, and so on. Manufacturers set a precise breakdown voltage by careful doping, allowing the diode to be used as a precision voltage reference.

Some devices labeled as high-voltage Zener diodes are avalanche diodes. Avalanche diodes conduct in the reverse direction when the reverse-biased voltage exceeds the breakdown voltage. Avalanche diodes are electrically very similar to Zener diodes and are often mistakenly called Zener diodes, but they break down by the avalanche effect. This occurs when the reverse electric field across the p-n junction causes a wave of ionization, reminiscent of an avalanche, leading to a large current. In this region, the high electric field in the middle of the depletion layer accelerates electrons that are produced there (by action of heat or light) so much that they garner enough energy to "dislodge" electrons from the valence band into the conduction band. When one energetic electron can "dislodge" two or more such electrons, it starts a "chain reaction" in which these electrons can produce even more conduction electrons. Avalanche diodes are designed to break down at a well-defined reverse voltage without being destroyed. Avalanche diodes are sensitive to very low radiation because of the "multiplication" of current by the avalanche effect.

The difference between the avalanche diode (which has a reverse breakdown above about 6.2 V) and the Zener diode is that the channel length of the former exceeds the "mean free path" of the electrons, which means there are collisions between them on the way out. The only practical difference is that the two types of diodes have temperature coefficients of opposite polarities.

Yet another special-purpose diode is the tunnel diode (or Esaki diode), which, like the Zener diode, exploits tunneling across a potential barrier. The tunnel diode has the unusual characteristic that the current decreases with the applied bias voltage. Because of this negative resistance region at low voltage and its rapid response, it is capable of amplification, oscillation, and switching in electronic circuits at high frequencies. The Esaki or tunnel diode is designed from very heavily doped n and p regions. The high doping ensures a very large

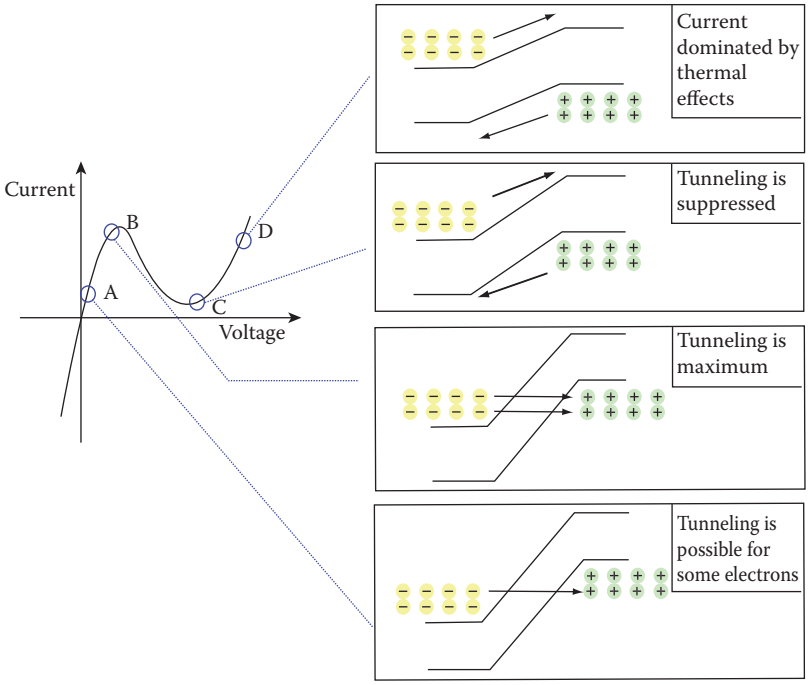


FIGURE 4.44 The tunnel or Esaki diode. For explanation, see text.

electric field in the junction so that band-to-band tunneling is possible at all biasing conditions (see Figure 4.44). The discovery of the Esaki diode was very significant as it was the first quantum mechanical electronic device where electron tunneling was clearly manifested in semiconductors, and it opened up a new field of research: tunneling in solids. Esaki described the tunnel diode in his thesis in 1957 and received the 1972 Nobel Prize for it. Shockley was so impressed by Esaki's work that at a symposium of the International Union of Pure and Applied Physics in Brussels in 1958, he announced (Figure 4.45):

> The most beautiful demonstration of the Zener effect so far achieved is presented at this symposium by L. Esaki of Tokyo. Esaki has studied p-n junctions which are very heavily doped on both sides, so that there is a built-in field of the order of 5×10^5 volt/cm. Under these conditions, current flows by the Zener mechanism even at zero voltage. Esaki finds that as forward bias is increased across the junction, this possibility of direct tunneling disappears and, as a consequence, he observes a predictable negative resistance region.

The tunnel diode couples tunneling with clever energy band alignment to create negative differential resistance as illustrated in Figure 4.44.

With reference to Figure 4.44, the operation of the Esaki diode is as follows. When applying a small forward bias, electrons on the n-side of the diode tunnel into holes on the p-side. As that bias is increased, the band alignment enables yet more electrons to find holes to tunnel into, and the current increases (A). Beyond the point (B), empty states on the p-side move higher in energy with respect to the electron energies on the n-side. Eventually, at point C, there are no empty states (holes) available

FIGURE 4.45 W. Shockley (age 48 years) and L. Esaki (age 33 years) in 1958 in Brussels for the symposium of the International Union of Pure and Applied Physics. This coincided with the World Expo held there.

for electrons to tunnel into. This causes the current to decrease to a minimum value. Eventually, the current increases again because the electrons and holes are able to overcome the reduced barrier because of their thermal distribution.

Transistors

Introduction

An integrated circuit (IC) is an electronic network fabricated in a single piece of a semiconductor material. In IC fabrication, the semiconductor surface is subjected to various processing steps in which impurities and other materials are added with specific geometrical patterns. The fabrication steps are sequenced to form 3D regions that act as transistors and interconnects that form the switching or amplification network. In this section, we detail different types of transistors.

Transistors are three-terminal devices, where the current or voltage at one terminal, the input terminal, controls the flow of current between the two remaining terminals. In Chapter 1 we saw how the name "transistor," a name coined by John R. Pierce, a scientist from Bell Laboratories, is a contraction of the terms *trans* and *resistor*. This term was adopted because it best describes the operation of the transistor—the transfer of an input signal current from a low-resistance circuit to a high-resistance circuit. Basically, the transistor is a solid-state device that amplifies by controlling the flow of current carriers through its semiconductor materials. Electronic devices made up of transistors are more interesting and have more applications than diodes.

Two types of transistors are widely used: the bipolar junction transistor (BJT), which is physically like two junction diodes back to back, and the field-effect transistor (FET), in which one uses the voltage on a control (gate) electrode to control the shape and hence the conductivity of a channel of one type of charge carrier in a semiconductor material. Only weeks after the point contact transistor was conceived by Shockley, Bardeen, and Brittain (see Chapter 1), Shockley came up with the BJT in 1948. Bardeen was the physicist; Walter Brittain, the experimentalist; and William Shockley, who became involved later in the development, was the instigator and idea man. The team won the 1956 Nobel Prize in physics for their efforts. The transistor demonstrated for the first time that amplification in solids was possible. The device was initially known simply as the junction transistor, and it did not become practical until the early 1950s. The term *bipolar* was tagged onto the name to clarify the fact that both carrier types play important roles in its operation. Historically the eldest, it was in use for many years for large integrated devices and high speed at the expense of low density and higher power consumption. The dominance of BJTs through the 1960s and 1970s was quickly overwhelmed by the simplicity of the metal–oxide–semiconductor field-effect transistor (MOSFET) (in terms of both circuit design and fabrication processes) in the 1980s. This is especially true in the very large scale integration (VLSI) era. In a MOSFET, power consumption is very low, and the current flow depends on the majority carriers only; thus, it is unipolar. The BJT's high current drive capability and superior analog performance continue to drive applications, however, and a recent trend is to combine the best of MOSFETs and bipolar devices in bipolar CMOS (BiCMOS) processes. BiCMOS is a type of IC that uses both bipolar and CMOS technologies. The logic gates are primarily made of CMOS, but their output stages use BJTs. Bipolar transistors consume more current (and therefore dissipate more power) but switch faster (see Table 4.9).

Bipolar Transistors

The BJT structure contains two p-n diodes: one between the base and the emitter, and one between the base and the collector. There are many different types of transistors, but the basic theory of their operation is all the same. The three elements of the two-junction transistor are 1) the EMITTER, which gives off, or *emits*, current carriers (electrons or holes); 2) the BASE, which controls the flow of current carriers; and 3) the COLLECTOR, which collects the current carriers (Figure 4.46). Early BJTs were fabricated using alloying—a complicated and unreliable process (Figure 4.46a). Figure 4.46a, also shows the symbol used for a BJT transistor. The "planar process," developed by Fairchild in the late 1950s, shaped the basic structure of the BJT, which is still seen even up to the present day. In the planar

TABLE 4.9 Bipolar, CMOS, and BiCMOS Compared

Bipolar	CMOS	BiCMOS
Better analog capability		High-performance analog
High power dissipation	Lower static power dissipation	Lower power dissipation than purely bipolar technology (simplifying packaging and board requirements)
Generally better noise performance and better high-frequency characteristics	Higher noise margins	
Improved I/O speed (particularly significant with the growing importance of package limitations in high-speed systems)		Flexible I/Os technology is well suited for I/O intensive applications
Low packing density	Higher packing density	
	Lower manufacturing cost per device	
Lower input impedance (high drive current)	High input impedance (low drive current)	
Low delay sensitivity to load	High delay sensitivity to load (fan-out limitations)	
Low voltage swing logic	Scaleable threshold voltage V_T	
Higher current drive per unit area, higher gain	Low output drive current (issue when driving large capacitive loads)	
High transconductance g	Low transconductance g	
Essentially unidirectional	Bidirectional capability (drain and source are interchangeable)	
Higher switching speed	A near ideal switching device	Improved speed over purely CMOS technology

process, all steps are performed from the surface of the wafer (Figure 4.46b).

A forward-biased p-n junction is comparable to a low-resistance circuit element because it passes a high current at a given voltage. In turn, a reverse-biased p-n junction is comparable to a high-resistance circuit element. By using the Ohm's law formula for power ($P = I^2R$) and assuming the current is held constant, it can be concluded that the power developed across a high resistance is greater than that developed across a low resistance. Thus, if a crystal were to contain two p-n junctions (one

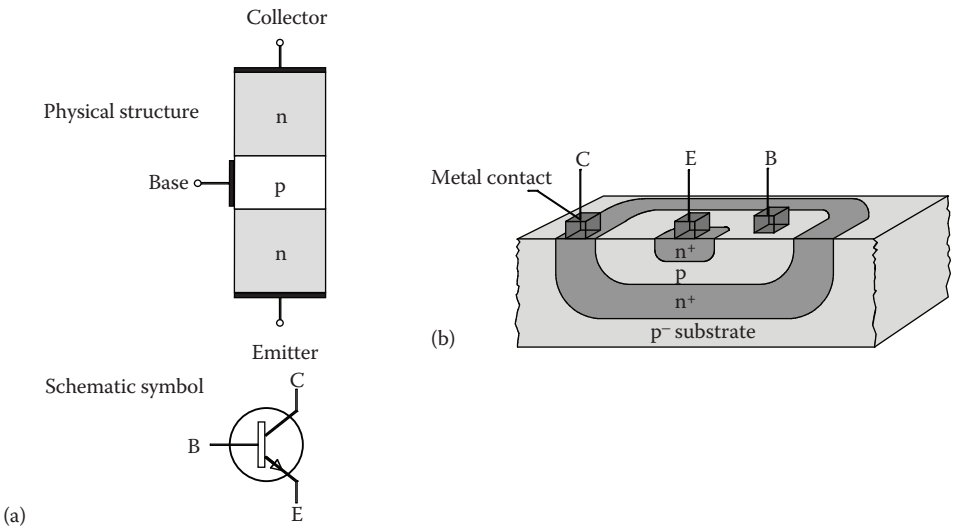


FIGURE 4.46 Bipolar npn transistor made by alloying a BJT symbol (a) and a BJT made in a planar process (b).

forward-biased and the other reverse-biased), a low-power signal could be injected into the forward-biased junction and produce a high-power signal at the reverse-biased junction. In this manner, a power gain would be obtained across the crystal. This concept is the basic theory behind transistor amplification.

MOSFETS

The most common field-effect transistor (FET) in both digital and analog circuits is the MOSFET. Nomenclature is different here than for a bipolar transistor with the base replaced by a gate (*G*), the emitter by a source (*S*), and the collector by a drain (*D*) (see Figure 4.47). The physical structure shown here consists of a p-type Si substrate, n^+ source, n^+ drain, a gate oxide (SiO_2), a polysilicon gate, a chemical vapor deposition (CVD) oxide, and a contact metal 1 (usually Al). Notice that the effective gate length is shorter than the one drawn in the mask layout ($L_{eff} < L_{drawn}$) as a result of lateral doping effects.

Silicon is the main choice of semiconductor used in MOSFET manufacture, but SiGe is also used more and more by some manufacturers (see page 269). Some other semiconductors such as GaAs are not useful in MOSFET manufacture because they do not form good insulating gate oxides. A thin oxide layer between the gate (e.g., a PolySi or so-called *poly gate*) and the semiconductor characterizes the MOSFET. This high resistivity oxide layer ($\rho > 10^{16}$ Ωcm) prevents current flow from the semiconductor to the gate. Silicon-based MOSFETs rely on the extremely high-quality interface between SiO_2, the standard gate dielectric, and silicon.

To understand a MOSFET, it is essential to recall the metal oxide semiconductor (MOS) capacitor fundamentals. In an idealized MOS capacitor, without any applied voltage, the metal work function, ϕ_m, is equal to the semiconductor work function, ϕ_s. Therefore, the Fermi level of the semiconductor, E_{FS}, is aligned with the Fermi level of the gate, E_{Fm}. As a consequence, there is no band bending in any region of the MOS capacitor. We also assume here that the oxide is free of any charges, and the semiconductor is uniformly doped. In the example MOS structure in Figure 4.48a, we use a p-type semiconductor with a negative gate bias applied. With this negative gate bias (V_{GB}), the gate charge, Q_G, is negative. Because charge neutrality must be preserved, a net positive charge, Q_C, is generated in the silicon substrate to counterbalance the negative charge on the gate. This is achieved by an accumulation of majority carrier holes under the gate (because we are dealing with a p-type semiconductor, the majority carriers are holes). The condition, where the majority carrier concentration is greater near the Si-SiO_2 interface compared with the bulk concentration, is called *accumulation*. Under this applied negative gate bias, the Fermi level of the gate is raised with respect to the Fermi level of the substrate by an amount qV_{GB}.

The energy bands in the p-type semiconductor are bent upward, bringing the valence band closer to the Fermi level, which is indicative of a higher hole concentration at the semiconductor surface right under the dielectric. Because no current flows through the device due to the presence of the insulating oxide layer, the Fermi level in the substrate

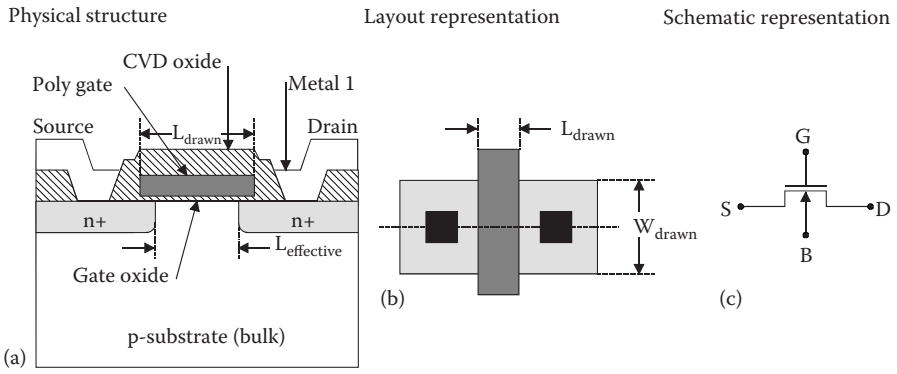
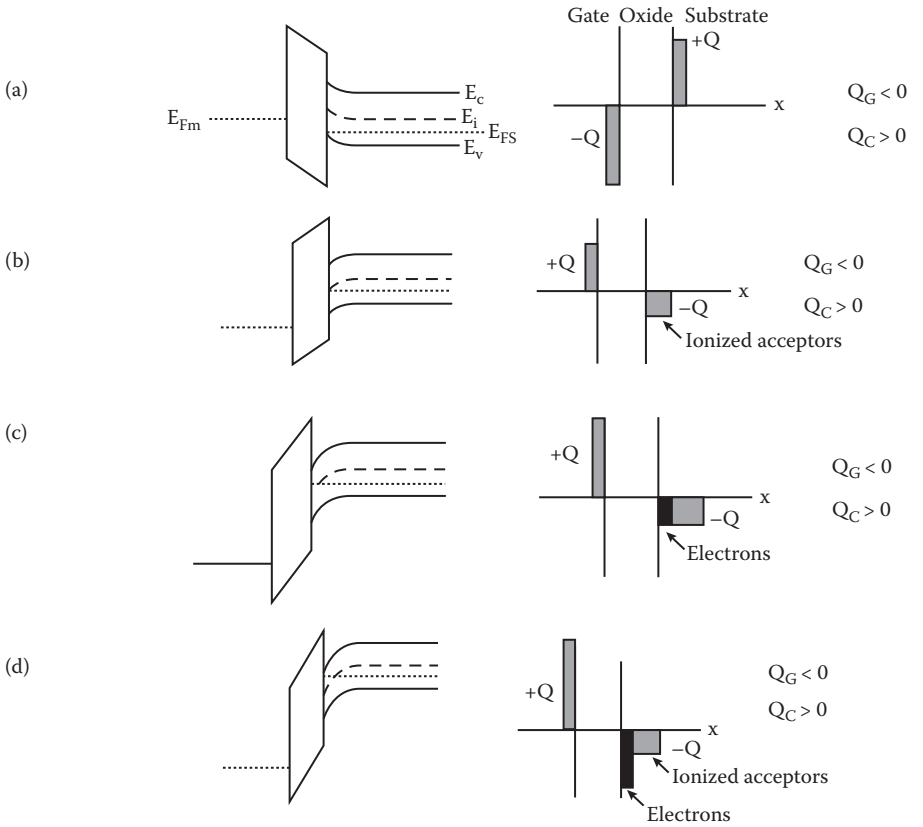


FIGURE 4.47 A MOSFET with nomenclature (a). The geometry of the areas on the Si surface is known as the layout of the chip (b). Schematic representation of a MOSFET (c).



Accumulation	$\psi_s < 0$	$V_{GB} < V_{FB}$	$Q_C > 0$
Depletion	$\psi_s > 0$	$V_{GB} > V_{FB}$	$Q_C < 0$
Inversion	$\psi_s > 0$	$V_{GB} >> V_{FB}$	$Q_C < 0$

FIGURE 4.48 A MOS capacitor with a metal gate, an insulator, and a p-type semiconductor. (a) Accumulation; (b) depletion; (c) inversion; (d) strong inversion. E_i is the Fermi level for the intrinsic material, and E_{FS} is that of the doped material (because it is a p-type material the Fermi level is close to the valence band).

(the semiconductor) remains invariant even with an applied bias. The applied gate voltage (V_{GB}) divides between the gate dielectric and the semiconductor such that:

$$V_{GB} = V_{ox} + V_s \tag{4.54}$$

where V_{ox} and V_s are the voltages that drop across the oxide and the semiconductor, respectively. The bands in the oxide drop linearly over the bulk of the oxide thickness an amount equal to qV_{ox}, and the electric field in the oxide can be expressed as:

$$E_{ox} = V_{ox}/t_{ox} \tag{4.55}$$

where t_{ox} is the oxide thickness. The amount of band bending in the semiconductor is principally located at the surface and is equal to $q\psi_\sigma$, where ψ_σ is the surface potential and is negative when the band bend is upward.

When applying a positive gate bias V_{GB} to counterbalance the positive gate charge, the holes under the gate are pushed away, leaving behind ionized, negatively charged acceptor atoms (N_A^-), which create a depletion region (Figure 4.48b). The charge in the depletion region of width W and area A is exactly equal to the charge on the gate to preserve charge neutrality [$Q_G(+) = Q_c(-) = -qN_A^-AW$]. With a positive gate bias, the Fermi level of the gate is now lowered with respect to the Fermi level of the semiconductor. The bands bend downward, resulting in a positive surface potential. Under the gate, the valence band moves away from the Fermi level, which is indicative of hole depletion. When the band bending at the

surface is such that the intrinsic level (E_i) coincides with the doped semiconductor Fermi level (E_{FS}), the surface region resembles an intrinsic material. The surface potential for this condition is given by:

$$\psi_s = \phi_F = \frac{1}{q}(E_i - E_F) \tag{4.56}$$

with ϕ_F, the Fermi potential, given by:

$$\phi_F = \frac{kT}{q}\ln\frac{N_A}{n_i} \tag{4.57}$$

where n_i is the charge carrier density of the intrinsic material (see Equation 4.17: $n_i \approx 1.5 \times 10^{10}$ cm^{-3} at 300 K in pure Si). Applying a larger positive gate bias, the positive charge on the gate increases further, and thermally generated electrons start collecting under the gate. With these electrons, the intrinsic surface region begins to convert into an n-type inversion layer. The negative charge in the semiconductor is now composed of both ionized acceptor atoms in the depletion layer and free electrons in the inversion layer (Figure 4.48c). In this condition, the electron concentration at the surface is still less than the hole concentration in the neutral bulk of the semiconductor, and this regime is called weak inversion. The point where $\psi_s = \phi_F$ is defined as the onset of weak inversion.

As the gate bias is further increased, the band bending also continues to increase. This causes the depletion region to become wider while the electron concentration in the inversion layer increases more and more. At the gate bias where the electron concentration in the inversion region is equal to the hole concentration in the bulk, a strong inversion layer is created (Figure 4.48d). The surface potential to achieve this strong inversion is given by:

$$\psi_s = 2\phi_F = \frac{2}{q}(E_i - E_F) \tag{4.58}$$

When the gate bias is increased beyond this value, new electrons in the inversion layer readily compensate extra positive charge on the gate and the depletion layer width does not further increase, and hence the band bending increases only very slightly. For simplicity, $2\phi_F$ is generally taken as the maximum value for ψ_s. The maximum depletion layer width, W_d, at the onset of strong inversion is then given by:

$$W_d = \sqrt{\frac{4\phi_F\varepsilon_s}{qN_A}} \tag{4.59}$$

In a real MOS device, the work function of the metal, ϕ_m, is not likely to be equal to that of the semiconductor, ϕ_s. Consider a MOS capacitor with $\phi_m < \phi_s$ at zero gate bias. In this case, electrons in the metal reside at energy levels above the conduction band of the semiconductor. Thus, electrons will flow from the metal into the semiconductor until a potential counterbalances the difference in the work functions between the two plates of the MOS capacitor. This induces a negative charge under the gate dielectric, accompanied by a downward band bending and hence a preexisting positive surface potential ψ_{ms}. With an external voltage, equal to this difference ψ_{ms}, applied to the gate, the net charge in the semiconductor disappears, and the bands return to their flat position. This applied voltage is defined as the flatband voltage, V_{FB}, and is expressed as:

$$V_{FB} = q\psi_{ms} = q(\phi_s - \phi_s) \tag{4.60}$$

The oxide and the semiconductor/oxide interface are not perfect, and another modification to the ideal picture of a MOS capacitor comes from charges that reside in the dielectric or at its interface. Such charges may originate from processing, defects in the bulk, or charges that exist at the interface between Si and SiO_2 (for an overview of the type of oxide defects, see Figure 4.35). Charges in the bulk oxide, or at the interface, induce opposite charges in the semiconductor and the gate. This again causes the bands to bend up or down, and the flatband voltage has to be adjusted to take this into account. The flatband voltage can then be expressed as:

$$V_{FB} = q\phi_{ms} - \frac{Q_o}{C_{ox}} \tag{4.61}$$

where Q_o is the oxide charge, and C_{ox} is the oxide capacitance (we also have introduced $\psi_{ms} = \phi_{ms}$). Defects located at the Si/SiO_2 interface may not be fixed in their charged state, and the amount of charges may also vary with the surface potential of

the semiconductor. These defects are referred to as fast interface traps, and they impact the switching characteristics of MOSFETs.

We now put a MOS capacitor between a source (*S*) and a drain (*D*). In Figure 4.49A we show an n-channel and a p-channel device. Both n-channel and p-channel MOSFETs are extensively used. In fact, CMOS IC technology relies on the ability to use both devices on the same chip. As shown in Figure 4.49B, to create an nMOSFET and a pMOSFET on the same p-type substrate, the pMOSFET must be provided with n-type well regions. N-wells are formed by ion implantation or by diffusion. Lateral diffusion limits the proximity between structures, and ion implantation results in shallower wells compatible with today's fine-line processes. Metal 1 makes contact to the nMOSFET and metal 2 to the pMOSFET. Metals in vias connecting to transistors are often made of tungsten metal plugs, and such plugs

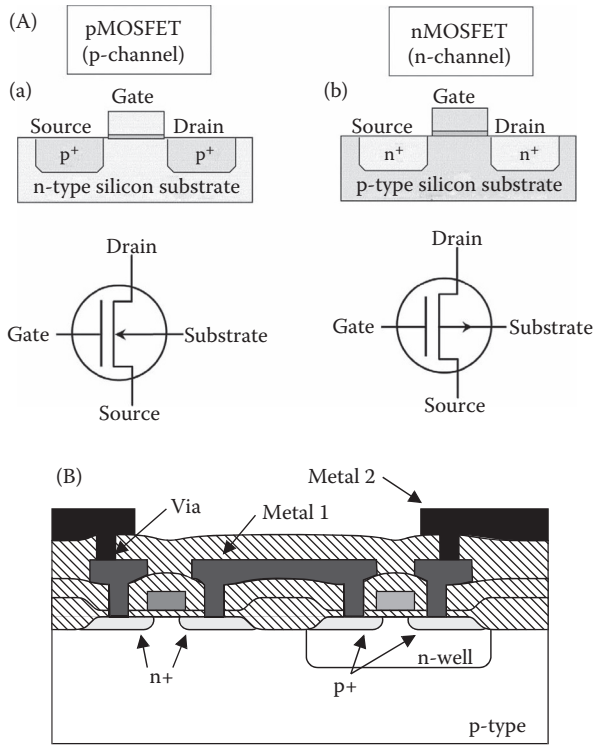


FIGURE 4.49 (A) pMOS and nMOS with symbols. (a) pMOS with p-channel: a negative voltage is applied to the gate, thus attracting holes that form a channel between the two p^+ source and drain regions. (b) nMOS with n-channel: a positive voltage is applied to the gate, thus attracting electrons that form a channel between the two n^+ source and drain regions. (B) To create an nMOSFET and a pMOSFET on the same p-type substrate, the pMOSFET must be provided with an n-well.

are also used to fill vias between successive layers of metal (Cu or Al) (see Figure 4.62 and Volume III, Figure 4.3). Tungsten is used because of the extraordinarily good conformality of CVD from WF_6. The tungsten in turn contacts the Si via an adhesion/barrier layer such as Ti/TiN. The Ti layer improves the *W* contact to the transistor and protects the underlying Si from attack by fluorine.

Depending on the type of MOSFET (i.e., p-channel or n-channel), two different situations arise. The voltage applied between the gate and the substrate (V_{GS}) generates an electric field through the oxide layer, creating an inversion channel under the gate oxide. In the case of a pMOSFET (with a p-channel) transistor, the n-type substrate is inverted to p-type in a thin layer under the gate oxide, and a p-type substrate is similarly inverted to n-type in an nMOSFET (with an n-channel) device. Because the inversion channel is of the same type as the source and drain, current can now pass from source to drain. The pMOS transistors are 2.5 times slower than nMOSFETs because of the mobility difference between electrons and holes.

By varying the potential between the gate and substrate, the inversion channel in which current flows can be altered to control the magnitude of the current, or be turned off completely. Thus, the magnitude of the drain source current, I_D, through the channel is controlled by the potential difference V_{GS} applied between the source *S* and the gate *G*. The threshold voltage, V_T, is the value of V_{GS} where the drain current just begins to flow. From Figure 4.50 typical values range from 0.3 to 0.8 V.

The drain current (I_D) versus the voltage between drain and source (V_{DS}) for different values of V_{GS} is illustrated in Figure 4.51. As shown at low V_{DS}, the drain current increases almost linearly with V_{DS}, resulting in a series of straight lines with slopes increasing with V_{GS}. At high V_{DS}, the drain current saturates and becomes independent of V_{DS}.

Using the nMOSFET in Figure 4.52, the curves in Figures 4.50 and 4.51 are explained in a bit more detail. To make an n-channel MOSFET, as shown in Figure 4.52a, we diffuse donors in the source, drain regions (forming n-regions) in an p-type substrate, and grow an oxide layer that acts as the gate oxide. With a small positive voltage on the drain and no

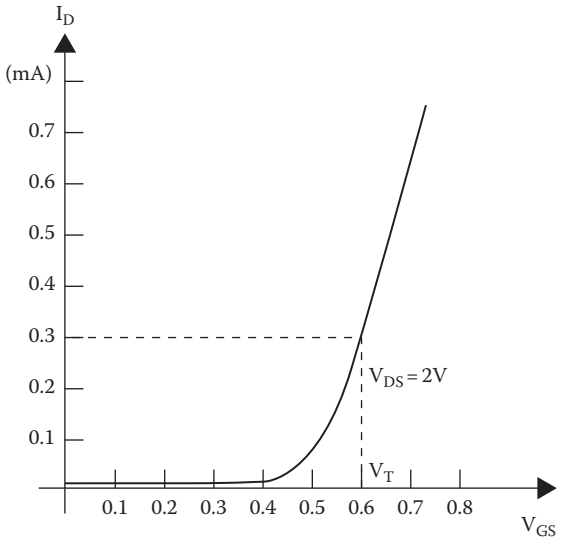


FIGURE 4.50 The value of V_{GS} where the drain current, I_D, just starts flowing is called V_T or the threshold voltage.

bias on the gate, i.e., $V_{DS} > 0$ and $V_{GS} = 0$, the drain is a reverse-biased p-n junction. Conduction band electrons in the source region encounter a potential barrier determined by the built-in potential of the source junction. As a result, no electrons enter the channel region, and hence no current flows from the source to the drain. This is referred to as the "off" state and is illustrated in Figure 4.52b.

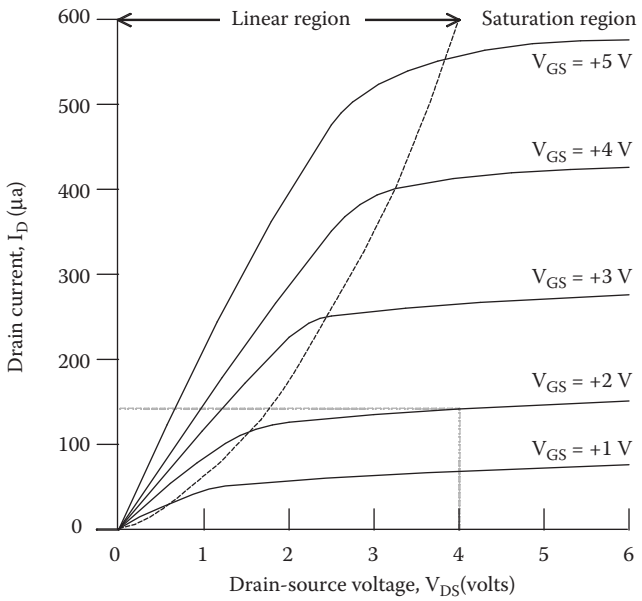


FIGURE 4.51 The drain current, I_D, versus the voltage between drain and source, V_{DS}, for different values of V_{GS}. The boundary between linear and saturation (active) modes is indicated by the upward-curving parabola.

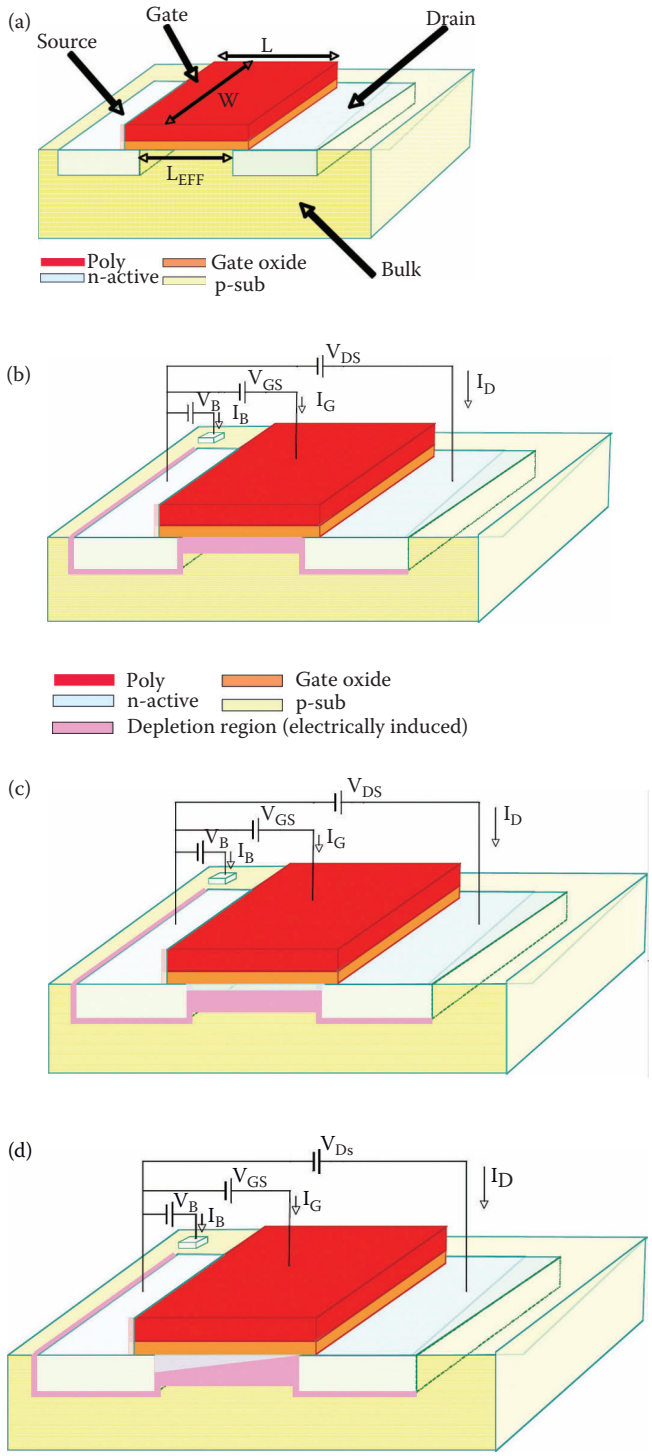


FIGURE 4.52 (a) nMOSFET. (b) Apply small V_{GS} (V_{DS} and V_B assumed to be small). Depletion region electrically induced in channel, termed "cutoff" region of operation. (c) At a critical value of V_{GS}, termed the threshold voltage (V_T), an inversion layer is created. The inversion layer supports current flow from *D* to *S*. The channel behaves as a thin-film resistor. (d) With $|V_{DS,\,max}| > |V_{GS} - V_T|$, the channel is flowing as much current as it can, so we call it pinched off (even though current continues to flow). The semiconductor bulk potential is V_B.

With a small positive bias on the gate, band bending in the channel region ($\psi_s > 0$) brings the conduction band in the channel region closer to the conduction band in the source region, reducing the height of the potential barrier to electrons. At a critical value of V_{GS}, termed the threshold voltage (V_T), an inversion layer is created (see Figure 4.52c). The inversion layer supports current flow from *D* to *S*, and electrons can now enter the channel and flow from source to drain; we say that the transistor is on. In the low drain bias regime, the drain current increases almost linearly with drain bias, and like an ideal resistor it obeys Ohm's law (see Figure 4.51).

The channel resistance is determined by the electron concentration in the channel, which is a function of the gate bias. As the gate bias is increased, the slopes of the linear portion of the I–V characteristics in Figure 4.51 gradually increase as a result of the increasing conductivity of the channel. This region, where the channel behaves like a resistor, is called the *linear region of operation*. The drain current I_D is given by:

$$I_D = \frac{1}{2}\mu_e C_{ox}\frac{W}{L}[2(V_{GS} - V_T)V_{DS} - V_{DS}^2] \qquad (4.62)$$

The dimensions *W*, *L* can be identified from Figure 4.52a, and μ_e is the effective channel mobility of the electrons, which differs from the bulk electron mobility. For small V_{DS} ($V_{DS} << V_{GS} - V_T$), the second term in parentheses can be ignored, and the expression for drain current reduces to a straight line with a slope equal to the gain factor of the device (see also below). We shall deal with the concept of effective channel mobility later. In transistors of the 1980s the V_{GS} was 5 V; in modern processes this has come down to 3.3 V, then 2.5 V, 1.8 V, 1.5 V, 1.2 V, and nowadays 1 V. The higher voltages used in the past would cause breakdown in the latest transistors. Importantly, lower V_{GS} saves power. The threshold voltage, V_T, of a MOSFET is the gate voltage, V_{GS}, required to create strong inversion (i.e., $\psi_s = 2\phi_F$; Equation 4.58) under the gate. When the semiconductor (body) is at ground potential ($V_B = 0$), the threshold voltage, V_T, equals the sum of the flatband voltage V_{FB} (Equation 4.61), the voltage across the oxide due to the depletion layer charge ($Q_d = -qN_AW_d$ with W_d given by Equation 4.59), and twice the bulk potential π_F (Equation 4.58), or:

$$V_T = q\phi_{ms} - \frac{Q_o}{C_{ox}} - \frac{Q_d}{C_{ox}} + 2\phi_F \quad \text{(nMOSFET) with}$$

$$\phi_F = \frac{kT}{q}\ln\frac{N_A}{n_i}$$

$$V_T = q\phi_{ms} - \frac{Q_o}{C_{ox}} - \frac{Q_d}{C_{ox}} - 2\phi_F \quad \text{(pMOSFET) with}$$

$$\phi_F = \frac{kT}{q}\ln\frac{N_D}{n_i}$$

and with

$$Q_d(\text{nMOSFET}) = -2\sqrt{qN_A\phi_F\varepsilon_s}$$

$$Q_d(\text{pMOSFET}) = 2\sqrt{qN_D\phi_F\varepsilon_s} \qquad (4.63)$$

Control of the threshold voltage is crucial in making MOSFETs. By using highly doped polysilicon as the gate material instead of Al, ϕ_{ms} is now just the difference in Fermi levels of the two silicon regions. Polysilicon is also more process friendly as it can withstand higher temperatures than Al. In Equation 4.63, C_{ox} (F/m^2) is given by:

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \qquad (4.64)$$

where t_{ox} is the oxide thickness. A larger C_{ox} leads to a smaller threshold. We would like a small V_T under the gate, but elsewhere we would like a large V_T to prevent channels from forming between transistors. To grow the most perfect SiO_2 layer, [100]-oriented wafers are used (less dangling bonds), slower growth rates in dry oxygen are used (leads to higher quality layer), and HCl is often added to the oxygen steam to reduce sodium incorporation in the SiO_2. In a 0.24-μm MOSFET process, t_{ox} is 5 nm (~10 atom layers), and C_{ox} is 5.6 fF/μm^2. With even thinner oxide films, say 2 nm, large tunneling currents lead to leakage and high power consumption. One solution is to look at dielectrics that feature a higher dielectric constant ε_{ox}, such as ZrO_2, HfO_2, and Al_2O_3. Finally, there is a substrate bias effect. By adding a contact to the semiconductor substrate (normally the source and body are tied together to

ground), another means of controlling the threshold voltage through changing V_B results (making the transistor a four-terminal device: gate, source, drain, and substrate body):

$$\Delta V_T = \frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}\left[(-V_B)^{\frac{1}{2}}\right](\text{nMOS})$$
$$\Delta V_T = \frac{\sqrt{2\varepsilon_s q N_D}}{C_{ox}}\left[(V_B)^{\frac{1}{2}}\right](\text{pMOS}) \quad (4.65)$$

where $\frac{\sqrt{2\varepsilon_s q N_D}}{C_{ox}}$ (for nMOS) and the same expression with N_D for pMOS are the body-effect coefficients (impact of changes in V_B) ($\varepsilon_s = 1.053 \times 10^{-10}$ F/m is the permittivity of silicon; $C_{ox} = \varepsilon_{ox}/t_{ox}$ is the gate oxide capacitance with $\varepsilon_{ox} = 3.5 \times 10^{-11}$ F/m). In the case of an n-channel device, one requires a gate voltage more positive than V_T to create an electron channel. For a p-channel, one requires a gate voltage more negative than V_T to create an electron channel.

As we saw earlier for small V_{DS}, the second term in parentheses in Equation 4.62 can be ignored, and the expression for drain current reduces to:

$$I_D = \mu_e C_{ox} \frac{W}{L}\left(V_{GS} - V_T\right) V_{DS} \quad (4.66)$$

which is a straight line with a slope equal to the gain factor of the device:

$$g = \mu_e C_{ox} \frac{W}{L} \quad (4.67)$$

and $\mu_e C_{ox}$ is the device's transconductance g'. When the drain voltage is made comparable to the gate bias voltage, the drain current increases sublinearly as predicted by Equation 4.62. This is due to a reduction of the inversion layer at the drain end of the channel. We delve into this reduction of the inversion layer length a bit deeper now.

For these larger drain biases, the drain current saturates and becomes independent of the drain bias. Naturally, this region is referred to as the saturation region. The drain current saturation is derived from Equation 4.62, which is parabolic with a maximum occurring at $|V_{DS,\,max}| = |V_{GS} - V_T|$, where $V_{DS,max}$ is also called the *pinch-off voltage*. To calculate the drain current, I_D, in this saturation region we substitute $|V_{DS,\,max}| = |V_{GS} - V_T|$ in Equation 4.62 to obtain:

$$I_D = \mu_e C_{ox} \frac{W}{L} V_{DS,max}^2 \quad (4.68)$$

When $|V_{DS,\,max}| = |V_{GS} - V_T|$, the channel is flowing as much current as it can, and the inversion layer now starts changing shape. Let us analyze why the inversion layer changes shape and why it is eventually pinched off. This is illustrated in Figures 4.52d and in more detail in Figure 4.53. The potential drop across the gate oxide determines the amount of positive charge on the gate in the nMOS structures of Figures 4.52 and 4.53. When the current flowing through the channel increases, a finite voltage drop occurs along it since it comes with its own finite resistance. This voltage opposes the applied gate bias, and as a consequence the amount of charges in the inversion layer becomes smaller at the drain end compared with the charge induced at the source end of the channel. This is illustrated in Figures 4.52d

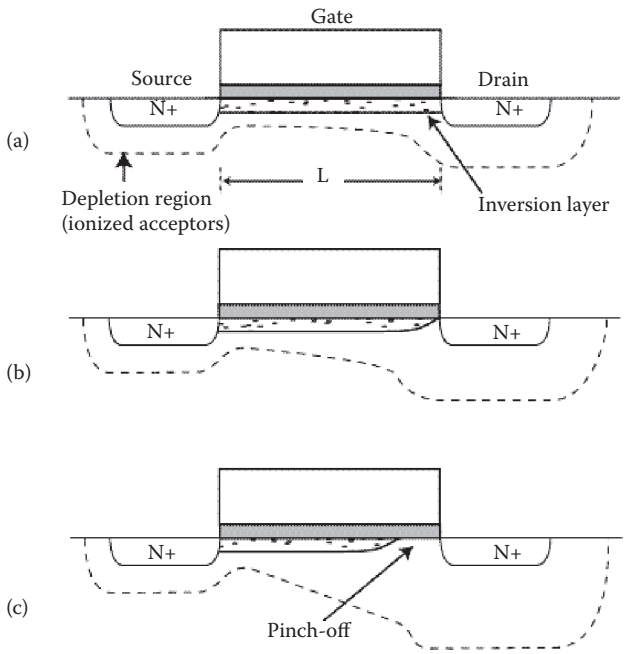


FIGURE 4.53 The pinch-off process in an nMOSFET. (a) The drain voltage is small compared with the gate bias voltage, and the charge in the inversion layer is uniform throughout the channel. (b) Increasing the drain voltage further reduces the inversion layer thickness and increases the depletion width around the reverse-biased drain junction. (c) Eventually the inversion layer completely disappears near the drain or the channel is pinched off.

and 4.53b by the smaller thickness for the channel inversion region close to the drain. Increasing the voltage on the drain also increases the depletion width around the reverse-biased drain junction, freeing more negative acceptor ions so that a fewer number of inversion layer electrons are needed to compensate the positive charges on the gate (see Figure 4.53b).

This means that the electron density in the inversion layer near the drain would decrease even if the charge density on the gate were constant. The reduction in number of carriers causes a reduction in the channel conductance as reflected in the smaller slope of I_D–V_{DS} curves as V_{DS} approaches $V_{D,sat}$ when the MOSFET enters the saturation region. Eventually, the inversion layer completely disappears near the drain. At this point, the induced charge is zero, and the conducting channel disappears or is pinched off (no channel exists in the vicinity of the drain region). In other words at a certain point there is not enough voltage difference between the gate and the semiconductor to cause a strong inversion layer at the drain end of the channel. This is shown in Figure 4.53c as a channel with zero thickness on the drain end, and this is referred to as the *channel pinch-off*.

The voltage difference over the induced channel remains fixed, and the current remains constant (or saturates). As V_{DS} is increased beyond $V_{D,sat}$, the width of the pinched-off region increases. However, the voltage that drops across the inversion layer remains constant and equal to $V_{D,sat}$. The portion of the drain bias in excess of $V_{D,sat}$ appears across the pinched-off region. In a long-channel MOSFET, the width of the pinched-off region is assumed to be small relative to the length of the channel. The applied voltage V_{DS} modulates the effective length of the conductive channel, so:

$$I_D = I'_D\,(1 + \lambda V_{DS}) \qquad (4.69)$$

where λ is the channel-length modulation (varies with the inverse of the channel length). In shorter transistors, the drain-junction depletion region presents a larger fraction of the channel, and the channel-length modulation effect is more pronounced.

In smaller devices the depletion width is similar in size to the gate length, and this affects the transistors current-voltage (I-V) curve. At short gate lengths, the electrostatic field no longer resembles that of a planar capacitor, and the FET suffers from threshold voltage shift, lack of pinch-off, increased leakage current, and increased output conductance. A linear I-V relation in a short-channel device replaces the quadratic dependence of I_D on V_{GS} in the saturation region of a long-channel device. As channel length decreases, space-charge regions at the source and drain encroach on areas normally controlled by the gate. In sufficiently small devices, the depletion regions of the two junctions can merge; this is known as punch-through. In punch-through the current is not controlled by the gate voltage but by V_{DS}^2 / L^3. Short-channel effect decreases the gate's ability to control the depletion region and allows current to flow as it normally would with a long-channel length. To defeat the short-channel effect, one must scale the device vertically, i.e., scale the lateral dimensions, but also the dimensions vertical to the wafer surface. By scaling all dimensions, the original field distribution can be recreated (see also below under "Constant Field and Constant Voltage Scaling"). However, the scaled structure has another problem now, namely, the closer proximity of source and drain-depletion region. As a consequence, the device is more susceptible to punch-through breakdown. To avoid this problem, the substrate doping is increased to reduce the depletion widths. Decreasing the junction depths also decreases the short-channel threshold voltage effects.

Complementary metal oxide semiconductor (CMOS) was first proposed in the 1960s but was not seriously considered until the severe limitations in power density and dissipation in nMOS circuits became apparent. Today CMOS is the dominant technology in IC manufacturing. CMOS uses both pMOS and nMOS transistors to form logic elements. The advantage of CMOS is that its logic elements draw significant current only during the transition from one state to another and very little current between transitions—hence power is conserved. In the case of an inverter, in either logic state one of the transistors is off. Because the transistors are in series, no current flows. A known deficiency of MOS technology is its limited load-driving capabilities (because of limited current sourcing and sinking abilities of pMOS and nMOS transistors). Bipolar

transistors have higher gain, better noise characteristics, and better high frequency characteristics.

BiCMOS combines bipolar and CMOS transistors in a single IC. By retaining benefit of bipolar and CMOS, BiCMOS is able to achieve VLSI circuits with speed-power-density performance previously unattainable with either technology individually. In Table 4.9 we summarize advantages/disadvantages of each of these transistor types.

CMOS Scaling Issues

Introduction

The driver behind the fantastic pace of miniaturization in electronics technology over the past 100 years (Figure 4.54a) is economics. This progress is projected to continue with 200 million transistors per chip today and 10 billion projected by 2016 (Figure 4.54b). In 1962 the semiconductor industry was worth more than $1 billion; by 1978 it was above $10 billion; and by 1994 it was more than $100 billion. From Appendix 1B, the IC market for 2003 was $166 billion worldwide, with 2004 projected at $241 billion (data from the World Semiconductor Trade Statistics; http://www.eetimes.com/ news/semi/showArticle.jhtml). From the same appendix, we learn that the IC business is feeding an electronic equipment business, which today is in the trillion-dollar range! Traditionally, the cost/function in IC reduced by 25% to 30%/year, and this enabled the industry to grow by 15%/year. To maintain this growth, the number of functions per IC has to be continuously increased. This necessitates an increase in transistor count, an increase in clock speed (= increased operations per unit time), and a decrease of feature size (Moore's first law: Chapter 1 and Figure 4.55, this volume, and Volume II, Chapter 2). After reviewing fundamental challenges in MOSFET miniaturization, we present a short treatise on scaling of CMOS devices. The downscaling factor α has typically been $\sim\sqrt{2}$. Today, the smallest features in a CMOS transistor are approaching atomic dimensions. In this context, we cover technical and practical limits involved in downscaling Si CMOS devices before quantum mechanics takes over. In response to that, we cover suggested Si solutions. Alternatives to Si-based MOSFETs include single-electron transistors, quantum wells, wires, dots, resonant tunneling diodes and transistors, and molecular electronics.

Fundamental Barriers to MOSFET Scaling

The rate of progress in downscaling Si-based electronics, perhaps the most important development of the 20th century, is said to be running out of steam. Today the presumed limit to scaling of MOSFETs shown on the International Technology Roadmap for Semiconductors [Appendix 1A and http://public.itrs.net (from 2003 and updated in 2004)] is often projected to occur in 2018 or 2020, when half-pitch spacing of metal lines will be 18 nm and device gate

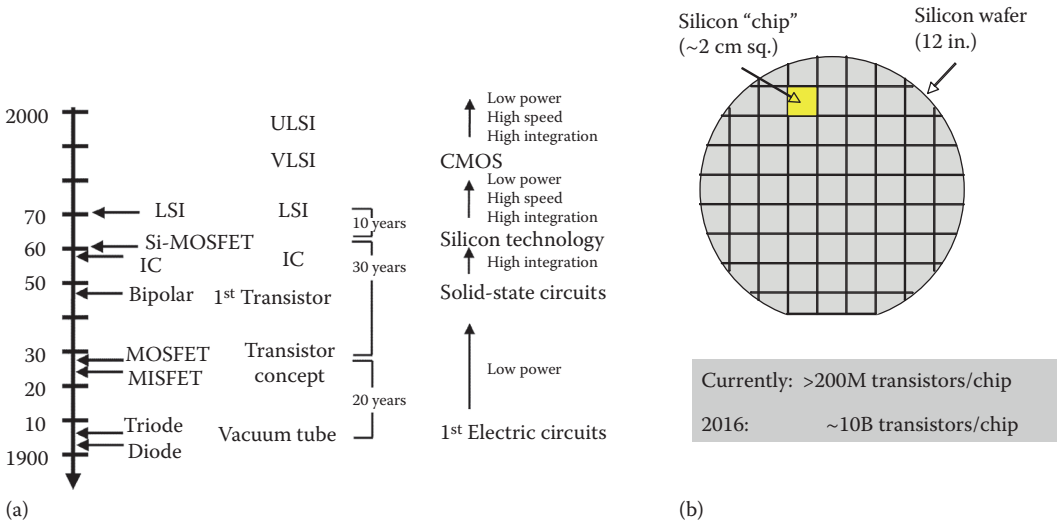


FIGURE 4.54 (a) History of electronics in the 20th century. In 100 years, the feature size was reduced by 1 million times. (b) Current and future (2016) transistors per chip.

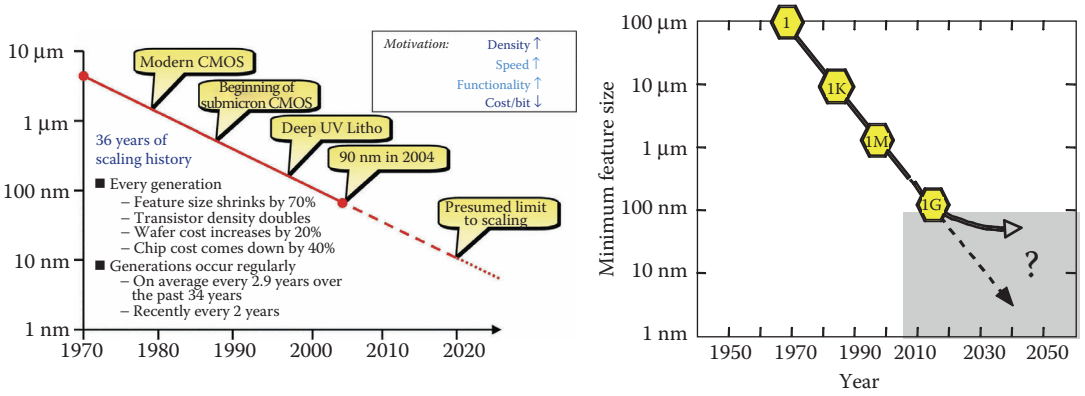


FIGURE 4.55 Two versions of Moore's first law. Scaling of minimum IC feature size.

lengths will be 7 nm (see Figure 4.55a). Technical limits are debatable and can be tackled, but fundamental limits are firm, and we discuss these ultimate limits first. Fundamental limits include electromagnetic, thermodynamic, and quantum mechanical limits.

From the electromagnetic point of view, the speed of light fundamentally limits the electron transport along IC wiring to ~60 ps/cm ($\tau > l/c$).* In today's technology, transmission times are getting longer than transistor switching times (now approaching 1 ps). In other words, as transistors are getting smaller and faster, on-chip interconnects (wires) are not getting any shorter or faster! This means that a larger percentage of chip area is devoted to metal interconnects as transistors get smaller. The capacitance of the wires is also not decreasing nearly as fast as the transistor sizes (in practice $\tau \sim RC \sim l^2$, with R and C proportional to l; see also "Constant Field and Constant Voltage Scaling").

The fundamental thermodynamic limit of any electrical device used as a switch (like MOSFETs in computers) is the energy required to produce a binary transition that can be distinguished: independent of the system architecture one must be able to distinguish between logic 0 and 1. CMOS uses an irreversible process with an inevitable intrinsic dissipation of energy. The von Neumann-Landauer[17] formula sets the theoretical limit for scaling of the energy in all irreversible technologies as (at 300 K):

$$E_{bit} > kT \ln 2 \approx 18 \text{ meV} \qquad (4.70)$$

It was Landauer who realized that with the binary value of a bit unknown, erasing that bit changes the logical entropy of the system from $k\log 2$ to $k\log 1 = 0$. If the physical representation of the bit is part of a thermalized distribution (so that thermodynamics does apply), then this bit erasure is associated with the dissipation of energy in a heat current of $kT\log 2$ per bit. The practical performance of a MOSFET is limited by its thermodynamic energy efficiency, i.e., the useful work accomplished per unit energy dissipated. If a MOSFET consumes an amount of free energy E and performs N useful operations, the energy efficiency is given by:

$$\eta = E/N \qquad (4.71)$$

with η in units of operations per unit energy, or operations/s/W. Using the thermodynamic energy limit per bit from Equation 4.70, this would enable an energy efficiency of $\eta = 3.5 \times 10^{20}$ operations/s/W. In the case of 90-nm VLSI technology, for minimal operations (e.g., conventional switching of a minimum-sized transistor), E is on the order of 1 fJ (femtojoule) and $\eta = 10^{15}$ operations/s/W. We see that the heat from the thermodynamic and logical entropy associated with a bit are currently five orders of magnitude lower than the electrostatic erasure energy.

From classical physics we expect a transistor gate to be a perfect barrier where no electrons flow through, but from quantum physics (Chapter 3) we know that electrons are extended wave-like objects

* It is important to note that the electric field propagates at the speed of light, but the electrons making up the electric current do not travel this fast themselves since they have mass. The individual electrons travel much more slowly. The electrons that go "into" one end of the wire are not the same electrons that "come out" of the other end.

that can "tunnel" through a thin gate, and this tunnel rate goes up exponentially as the gate gets thinner. MOSFET scaling is hampered by quantization of charge, which becomes important at a length scale of 10 nm in all materials, and energy level quantization, which becomes important in semiconductors also at a length scale of 10 nm and a factor of about 10 times lower for metals (see Chapter 3). The quantum mechanical limits on MOSFET scaling, based on the Heisenberg uncertainty principle (HUP), can be summarized as:

$$\Delta p_x \Delta x \geq \frac{h}{2\pi} = \hbar \text{ (see Equation 3.106)}$$

$$x_{min} = \frac{\hbar}{\Delta p_x} = \frac{\hbar}{\sqrt{2 m_e E_{bit}}} = \frac{\hbar}{\sqrt{2 m_e k_B T \ln 2}} = 1.5 \text{ nm} \tag{4.72}$$

which puts a fundamental limit on the integration density. So quantum mechanics allows for a very small switch indeed, but other factors, such as material temperature tolerance range or economics, will limit before this fundamental limit sets in. And also:

$$\Delta E \Delta t \geq \frac{h}{2\pi} = \hbar \text{ (see Equation 3.107)}$$

$$\Rightarrow \tau \sim 0.04 \text{ ps or } 25 \text{ Tbits/s} \tag{4.73}$$

which sets a fundamental limit to switching speed.

Constant Field and Constant Voltage Scaling

Scaling of transistors may be achieved by reducing all dimensions (lateral and vertical) by $1/\alpha$. Suppose we scale down all transistor dimensions this way, including the gate oxide thickness (t_{ox}); then, to maintain the original electric field distribution, including the gate field, one must decrease voltages to $1/\alpha$ and increase the channel doping by α. Typically $1/\alpha = 0.7$ (30% reduction in the dimensions). The scaling variables are summarized in Table 4.10. This is called *constant field scaling* because the electric field across the gate oxide does not change when the technology is scaled.

If the power supply voltage is maintained constant, the scaling is called constant voltage. In this case, the electric field across the gate oxide increases as the technology is scaled down. Because of gate oxide breakdown, at less than 0.8 µm only "constant field" scaling is used. As an example, let us consider the consequences of a 30% scaling effort in the constant field regime ($\alpha = 1.43$, $1/\alpha = 0.7$).

Table 4.10 Scaling of Variables

Scaling Variables	Before	After
Voltage	V	V/α
Gate length	L	L/α
Gate width	W	W/α
Gate oxide thickness	t_{ox}	t_{ox} /α
Junction depth	X	X/α
Channel doping	N	Nα

- Device/die area: $W \times L \Rightarrow (1/\alpha)^2 = 0.49$. In practice, microprocessor die size grows about 25% per technology generation! This is a result of added functionality.
- Transistor density: (unit area)/$(W \times L) \Rightarrow \alpha^2 = 2.04$. In practice, memory density has been scaling as expected (not true for microprocessors…).
- Gate capacitance: $W \times L/t_{ox} \Rightarrow 1/\alpha = 0.7$
- Drain current: $(W/L)\ (V^2/t_{ox}) \Rightarrow 1/\alpha = 0.7$
- Gate delay: $(C \times V) / I \Rightarrow 1/\alpha = 0.7$ (see Figure 4.56)
- Frequency: $\alpha = 1.43$ (see Figure 4.56). In practice, microprocessor frequency has doubled every technology generation (2–3 years)!
- The processing power (number of gates/area times speed): goes up as (unit area)/$(W \times L) \times$ frequency $\Rightarrow \alpha^3 = 2.92$
- Power: $C \times V^2 \times f \Rightarrow (1/\alpha)^2 = 0.49$
- Power density: $1/t_{ox} \times V^2 \times f \Rightarrow 1$
- Active capacitance/unit area: Power dissipation is a function of the operation frequency, the power supply voltage, and the circuit size (number of devices). If we normalize the power density to $V^2 \times f$, we obtain the active capacitance per unit area for a given circuit. This parameter can be compared with the oxide capacitance per unit area: $1/t_{ox} \Rightarrow \alpha = 1.43$. In practice, for microprocessors, the active capacitance/unit area only increases between 30% and 35%. Thus, the twofold improvement in logic density between technologies is not achieved.

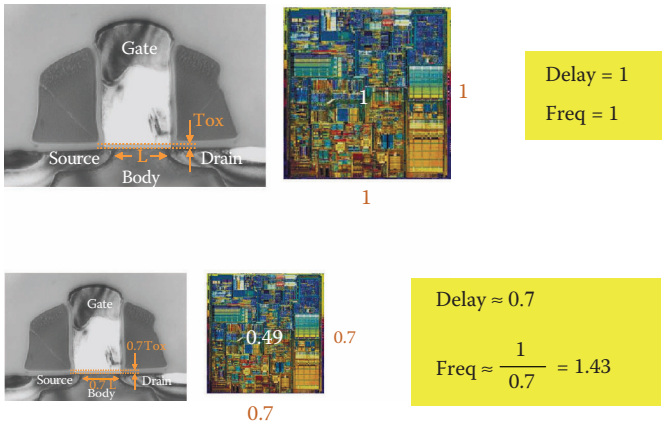


FIGURE 4.56 Scaling of delay and frequency in a MOSFET.

From the preceding, if we succeed in constant field scaling, the speed of the transistor goes up by α and the number of devices/area goes up by α^2, the processing power (number of gates/area times speed) goes up by α^3, and the heat generation/area (power density) remains unchanged (=1). Unfortunately, you cannot scale down voltage indefinitely. There are fixed voltage characteristics in MOSFETs, such as the 59-mV/decade turn-on current relation so the industry scales down voltage less aggressively than feature size and as a consequence the electric fields in the gate oxide have been increasing. If the electric field increases by E and the features are scaled down by α, then the doping goes up by $E \times \alpha$, the power density (heat) goes up by E^2, the gate leakage current increases, and transistors are more likely to fail.

As pointed out earlier, higher transistor densities are only possible if the interconnects also scale. Typical wires have a capacitance of ~0.2 fF/μm compared with 2 fF/μm for a typical gate capacitance. Reduced interconnect wire width leads to increased resistance, and denser interconnects also lead to a higher capacitance. For a wire interconnect, R and C are proportional to l, whereas the RC delay is proportional to l^2. This leads to unacceptably great delays for long wires. To account for increased parasitics and integration complexity, more interconnection layers are added with thinner and tighter layers for local interconnections and thicker and sparser layers for global interconnections and power. One also breaks long wires into N shorter segments and drives each one with an inverter or buffer in so-called *repeaters*.

In Table 4.11 constant field and constant voltage scaling of a MOSFET are compared.

Technical and Economic (Practical) Barriers to Downscaling and Possible Solutions

Introduction Besides the fundamental limits to CMOS scaling (see above), there are several technical limits to consider, and ultimately the economics might decide the ultimate fate of CMOS. However, Si devices are the smallest devices currently available

TABLE 4.11 Constant Field and Constant Voltage Scaling of a MOSFET

Parameter	Constant	Field	Constant	Voltage
Supply voltage (V)		1/α		1
Length (L)	1/α		1/α	
Width (W)		1/α		1/α
Gate oxide thickness (t_{ox})		1/α		1/α
Junction depth (X_j)		1/α		1/α
Substrate doping (N_A)	α		α	
Electric field across gate oxide (E)	1		α	
Depletion layer thickness		1/α		1/α
Gate area (Die area)	$1/\alpha^2$		$1/\alpha^2$	
Gate capacitance (load) (C)		1/α		1/α
Drain current (I_d)		1/α		α
Transconductance (g)	1		α	
Gate delay	1/α		$1/\alpha^2$	
Current density		α		α^3
DC and dynamic power dissipation	$1/\alpha^2$		α	
Power density		1		α^3
Power-delay product	$1/\alpha^3$		1/α	

in the market, and there are no other candidates smaller than Si in the near future. Introducing new materials, structures, and processes will solve problems of downsizing.

Mass Manufacturability and Economics As we scale down components, they have fewer atoms, electrons, and holes, and unless one has perfect control of the fabrication process, statistical fluctuations start coming in. Obviously MOSFET fabrication lines do not have perfect control, and statistical fluctuations get worse as the device gets smaller (the tyranny of small numbers). As a consequence, as transistors get smaller, there are greater variations in device-to-device characteristics, such as gate threshold voltage (V_T) and output current (I_D). Combined with fundamental voltage fluctuations (Johnson noise) and current fluctuations (shot noise), this leads to a greater probability of errors. As on-chip circuitry gets smaller and more complex, it is also increasingly sensitive to smaller defects. This lowers the yield and drives up costs even further. The demise of CMOS on technical grounds has been announced several times before.* However, it is more likely that one will hit an economic limit before the ultimate technical limit is reached. Because of increased sophistication of lithography tools, cost of IC fabrication lines is increasing exponentially with time (Moore's second law) and is now in the multi-billion-dollar range (1967 cost of a silicon fabrication laboratory: $2 million; 2002 cost: ~$3 billion; projected 2015 cost: ~$100 billion). As an example of mass manufacturability and economics, consider resolution-enhancing techniques (RET) that enable subwavelength printing of MOSFETs (see Volume II, Chapters 1 and 2). Basically, RET is a set of optical wave-front engineering techniques that allows for the pattern transfer of features smaller than the lithography wavelength. Feature size scaling is now outpacing lithography wavelength advances, forcing design compromises and complex and costly optical wave-front management techniques. It is important to recognize that at less than 90 nm there is no yield without RET, but this unfortunately vastly increases mask costs and severely constrains design flexibility. Implementing all of these technology innovations may increase manufacturing and development costs to beyond the reach of all but a few global entities. This subwavelength crisis is summarized in Figure 4.57.

A practical limit may thus be reached because of cost, reliability, and performance considerations. It is safe to say that it is not really known where that practical limit really lies, and, as remarked earlier, it is certainly too early to give up on Si. It is also clear that pushing CMOS to its ultimate limits will require revolutionary materials and device innovations that pose significant scientific and engineering barriers. These problems have caused more changes in IC technology and materials in the past 5 years than in the previous 40 years with work in SiGe, silicon on insulator (SOI), strained Si, alternative dielectrics, new metallization systems, and so on.

Carrier Mobility Limitations The effective mobility of charge carriers in the channel of a MOSFET differs from bulk mobility as a result of additional scattering mechanisms associated with the channel surface, including interface charge scattering and surface roughness scattering. These mechanisms also tend to depend on the vertical electrical field as illustrated in Figure 4.58, where E_{EFF} represents the effective electrical field. As gate length shrinks, mobility decreases, and increased doping is often used to combat this short-channel effect. Another

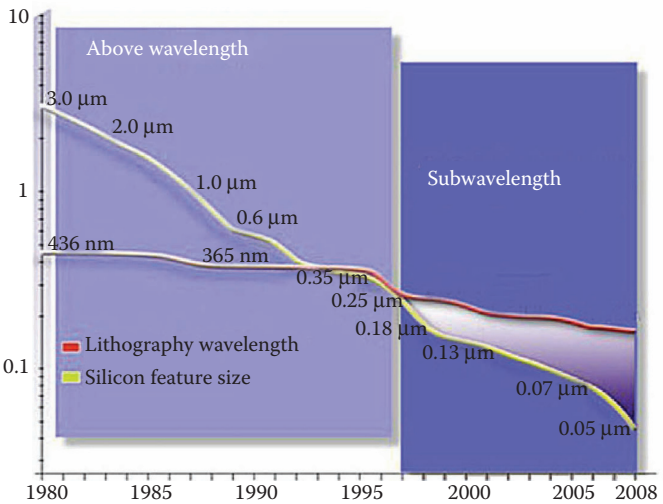


FIGURE 4.57 The subwavelength crisis.

* In the late 1980s, 0.1 µm was called the "brick wall"; today 50 nm is called the "red brick wall" and 100 nm the "fundamental limit."

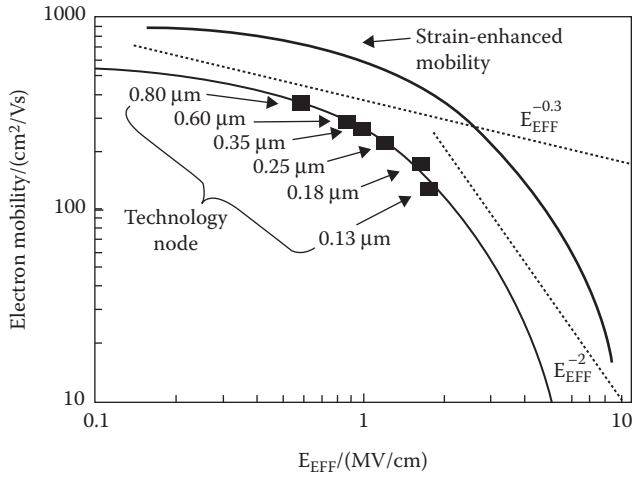


FIGURE 4.58 Mobility versus technology scaling trend for Intel process technologies. The *x*-axis is the effective electrical field E_{EFF}. (From Thompson, S.E., M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, et al. 2004. A 90-nm logic technology featuring strained-silicon. *IEEE Trans Electron Devices* 51:1790–7.[49])

possible solution involves electron and hole mobility enhancement by using strained Si. This method increases mobility of electrons and holes in the channel of a FET by applying tensile (to increase electron mobility) or compressive strains (to increase hole mobility) to the channel. Strain alters the energy band structure and enables charge carriers to move with less resistance through silicon. The MOSFET channel is very close to the device surface, so that the strain must occur in the topmost ~0.01 µm of the silicon to realize a mobility enhancement in the channel. The effect of strain on mobility was noticed in the early 1950s, the exploitation of the phenomenon in an actual device was first carried out at MIT in the early 1990s (Gene Fitzgerald, who founded AmberWave in 1998). By August 2002, Intel used strained silicon in their HT Prescott P4 Processor. Models suggest that a strained silicon lattice can give hole mobilities up to 4× the unstrained value and electron mobilities up to 1.8× the unstrained value.[18]

One way to introduce strain in Si is to use a $Si_{1-x}Ge_x$ crystal alloy substrate. Such an alloy is possible because both materials create diamond-type lattices and their lattice constants (a) are close. Vegard's law determines the lattice constant of the alloy, which is a linear average between the lattice constants of Si (5.431 Å) and Ge (5.658 Å):

$$a_{alloy} = (1 - x) \cdot a_{Si} + x \cdot a_{Ge} \qquad (4.74)$$

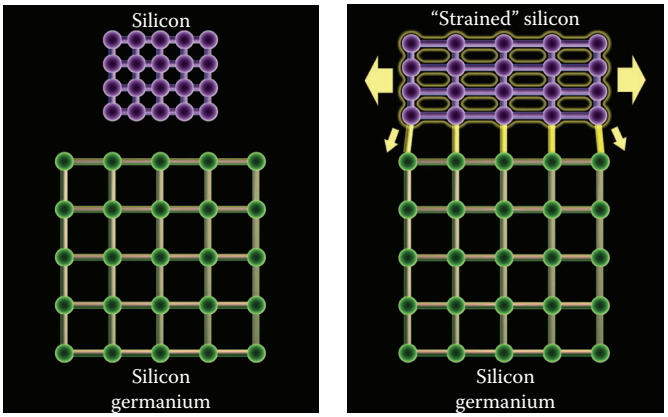


FIGURE 4.59 Formation of a strained Si lattice using an epitaxial strain-inducing template.

First, one introduces Ge in a Si lattice, and then the Si gate material is grown epitaxially using molecular beam epitaxy (MBE) on top of this $Si_{1-x}Ge_x$; the Si atoms align with the underlying lattice, causing a lattice strain in the silicon (see Figure 4.59).

The electrons flow up to 70% faster through strained silicon, resulting in a 17% increase in transistor speed and a 34% reduction in power consumption.

The SiGe devices market has grown more rapidly than was anticipated, and III–Vs producers are concerned about the switching from Si to SiGe instead of using GaAs in a number of key high-speed applications. With SiGe matching GaAs performance and beating it on cost, GaAs still edges out SiGe in terms of power-added efficiency (PAE)* (typically 40% vs. <20%). So for those applications where efficiency is less important—such as for shorter-range wireless networks like WiFi and Bluetooth—SiGe is competing hard with GaAs (see also Volume III, Chapter 5 where we compare GaAs with SiGe). An example SiGe process is the 130-nm SiGe BiCMOS technology designed for high-speed wireless and optical communications applications from Jazz Semiconductor (http://www.jazzsemi.com/process_technologies/sige.shtml). This process combines industry standard 130-nm CMOS with 200-GHz SiGe heterojunction bipolar transistors (HBTs) and npn transistors for high-performance RF and millimeter-wave ICs. This technology enables the design of the highest

* In an RF power amplifier, PAE is defined as the ratio of the difference of the output and input signal power to the DC power consumed.

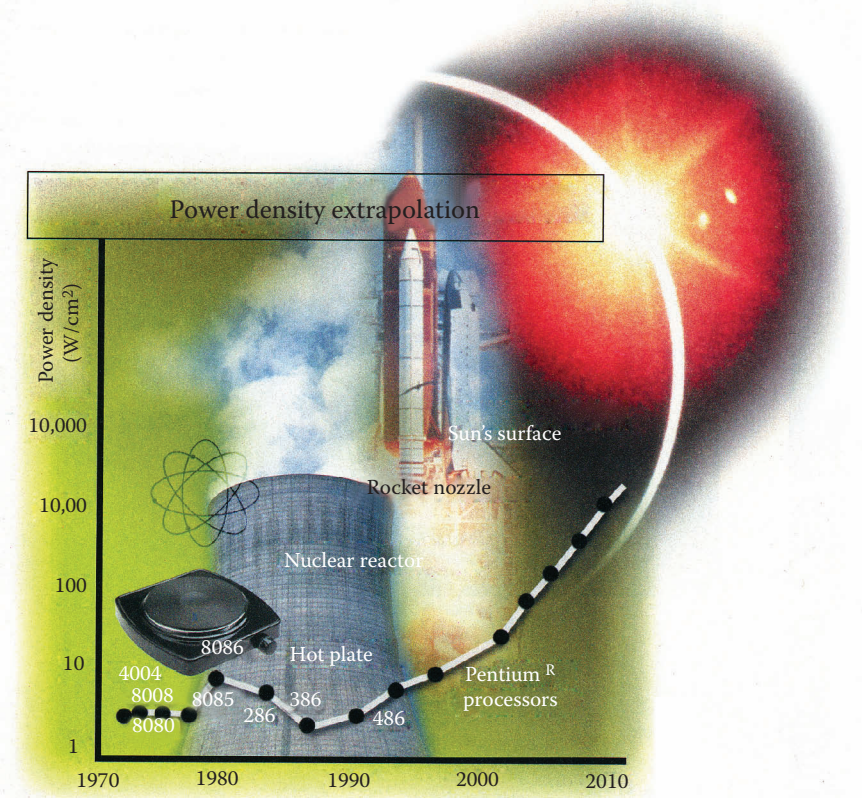


FIGURE 4.60 Power densities of various MOSFET technologies, one of the biggest challenges to further miniaturization.

performance circuits in advanced high-speed optical, wireless, and millimeter-wave applications.

Power Budget The attempts to increase the carrier mobility (see above), reduce the threshold voltage, minimize current leakage in the off state, and minimize losses in the interconnect system of MOSFETs are all somewhat interrelated. The further miniaturization of MOSFETs is limited mostly by both static (passive) and dynamic (active) power. Dynamic power is required to charge and discharge load capacitances when transistors switch and was calculated earlier as $C \times V^2 \times f$. From Figure 4.60 the latest Pentium chips dissipate about 100 W/cm^2—more heat than a stove-top cooking surface.

Static power is consumed even when the chip is quiescent and is calculated from $P_{off} = I_D(off)V_{DS}$. Gate current leakage, $I_D(off)$, is parasitic, on all the time, and uses extra power and generates extra heat. Over the years this gate leakage (10× increase per technology node) has caused the off-current [$I_D(off)$] to encroach on the on-current [$I_D(on)$] (Figure 4.61). Gate current reduces the on/off current ratio, which reduces operating margins.

Direct tunneling through the gate insulator is the dominant cause of static power dissipation. Single atom defects can cause local leakage currents 10–100 times higher than the average current, impacting reliability and generating unwanted variation between devices.

To reduce resistance and thus power consumption, both Ti and Co silicides are used as source/drain contact materials. Silicides are formed at the Si interface by an approach referred to as self-aligned-silicide (salicide), which selectively forms the silicide on the drain and source junctions as well as on the polysilicon gate. The salicide process is detailed in Volume II, Chapter 1. An important advantage of the

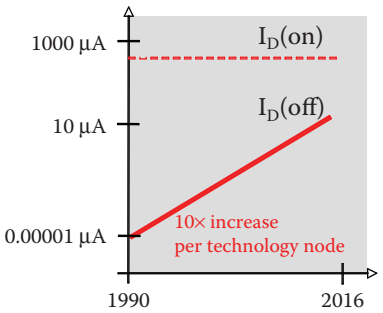


FIGURE 4.61 Evolution of the off-current [$I_D(off)$].

salicide process is that the entire junction area is used for contact formation, which translates into a lower overall contact resistance. The silicide contact materials on the transistors are then further contacted through chemical vapor-deposited (CVD) plugs of pure W filling the contact vias. CVD tungsten (from WF_6) exhibits extraordinarily good conformality, and the silicide adhesion/barrier layer protects the underlying Si from attack by fluorine during deposition; ensures adhesion of W to the silicon; and forms a good contact to the drain, source, and gate.

The vertical W plugs then make contact to horizontal Al (or Cu) interconnects as shown in Figure 4.62. To planarize the wafers between metallizations, IBM pioneered the damascene process, based on chemical-mechanical polishing (CMP). The damascene process is repeated many times to form the alternating layers of wires and vias that form the complete wiring system of a silicon chip. The damascene process is detailed in Volume II, Chapter 8 on single and dual damascene processes.

Solutions to reduce the excessive heat generation include using cooling fins and chip fans. Thus, it would appear that MEMS research for passive and active power management of ICs could present a tremendous new market opportunity. A wide variety of implementations of MEMS structures for cooling are feasible, but the associated costs, size, and practicality are the real issue.

Oxide Leakage and Device Isolation Today's MOSFET has evolved through many modifications to provide solutions for parasitic problems that emerged during downscaling. The gate dielectric is still mostly SiO_2-grown by thermal oxidation. As a result of continued scaling, that gate oxide has thinned down to the extent that direct tunneling has become a serious concern. As the thickness scales below 2 nm, leakage currents due to tunneling increase drastically. Despite the tremendous benefits that SiO_2 has presented toward the success of CMOS, scaling and high-power consumption are now urging process engineers to look for higher-k materials (k is the relative dielectric constant of the material, 3.9 for silicon dioxide).* High-k dielectric materials are used to minimize the tunneling current and to prevent the diffusion of dopants out from the gate. There are three major categories of high-k materials:

1. $4 < k < 10$; Si_3N_4/SiO_2 (5 ~ 6), Si_3N_4 (~7)
2. $10 < k < 20$; Ta_2O_5, Al_2O_3 (~10), TiO_2, $ZrSi_xO_y$, $HfSi_xOy$, $LaSi_xO_y$ (last three are silicates), $ZrAl_xO_y$ (aluminate), ZrN_xO_y (nitride)
3. $15 < k < 30$; ZrO_2, HfO_2, La_2O_3, Pr_2O_3, Gd_2O_3, Y_2O_3, crystal Pr_2O_3 (~30)

A prominent candidate for a gate dielectric of the future is Al_2O_3. Recent studies show that Al_2O_3 dielectrics less than 1 nm thick are superior to previously used Si_3N_4 and SiO_2 (Al_2O_3 film has been used to make 1-Gbit dynamic random access memory). As we will see in Volume II, Chapter 7, it may be deposited, for example, monolayer by monolayer with atomic layer deposition (ALD; see Volume II, Figure 7.47).

The most significant advances in reducing overall device size and packing density of transistors have come from improved isolation methods. The traditional junction isolation technique requires the p+ deep diffusion to be aligned to the n+ buried layer that is covered by a thick epitaxial layer. The area (and hence junction capacitance) is determined by alignment tolerance, area for side diffusion, and allowance for the spread of the depletion region. Modern isolation techniques such as oxide isolation and trench isolation eliminate most of these requirements. Oxide isolation processes were

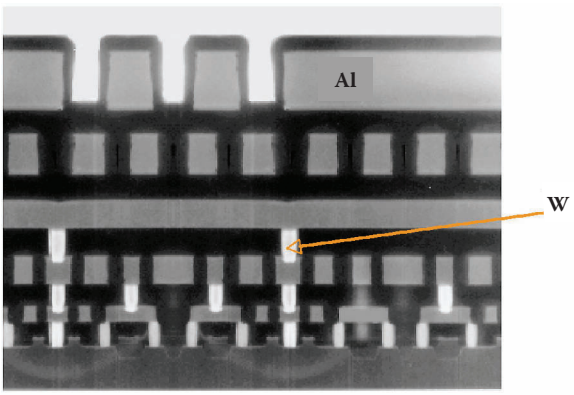


FIGURE 4.62 Multilevel aluminum metallization with tungsten plugs.

* There is no standard symbol for the dielectric constant; you may see it referred to as k, κ, ε, ε' or ε_r.

first introduced in the late 1970s. These first attempts used wet anisotropic etching (with KOH) of the [100] Si wafer with Si_3N_4 as a mask to etch V-grooves. Oxide is either deposited or grown (e.g., LOCOS) to fill the V-grooves. Because no side spread of the depletion region occurs, the packing density can be higher. These V-grooves require spacings of several tens of microns between devices, and that is not compatible with the required densities in VLSI. To further reduce the area between adjacent transistors, trench isolation is used, making use of dry etched vertical trenches. The trenches are typically 2 μm wide and 5 μm deep. The process is illustrated in Figure 4.63. After a patterning step (a), a channel stop (Chanstop) trench area is dry etched into the substrate (b); the trench is refilled with oxide (c); and chemical-mechanical polishing (CMP) removes the oxide on the nitride to get a flat surface (d).

In this same regard, silicon-on-insulator (SOI) wafers are used more and more in MOSFET construction. Differences from bulk CMOS include the fact that the transistor source, drain, and body are now surrounded by insulating SiO_2 rather than by a Si substrate (well). Advantages include reduced source and drain to substrate capacitance, lower passive current, and denser layout. Because of the high cost of SOI wafers, only some microprocessors, which command high prices and compete on speed, have embraced this technology (March 2004, Apple's Xserve G5 and AMD 90-nm processor at the end of 2004). An important SOI extension is to use

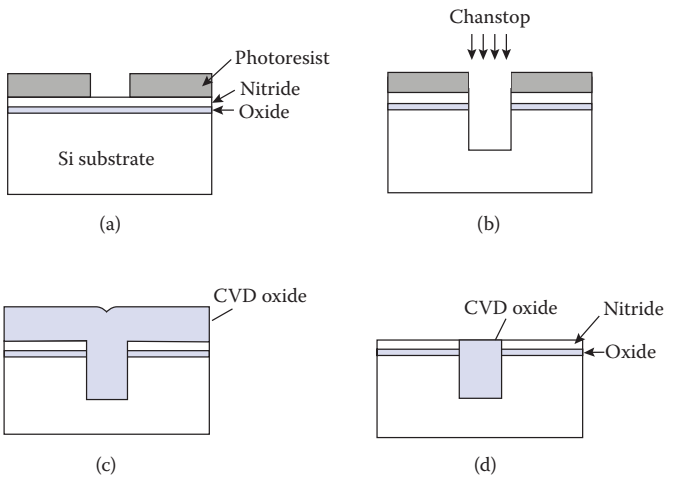


FIGURE 4.63 Trench isolation channel stops. See text for details.

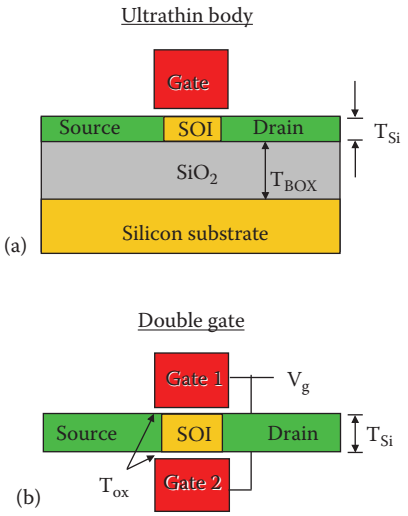


FIGURE 4.64 Ultrathin-body SOI FET (a) and a double-gate FET (b).

fully depleted SOI (Figure 4.64a). In this case, an ultrathin silicon body is used so that the entire body is depleted before the threshold voltage is reached (the body is thinner than the channel depletion width). Ultrathin-body MOSFETs feature very low leakage currents and a raised source/drain structure leading to low resistance and a ~12-nm scaling limit because of the low channel doping. The double-gate transistor (DGT) structure, proposed in the 1980s, further improves this situation. The DGT is composed of a conducting channel (usually undoped), surrounded by gate electrodes on either side (Figure 4.64b). This ensures that no part of the channel is far away from a gate electrode. Advantages of this design include the reduction of short-channel effects (e.g., a decreased V_T resulting from a reduced channel depletion charge), higher current drive capability, good electrical characteristics with a high on/off current ratio, and a sharp I–V slope. Reducing the body thickness further decreases gate current leakage, I_D(off), but has the tradeoff of increasing series resistance. This series resistance is minimized using a raised source/drain type of structure as in an ultrathin-body MOSFET or FinFET (see next section). It is projected that this technology will enable sub-10 nm scaling.

3D Transistors: Go Up, Young Man! A viable approach to stay on the Moore's curve is to deposit additional Si layers and fabricate transistors in them; in other words, start growing in the third dimension.

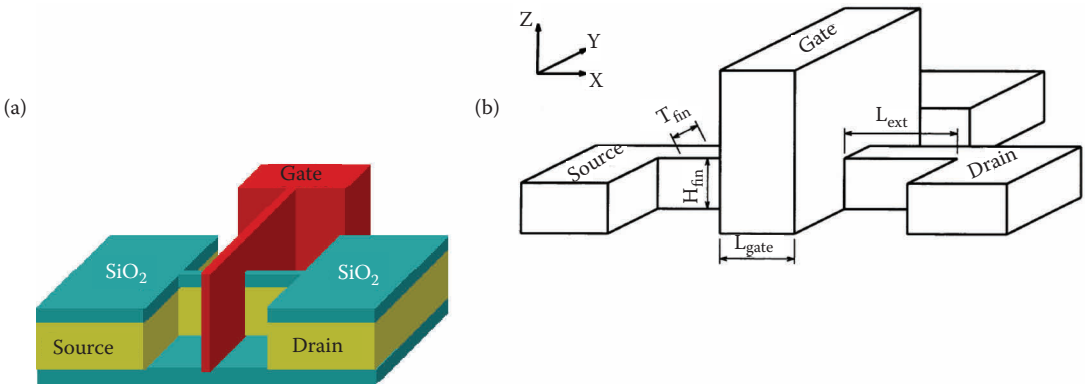


FIGURE 4.65 (a) FinFET structure. (b) Dimensioning: $H_{fin} >> T_{fin}$, top gate oxide thickness >> sidewall oxide, effective channel length $L_{eff} = L_{gate} + 2 \times L_{ext}$ thickness, effective channel width $W = T_{fin} + 2 \times H_{fin}$.

Present MOSFET designs are planar, "buried" in the flat upper surface of the Si substrate. Each time you double circuit density by stacking, you can move a whole Moore's law generation without using smaller features. But interconnection is tricky, as are connections to the outside world. It is also more difficult to make high-performance transistors from deposited films, and additional processing tends to spoil previous layers. Source and drain are getting too close for reliable performance, and the gate between them is getting harder to define. 3D techniques can extend microprocessor design for several more generations, but after that, a new design paradigm will be needed to deal with increasing errors and defects.

The raised drain and source design in SOI approaches is a step in the right direction. In the FinFET approach, this concept is pushed even further; the transistor structure is put on top of the chip substrate as shown in Figure 4.65a. The term *FinFET* was coined by University of California, Berkeley researchers to describe a nonplanar, double-gate transistor built on an SOI substrate. The FinFET is a 3D device structure with a channel Si fin, sitting on the substrate, surrounded by a gate that effectively makes two to three gates in parallel. The gate wraps around the fin to form a very small channel. The Si fin has an insulator on the top, and the gate has oxide on either side (two channels). The current flows parallel to the device surface, and the channel width depends on the fin height (for dimensioning see Figure 4.65b). A lightly doped channel in a FinFET affords good gate control. Spacers are used to self-align the gate to source/drain, and the fin height can be used to increase the on-current, limited by aspect ratio. Many benefits arise from this design. Short-channel effects are reduced, and there is very low leakage current, high on/off current ratio, low voltage operation, efficient gate design (less switching power), flexibility of using multiple fins for better performance, compatibility with current manufacturing processes, and scalability to sub-10 nm.

FinFET is the leading candidate set to replace the classic MOSFETs. It is based on a 3D architecture that uses double gates. It can be manufactured with very minor modifications to the current self-aligned process flow. However, there are many issues that have to be resolved before it can be used commercially. A 25-nm transistor operating on just 0.7 V was demonstrated in December 2002 by the Taiwan Semiconductor Manufacturing Company. The "Omega FinFET" design, named after the similarity between the Greek letter omega and the shape in which the gate wraps around the source/drain structure, has a gate delay of just 0.39 ps for the n-type transistor and 0.88 ps for the p-type.

In Figure 4.66, we sketch the predicted material and device evolution for the years 2004–2020.

From the preceding, it appears that making Si-based MOSFETs smaller than 50 nm is not a problem. But the "ultimate" FET may not be made of silicon. Alternative approaches to fabricate logic switches, which we analyze in this volume Chapters 3 and 5 and in Volume III, Chapter 3 on nanotechnology, include molecular electronics, resonant tunneling diodes, and single-electron transistors.

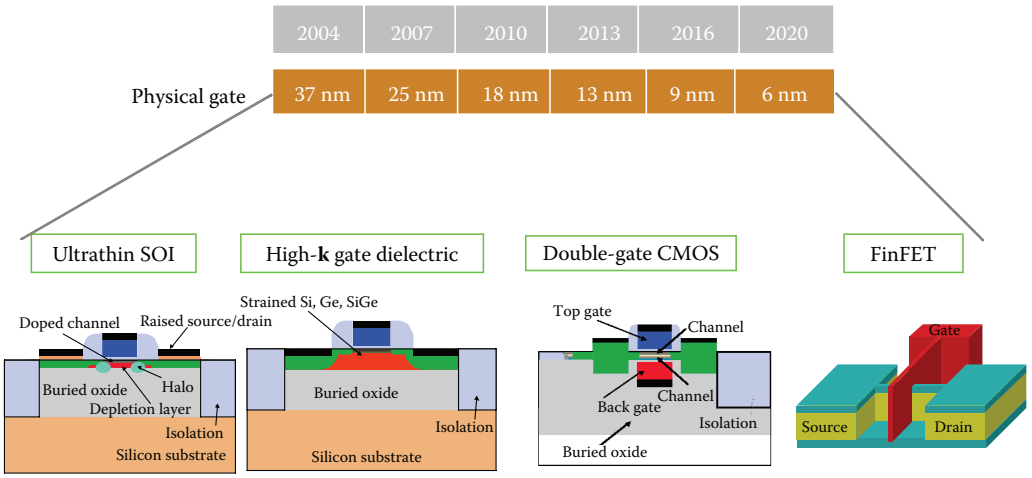


FIGURE 4.66 Evolving materials and device structures in MOSFET evolution. Halo from halo implant is a high-angle implant of well-type dopant species introduced into the device after transistor gate patterning.

Physicochemical Properties of Si

Introduction

In the previous sections, we mostly focused on Si as an electronic material. In this section we are looking at all the other properties that make Si such an attractive IC, MEMS, and NEMS material. We first consider Si as a simple substrate, where the semiconductor only plays the role as a flat convenient passive substrate. Next, we investigate the mechanical, thermal, and optical properties of Si. We finish this section with a short commentary on Si's biocompatibility.

Silicon as a Passive Substrate

Any IC, MEMS, or NEMS device needs a substrate. In some cases that substrate acts as a passive support only and does not play any other role in the operation of the device it supports. The use of Si as a passive substrate is discussed in more detail in Volume III, Chapter 1, which focuses on comparing nonlithography-based (traditional) and lithography-based (nontraditional) manufacturing methods. In Volume III, Table 1.9, we show a performance comparison of passive substrate materials in terms of cost, metallization ease, and machinability. We will see there that silicon, because of its extreme flatness, relative low cost, and well-established coating procedures, is often the preferred passive substrate—especially for thin films. The disadvantages of using Si substrates become more pronounced with increasing device size and low production volumes. An overwhelming determining factor for passive substrate choice is the final package of the device because the package often determines the volume and cost. Packaging is so important in the case of sensors that as a rule, their design should start from the package rather than from the sensor itself. For many mechanical sensor applications, single-crystal Si (SCS), based on its intrinsic mechanical stability and the feasibility of integrating sensing and electronics on the same substrate, presents an excellent substrate choice. But, as we will see in Volume III, Chapter 1, for chemical sensors, Si, with few exceptions, is merely the substrate and as such is not necessarily the most attractive option.

Silicon in Mechanical Sensors

Introduction

In mechanical sensors, the active structural elements convert a mechanical external input signal (e.g., pressure, acceleration) into an electrical signal output (voltage, current, or frequency). The transfer functions in mechanical devices describing this conversion are mechanical, electromechanical, and electrical.

In mechanical conversion, a given external load is concentrated and maximized in the active member of the sensor. Structurally active members are typically high-aspect-ratio elements such as suspended

beams or membranes. Electromechanical conversion is the transformation of the mechanical quantity into an electrical quantity, such as capacitance, resistance, charge, and so on. Often, the electrical signal needs further electrical conversion into an output voltage, frequency, or current. For electrical conversion into an output voltage, a Wheatstone bridge (see Figure 4.75) may be used as in the case of a piezoresistive sensor, and a charge amplifier may be used in the case of a piezoelectric sensor. To optimize all three transfer functions, detailed electrical and mechanical modeling is required. One of the most important inputs required for mechanical models is the experimentally determined, independent elasticity constants or moduli. In what follows, we describe in more detail what makes Si such an important structural element in mechanical sensors and present its elasticity constants.

Stress-Strain Curve and Elasticity Constants

Yield, tensile strength, hardness, and creep of a material all relate to the elasticity curve (the stress-strain diagram). Stress-strain curves for several types of materials are shown in Figure 4.67. For small

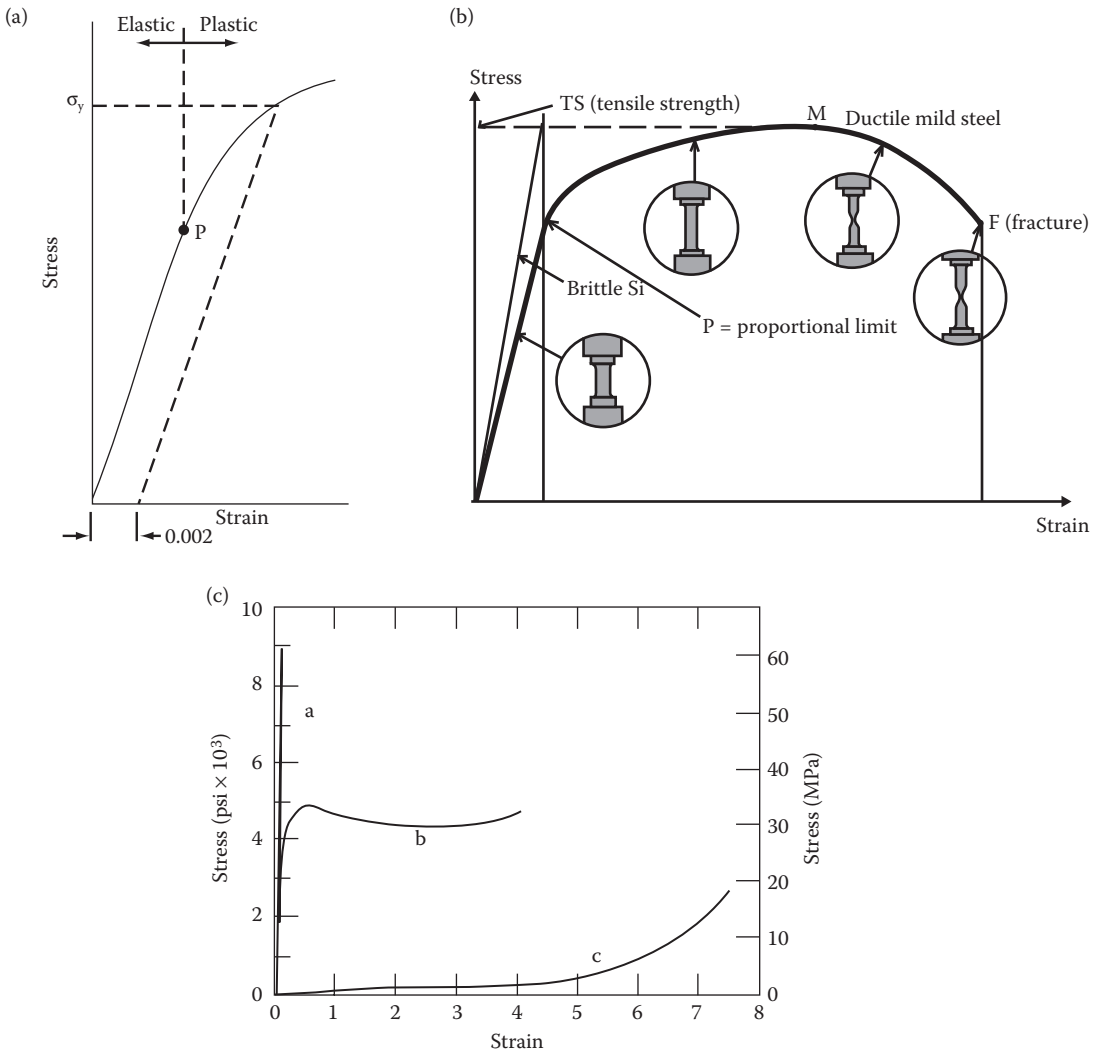


FIGURE 4.67 Typical stress-strain behaviors for different types of materials showing different degrees of elastic and plastic deformations. (a) Stress-strain curve for a typical material with the proportional limit *P*, and the yield strength σ_y, as determined using the 0.002 strain offset method. (b) Stress-strain curves for silicon and a ductile mild steel. The tensile strength (TS) of a metal is the stress at the maximum of the curve shown (*M*). Fracture is indicated by *F*. The various stages of a test metal sample are indicated as well. A high modulus material such as Si exhibits abrupt breakage; it is brittle with no plastic deformation region at all. (c) Stress-strain behavior for brittle (curve a), plastic (curve b), and highly elastic (elastomeric) (curve c) polymers.

strain values, Hooke's law (1635–1702) applies, that is, stress (force per unit area, N/m^2) and strain (displacement per unit length, dimensionless) are proportional, and the stress-strain curve is linear, with a slope corresponding to the elastic modulus E (Young's modulus, N/m^2). This regime, as marked in Figure 4.67a, is the elastic deformation regime (typically valid for $\varepsilon < 10^{-4}$). The magnitude of the Young's modulus ranges from 4.1×10^4 MPa (Pascal, Pa = N/m^2) for magnesium to 40.7×10^4 MPa for tungsten and 144 GPa for Invar,* whereas concrete is 45 GPa, aluminum is about 70 GPa, and elastomeric materials are as low as 10^{-3} to 10^{-2} GPa. Silicon is a stiff and brittle material with a Young's modulus of 160 GPa (Figure 4.67b). With increasing temperature, the elastic modulus diminishes.

For isotropic media, such as amorphous and polycrystalline materials, the tensile stress, σ_a, or the applied axial force per unit area and the axial or tensile strain, ε_a, are related as:

$$\sigma_a = E\varepsilon_a \tag{4.75}$$

with ε_a given by the dimensionless ratio of $(L_2 - L_1)/L_1$ $(\Delta L/L_1)$, that is, the ratio of the sample's elongation to its original length. The elastic modulus may be thought of as stiffness or a material's resistance to elastic deformation. The greater the modulus, the stiffer the material. A tensile stress usually leads to a lateral strain or contraction (Poisson effect), ε_l, given by the dimensionless ratio of $(D_2 - D_1)/D_1$ $(\Delta D/D_1)$, where D_1 is the original wire diameter and ΔD is the change in diameter under axial stress (see Figure 4.68). The Poisson ratio is the ratio of lateral over axial strain:

$$\nu = -\frac{\varepsilon_l}{\varepsilon_a} \tag{4.76}$$

The minus sign indicates a contraction of the material. For most materials, ν is a constant within the elastic range and fluctuates for different types of materials over a relatively narrow range. Generally, it is on the order of 0.25–0.35, and a value of 0.5 is the largest value possible. The latter is attained by materials such as rubber and indicates a material in which only the shape changes, whereas the volume remains constant. Normally, some slight volume change does accompany the deformation, and, consequently, ν is smaller than 0.5. The Poisson ratios for metals are typically around 0.33 (e.g., aluminum and cast steel are 0.34 and 0.28, respectively). For ceramics, it is around 0.25, and for polymers it is typically between 0.4 and 0.5. In extreme cases, values as low as 0.1 (certain types of concrete) and as high as 0.5 (rubber) occur.

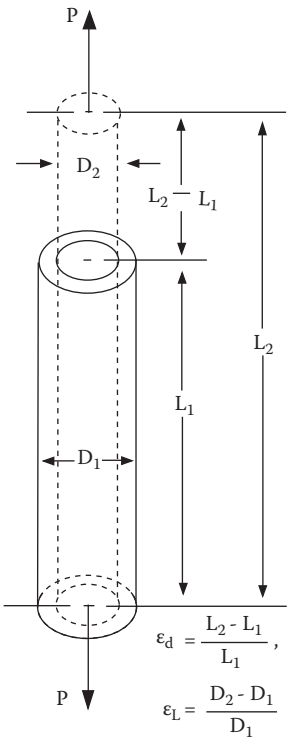


FIGURE 4.68 Metal wire under axial or normal stress; normal stress creates both elongation and lateral contraction.

For an elastic isotropic medium subjected to a triaxial state of stress, the resulting strain component in the x-direction, ε_x, is given by the summation of elongation and contraction:

$$\varepsilon_x = \frac{1}{E}\left[\sigma_x - \nu(\sigma_y + \sigma_z)\right] \tag{4.77}$$

and so on for the y- and z-directions (three equations in total).

For an analysis of mechanical structures, we must consider not only compressional and tensile strains but also shear strains. Whereas normal stresses create elongation plus lateral contraction with accompanying volume changes, shear stresses create shape

* Invar, an alloy formed of nickel and iron, has an extremely small coefficient of thermal expansion at room temperature. Invar is often called for in the design of machinery that must be extremely stable.

changes without volume changes (e.g., by twisting a body), that is, shear strains. The 1D shear strain, γ, is produced by the shear stress, τ, and has units of N/m^2. For small strains, Hooke's law may be applied again:

$$\gamma = \frac{\tau}{G} \tag{4.78}$$

where G is called the *elastic shear modulus* or the *modulus of rigidity*. For any 3D state of shear stress, three equations of this type will hold. Isotropic bodies are characterized by two independent elastic constants only because the shear modulus G, it can be shown,[19] relates the Young's modulus and the Poisson ratio as:

$$G = \frac{E}{2(1+\nu)} \tag{4.79}$$

Crystal materials, whose elastic properties are anisotropic, require more than two elastic constants, the number increasing with decreasing symmetry. Cubic crystals, for example, require 3 elastic constants, hexagonal crystals require 5, and materials without symmetry require 21[19,20] (see also Table 2.1). The relation between stresses and strains is more complex in this case and depends greatly on the spatial orientation of these quantities with respect to the crystallographic axes (see Chapter 2 on crystallography). Hooke's law in the most generic form is expressed in two formulas:

$$\sigma_{ij} = E_{ijkl} \cdot \varepsilon_{kl} \text{ and } \varepsilon_{ij} = S_{ijkl} \cdot \sigma_{kl} \tag{4.80}$$

where σ_{ij} and σ_{kl} = stress tensors of rank 2 expressed in N/m^2

ε_{ij} and ε_{kl} = strain tensors of rank 2 and are dimensionless

E_{ijkl} = stiffness coefficient tensor of rank 4 expressed in N/m^2 with at the most $3 \times 3 \times 3 \times 3 = 81$ elements

S_{ijkl} = compliance coefficient tensor of rank 4 expressed in m^2/N with at the most $3 \times 3 \times 3 \times 3 = 81$ elements

The first expression is analogous to Equation 4.75, and the second expression is the inverse, giving the strains in terms of stresses. The tensor representations in Equation 4.80 can also be represented as two matrices:

$$\sigma_m = \sum_{n=1}^{6} E_{mn}\varepsilon_n \text{ and } \varepsilon_m = \sum_{n=1}^{6} S_{mn}\sigma_n \tag{4.81}$$

Components of tensors E_{ijkl} and S_{ijkl} are substituted by elements of the matrices E_{mn} and S_{mn}, respectively. To abbreviate the *ij* indices to *m* and the *kl* indices to *n*, the following scheme applies:

11 → 1, 22 → 2, 33 → 3, 23 and 32 → 4, 13 and 31 → 5, 12 and 21 → 6, $E_{ijkl} \rightarrow E_{mn}$ and $S_{ijkl} \rightarrow S_{mn}$ when m and n = 1, 2, 3; $2S_{ijkl} \rightarrow S_{mn}$ when m or n = 4, 5, 6; $4S_{ijkl} \rightarrow S_{mn}$ when m and n = 4, 5, 6; $\sigma_{ij} \rightarrow \sigma_m$ when m = 1, 2, 3; and $\varepsilon_{ij} \rightarrow \varepsilon_m$ when m = 4, 5, 6.

Thus, with these reduced indices, there are six equations of the type:

$$\sigma_x = E_{11}\varepsilon_x + E_{12}\varepsilon_y + E_{13}\varepsilon_z + E_{14}\gamma_{yz} + E_{15}\gamma_{zx} + E_{16}\gamma_{xy} \tag{4.82}$$

And hence there exist 36 moduli of elasticity or E_{mn} stiffness constants. There are also six equations of the type:

$$\varepsilon_x = S_{11}\sigma_x + S_{12}\sigma_y + S_{13}\sigma_z + S_{14}\tau_{yz} + S_{15}\tau_{zx} + S_{16}\tau_{xy} \tag{4.83}$$

defining 36 S_{mn} constants, which are called the *compliance constants*. It can be shown that the matrices E_{mn} and S_{mn}, each composed of 36 coefficients, are symmetrical ($E_{mn} = E_{nm}$ and $S_{mn} = S_{nm}$); hence a material without symmetrical elements has 21 independent constants or moduli. Because of symmetry of crystals, several more of these may vanish until, for an isotropic medium, they number only two (E and ν). The stiffness coefficient and compliance coefficient matrices for cubic-lattice crystals with the vector of stress oriented along the [100] axis are given as:

$$E_{mn} = \begin{bmatrix} E_{11} & E_{12} & E_{12} & 0 & 0 & 0 \\ E_{12} & E_{11} & E_{12} & 0 & 0 & 0 \\ E_{12} & E_{12} & E_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & E_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & E_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & E_{44} \end{bmatrix} \tag{4.84}$$

$$S_{mn} = \begin{bmatrix} S_{11} & S_{12} & S_{12} & 0 & 0 & 0 \\ S_{12} & S_{11} & S_{12} & 0 & 0 & 0 \\ S_{12} & S_{12} & E_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & S_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & S_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & S_{44} \end{bmatrix} \quad (4.85)$$

In cubic crystals, the three remaining independent elastic moduli are usually chosen as E_{11}, E_{12}, and E_{44}. The S_{mn} values can be calculated simply from the E_{mn} values. Expressed in terms of the compliance constants, one can show that $1/S_{11} = E$ = Young's modulus, $-S_{12}/S_{11} = \nu$ = Poisson's ratio, and $1/S_{44} = G$ = shear modulus. In the case of an isotropic material, such as a metal wire, there is an additional relationship:

$$E_{44} = \frac{E_{11} - E_{12}}{2} \quad (4.86)$$

With this expression the number of independent stiffness constants reduces to two. The anisotropy coefficient α is defined as:

$$\alpha = \frac{2E_{44}}{E_{11} - E_{12}} \quad (4.87)$$

making $\alpha = 1$ for an isotropic crystal. For an anisotropic crystal, the degree of anisotropy is given by the deviation of α from 1. SCS has moderately anisotropic elastic properties,[21,22] with $\alpha = 1.57$, and very anisotropic crystals may have a value larger than 8. Brantley[21] gives the nonzero Si stiffness components, referred to the [100] crystal orientation, as $E_{11} = E_{22} = E_{33} = 166 \times 10^9$ N/m^2, $E_{12} = E_{13} = E_{23} = 64 \times 10^9$ N/m^2, and $E_{44} = E_{55} = E_{66} = 80 \times 10^9$ N/m^2:

$$\begin{vmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \tau_{xy} \\ \tau_{xz} \\ \tau_{yz} \end{vmatrix} = \begin{vmatrix} 166(E_{11}) & 64(E_{12}) & 64(E_{12}) & 0 & 0 & 0 \\ 64(E_{12}) & 166(E_{11}) & 64(E_{12}) & 0 & 0 & 0 \\ 64(E_{12}) & 64(E_{12}) & 166(E_{11}) & 0 & 0 & 0 \\ 0 & 0 & 0 & 80(E_{44}) & 0 & 0 \\ 0 & 0 & 0 & 0 & 80(E_{44}) & 0 \\ 0 & 0 & 0 & 0 & 0 & 80(E_{44}) \end{vmatrix} \times \begin{vmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \\ \gamma_{xy} \\ \gamma_{xz} \\ \gamma_{yz} \end{vmatrix} \quad (4.88)$$

with σ normal stress, τ shear stress, ε normal strain, and γ shear strain. The values for E_{mn} in Equation 4.88 compare with a Young's modulus of 207 GPa for a low carbon steel. Variations on the values of the elastic constants on the order of 30%, depending on crystal orientation, doping level, and dislocation density, must be considered to have effects as well. In Volume II, Chapter 7, we learn that the Young's modulus for polycrystalline silicon is about 161 GPa. From the stiffness coefficients, the compliance coefficients of Si can be calculated as $S_{11} = 7.68 \times 10^{-12}$ m^2/N, $S_{12} = -2.14 \times 10^{-12}$ m^2/N, and $S_{44} = 12.6 \times 10^{-12}$ m^2/N.[23] A graphical representation of elastic constants for different crystallographic directions in Si and Ge is given in Worthman et al.[24] and is reproduced in Figure 4.69a–d displays E and ν for Ge and Si in planes (100) and (110) as functions of direction. Calculations show that E, G, and ν are constant for any direction in the (111) plane. In other words, a plate lying in this plane can be considered as having isotropic elastic properties. A review of independent determinations of the Si stiffness coefficients, with their respective temperature coefficients, is given in Metzger et al.[25] Some of the values from that review are reproduced in Table 4.12. Values for Young's modulus and the shear modulus of Si can also be found in Greenwood[26] and are reproduced in Table 4.13 for the three important crystal orientations.

Residual Stress in Single-Crystal Si

Most mechanical properties, such as the Young's modulus, for lightly and highly doped silicon are identical. Residual stress and associated stress gradients in highly boron-doped single-crystal Si (SCS) present some controversy. Highly boron-doped membranes, which are usually reported to be tensile, have also been reported as compressive.[27,28] From a simple atom-radius argument, a large number of substitutional boron atoms would be expected to create a net shrinkage of the lattice compared with pure silicon and that the residual stress would be tensile with a stress gradient corresponding to the doping gradient. For example, an etched cantilever would be expected to bend up out of the plane of the silicon wafer. Maseeh et al.[28] believe that the appearance of compressive behavior in heavily boron-doped single-crystal layers results from the use of an oxide etch mask. They suggest that plastic deformation of the p^+ silicon beneath the compressively stressed oxide can explain the observed behavior. Ding et al.[29] also

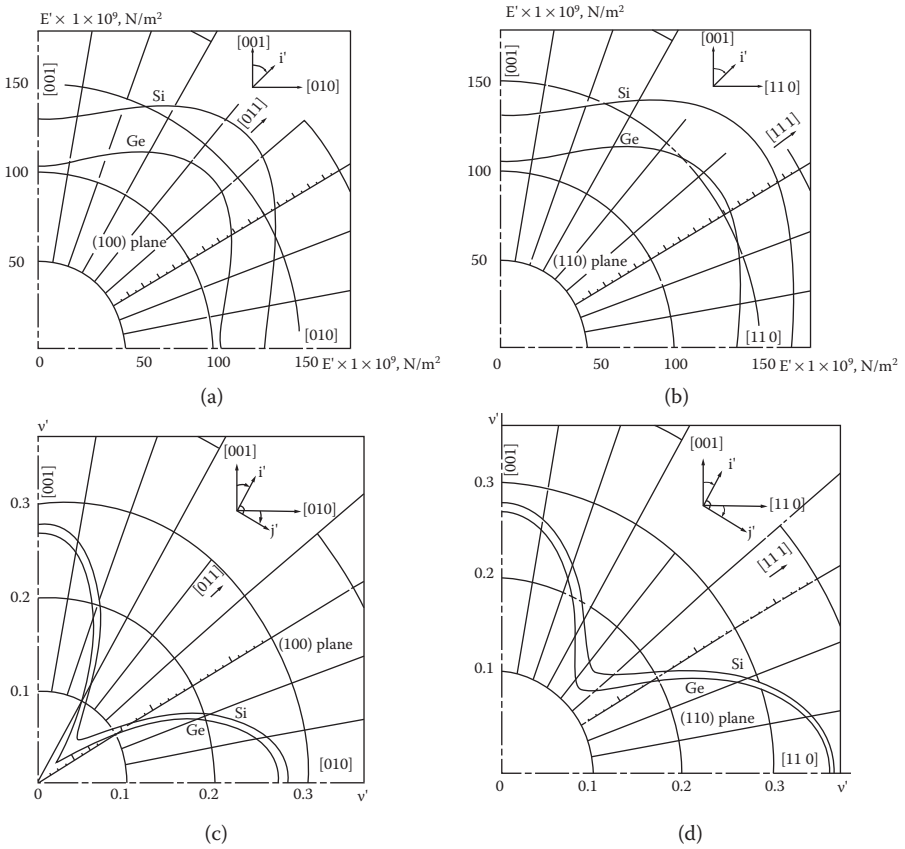


FIGURE 4.69 Elasticity constants for Si and Ge. (a) Young's modulus as a function of direction in the (100) plane. (b) Young's modulus as a function of direction in the (110) plane. (c) Poisson's ratio as a function of direction in the (100) plane. (d) Poisson's ratio as a function of direction in the (110) plane. (From Worthman, J. J., and R. A. Evans. 1965. *J Applied Physics* 36:153–56. With permission.[24])

found compressive behavior for nitride-covered p^+ Si thin membranes. These authors believe that the average stress in p^+ silicon is tensile, but great care is required to establish this fact because the combination of heavy boron doping and a high-temperature drive-in under oxidizing conditions can create an apparent reversal of both the net stress (to compressive) and of the stress gradient (opposite to the doping gradient). A proposed explanation is that, at the oxide/silicon interface, a thin compressively stressed layer is formed during the drive-in, which is not removed in buffered HF. It can be removed by reoxidation and etching in HF, or by etching in KOH.

The question of residual stress in polycrystalline Si is a heavily studied topic as we shall discover in Volume II, Chapter 7 on surface micromachining. Stress in thin polycrystalline films has dramatic effects on the reliable operation of surface micromachined mechanical elements (see also Volume II, Equation 7.29).

TABLE 4.12 Stiffness Coefficients and Temperature Coefficients of Stiffness for Si

Stiffness Coefficients in GPa (value in GPa = 10^9 N/m^2)	Temperature Coefficient of Young's Modulus (10^{-6} K^{-1})
$E_{11} = 164.8 \pm 0.16$	−122
$E_{12} = 63.5 \pm 0.3$	−162
$E_{44} = 79.0 \pm 0.06$	−97

Source: Metzger, H., and F.R. Kessler. 1970. Der Debye-Sears Effect zur Bestimmung der Elastischen Konstanten von Silicium. *Z Naturf A* 25:904.

Yield, Tensile Strength, Hardness, and Creep

As a material is deformed beyond its elastic limit, yielding or plastic deformation (permanent, nonrecoverable deformation) occurs. The point of yielding in Figure 4.67a is the point of initial departure from linearity of the stress-strain curve and is sometimes called the *proportional* limit indicated by a letter *P*. Young's modulus of mild steel is ±30,000,000 psi, and its proportional limit (highest stress in the

Table 4.13 Derived Values for Young's Modulus and Shear Modulus for Si

Crystal Orientation	Young's Modulus E (GPa)	Shear Modulus G (GPa)	Temperature Coefficient (10^{-6} K^{-1})
[100]	129.5	79.0	–63
[110]	168.0	61.7	–80
[111]	186.5	57.5	–46

Source: Greenwood, J.C. 1988. Silicon in mechanical sensors. *J Phys E Sci Instrum* 21:1114.

elastic range) is approximately 30,000 psi.* Thus, the maximum elastic strain in mild steel is about 0.001 under a condition of uniaxial stress. This gives an idea as to the magnitude of the strains we are dealing with. A convention has been established wherein a straight line is constructed parallel to the elastic portion of the stress-strain curve at some specified strain offset, usually 0.002. The stress corresponding to the intersection of this line and the stress-strain curve as it bends over in the plastic region is defined as the yield strength, σ_y (see Figure 4.67a). The magnitude of the yield strength of a material is a measure of its resistance to plastic deformation. Yield strengths may range from 35 MPa (5,000 psi) for a soft and weak aluminum to more than 1,400 MPa (200,000 psi) for high-strength steels. The tensile strength is the stress at the peak of the stress-strain curve. This corresponds to the maximum stress that can be sustained by a structure in tension; if the stress is applied and maintained, fracture will result.

Crystalline silicon is a hard and brittle material, deforming elastically until it reaches its yield strength, at which point it breaks. For Si, the yield strength is 7 GPa, equivalent to a 700-kg weight suspended from a 1-mm^2 area. Both tensile strength and hardness are indicators of a metal's resistance to plastic deformation. Consequently, they are roughly proportional.[30] Material deformation occurring at elevated temperatures and static material stresses is termed *creep*. It is defined as a time-dependent and permanent deformation of materials when subjected to a constant load or stress.

Silicon exhibits no plastic deformation or creep below 500°C; therefore, Si sensors are very insensitive to fatigue failure when subjected to high cyclic loads. Silicon sensors have been cycled in excess of 100 million cycles with no observed failures. This ability to survive a very large number of duty cycles is because there is no energy-absorbing or heat-generating mechanism caused by intergranular slip or movement of dislocations in silicon at room temperature. However, single-crystal Si (SCS), as a brittle material, will yield catastrophically when stress beyond the yield limit is applied rather than deform plastically as metals do (see Figure 4.67b). At room temperature, high modulus materials, such as Si, SiO_2, and Si_3N_4, often exhibit linear-elastic behavior at lower strain and transition abruptly to brittle-fracture behavior at higher strain. Plastic deformation in metals is based on stress-induced dislocation generation in the grain boundaries and a subsequent dislocation migration that results in a macroscopic deformation from intergrain shifts in the material. No grain boundaries exist in SCS, and plastic deformation can only occur through migration of the defects originally present in the lattice or of those generated at the surface. As the number of these defects is very low in SCS, the material can be considered a perfect elastic material at normal temperatures. Perfect elasticity implies proportionality between stress and strain (i.e., load and flexure) and the absence of irreversibilities or mechanical hysteresis. The absence of plastic behavior also accounts for the extremely low mechanical losses in SCS, which enable the fabrication of resonating structures that exhibit exceptionally high Q-factors. Values of up to 10^8 in vacuum have been reported. At elevated temperatures, and with metals and polymers at ordinary temperatures, complex behavior in the stress-strain curve can occur. Considerable plasticity can be induced in SCS at elevated temperatures (>800°C), when silicon softens appreciably and the mobility of defects in the lattice is substantially increased. Huff

* In the sensor area it is still mandatory to be versatile in the different unit systems especially with regard to pressure and stress. In this book we are using mostly Pascal, Pa (=N/m^2), but in industry it is still customary to use psi when dealing with metal properties, Torr when dealing with vacuum systems, and dyne/cm^2 when dealing with surface tension.

and Schmidt[27] report a pressure switch exhibiting hysteresis based on buckling of plastically deformed silicon membranes. To eliminate plastic deformation of Si wafers, it is important that during high-temperature steps the presence of films that could stress or even warp the wafer in an asymmetric way, typically oxides or nitrides, be avoided.

Piezoresistivity in Silicon

Piezoresistance is the fractional change in bulk resistivity induced by small mechanical stresses applied to a material. Lord Kelvin discovered the effect in 1856. Most materials exhibit piezoresistivity, but the effect, Smith found in 1954, is particularly important in some semiconductors (more than an order of magnitude higher than that of metals).[31] Monocrystalline silicon has a high piezoresistivity and, combined with its excellent mechanical and electronic properties, makes a superb material for the conversion of mechanical deformation into an electrical signal. The history of silicon-based mechanical sensors started with the discovery of the piezoresistance effect in Si (and Ge) more than four decades ago.[31] The piezoresistive effect in semiconductor materials originates in the deformation of the energy bands as a result of the applied stress. The deformed bands change the effective mass and mobility of the charge carriers (electrons and holes), hence modifying the resistivity. The two main classes of piezoresistive semiconductor sensors are membrane-type structures (typically pressure and flow sensors) and cantilever beams (typically acceleration sensors) with in-diffused resistors (boron, arsenic, or phosphorus) strategically placed in zones of maximum stress (Figure 4.70).

In Chapter 2 we saw how in a 3D anisotropic crystal, the electrical field vector (**E**) is related to the current vector, **J**, by a 3 × 3 resistivity tensor (see Equation 2.4).[23] Experimentally, the nine coefficients always reduce to six, and the symmetric tensor, is given by:

$$\begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \end{bmatrix} = \begin{bmatrix} \rho_1 & \rho_6 & \rho_5 \\ \rho_6 & \rho_2 & \rho_4 \\ \rho_5 & \rho_4 & \rho_3 \end{bmatrix} \begin{bmatrix} \mathbf{J}_1 \\ \mathbf{J}_2 \\ \mathbf{J}_3 \end{bmatrix} \qquad (4.89)$$

Electric Field — Resistivity Tensor — Curren

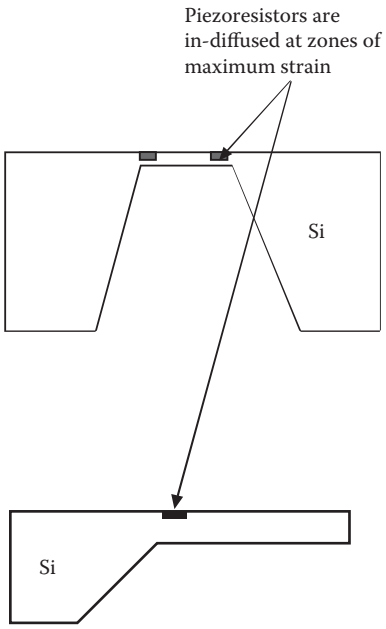


FIGURE 4.70 Si membrane and Si cantilever with in-diffused resistors.

For the cubic Si lattice, with the axes aligned with the <100> directions, ρ_1, ρ_2, and ρ_3 define the dependence of the electric field on the current along the same direction (one of the <100> directions). The cross-sensitivities, ρ_4, ρ_5, and ρ_6, relate the electric field to the current along a perpendicular direction.

The six resistivity components in Equation 4.89 depend on the normal (σ) and shear (τ) stresses in the material as defined earlier in this chapter. This can be represented as:

$$\rho = \rho_{\text{no stress}} + \Delta\rho(\sigma,\tau) \text{ or } \sigma_c\left(=\frac{1}{\rho}\right) = \sigma_{\text{no stress}} + \Delta\sigma_c(\sigma,\tau) \qquad (4.90)$$

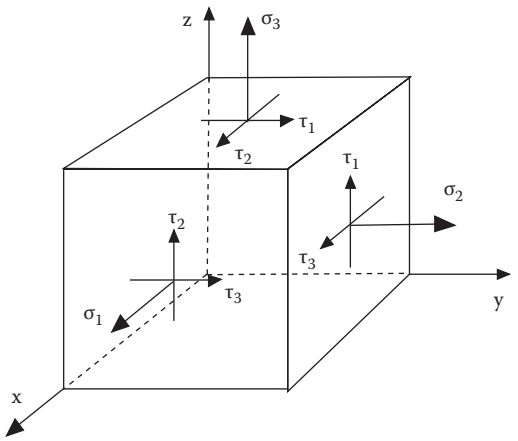


FIGURE 4.71 Piezoresistance in silicon.

where we use σ_c for conductivity to distinguish it from σ for the normal stress. For cubic crystals this can be written out as (see also Figure 4.71):

$$\begin{bmatrix} \Delta\sigma_{c,1} \\ \Delta\sigma_{c,2} \\ \Delta\sigma_{c,3} \\ \Delta\sigma_{c,4} \\ \Delta\sigma_{c,5} \\ \Delta\sigma_{c,6} \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{12} & 0 & 0 & 0 \\ \pi_{12} & \pi_{11} & \pi_{12} & 0 & 0 & 0 \\ \pi_{12} & \pi_{12} & \pi_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & \pi_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & \pi_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \pi_{44} \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} \tag{4.91}$$

Substituting the results of Equation 4.91 in Equation 4.89 one obtains the following three expressions for a large piece of bulk silicon:

$$\begin{aligned} E_1 &= \rho J_1 + \rho\pi_{11}\sigma_1 J_1 + \rho\pi_{12}(\sigma_2+\sigma_3)J_1 + \rho\pi_{44}(J_2\tau_3 + J_3\tau_2) \\ E_2 &= \rho J_2 + \rho\pi_{11}\sigma_2 J_2 + \rho\pi_{12}(\sigma_1+\sigma_3)J_2 + \rho\pi_{44}(J_1\tau_3 + J_3\tau_1) \\ E_3 &= \rho J_3 + \rho\pi_{11}\sigma_3 J_3 + \rho\pi_{12}(\sigma_1+\sigma_2)J_3 + \rho\pi_{44}(J_1\tau_2 + J_2\tau_1) \end{aligned} \tag{4.92}$$

For devices with finite dimensions, the influence of dimensional changes needs to be added through the Poisson ratio.

Smith[31] was the first to measure the resistivity coefficients for Si at room temperature. The coefficients are dependent on crystal orientation, temperature, and dopant concentration. Table 4.14 lists Smith's results for the three remaining independent resistivity coefficients, that is, π_{11}, π_{12}, and π_{44} of [100]-oriented Si at room temperature.[31] The piezoresistance coefficients are largest for π_{11} in n-type silicon and π_{44} in p-type silicon, about –102.2 10^{-11} and 138 10^{-11} Pa^{-1}, respectively. The piezoresistivity coefficients are related to the gauge factor (G_f) by the Young's modulus. The gauge factor, G_f, is the relative resistance change divided by the applied strain, or:

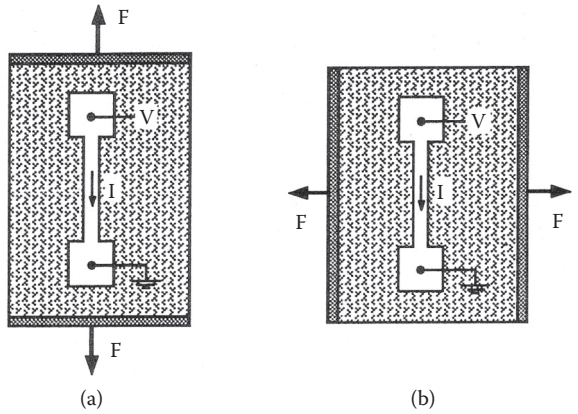


FIGURE 4.72 Longitudinal (a) and transverse (b) piezoresistors.

$$G_f = \frac{1}{\varepsilon}\frac{\Delta R}{R} \tag{4.93}$$

The gauge factor of a metal strain gauge is typically around 2; for single-crystal Si (SCS) it is 90; and for polycrystalline and amorphous Si, it is between 20 and 40.

The contribution to resistance changes from stresses that are longitudinal (σ_l) and transverse (σ_t) with respect to the current flow is given by (Figure 4.72):

$$\frac{\Delta R}{R} = \sigma_l\pi_l + \sigma_t\pi_t \tag{4.94}$$

where σ_l = longitudinal stress component, i.e., stress component parallel to the direction of the current
σ_t = transversal stress component, i.e., stress component perpendicular to the direction of the current
π_l = longitudinal piezoresistance coefficient
π_t = transversal piezoresistance coefficient

The piezoresistance coefficients π_l and π_t for (100) silicon as a function of crystal orientation are reproduced from Kanda in Figure 4.73a for

TABLE 4.14 Resistivity and Piezoresistivity Coefficients at Room Temperature, <100> Si Wafers and Doping Levels below 10^{18} cm^{-3} in 10^{-11} Pa^{-1}

	ρ (Ωcm)	Direction	π_{11}	π_{12}	π_{44}	π_t	π_l
p-Si	7.8	<100>				0	0
		<110>	+6.6	–1.1	+138.1	–66.3	71.8
n-Si	11.7	<100>				+ 53.4	–102.2
		<110>	–102.2	+53.4	–13.6	–17.6	–31.2

Source: Smith, C.S. 1954. Piezoresistance effect in germanium and silicon. *Phys Rev* 94:42; and Khazan, A.D. 1994. *Transducers and their elements.* Englewood Cliffs, NJ: PTR Prentice Hall.

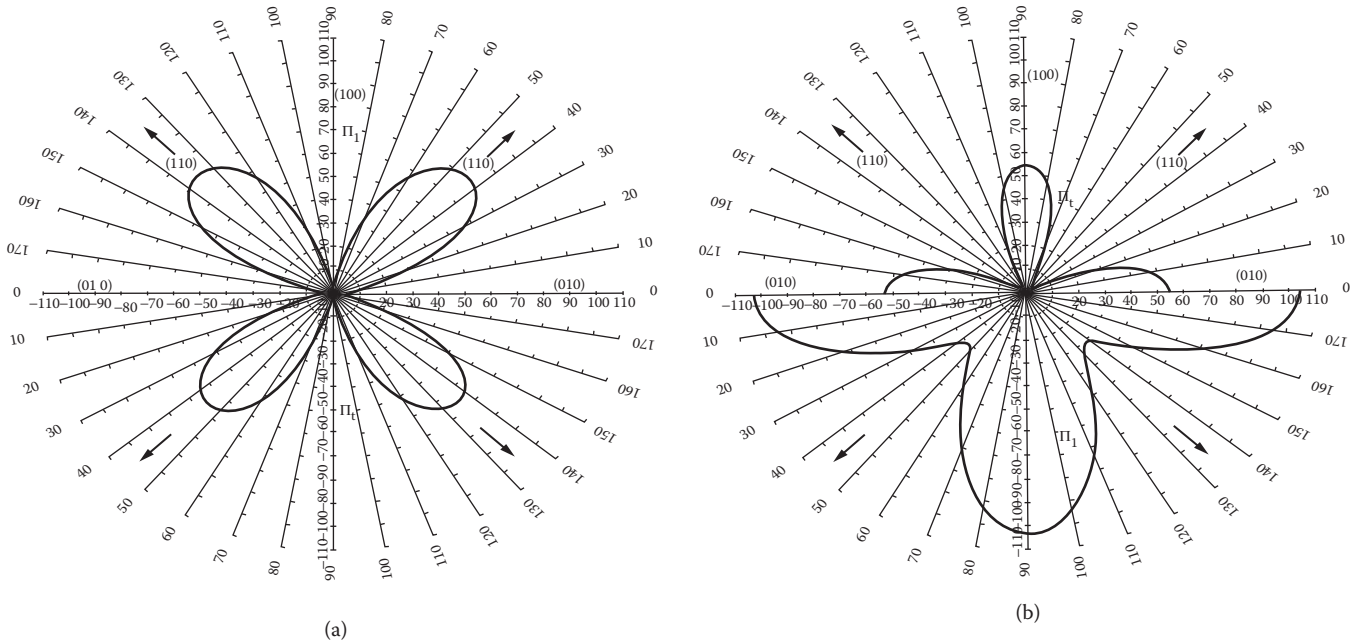


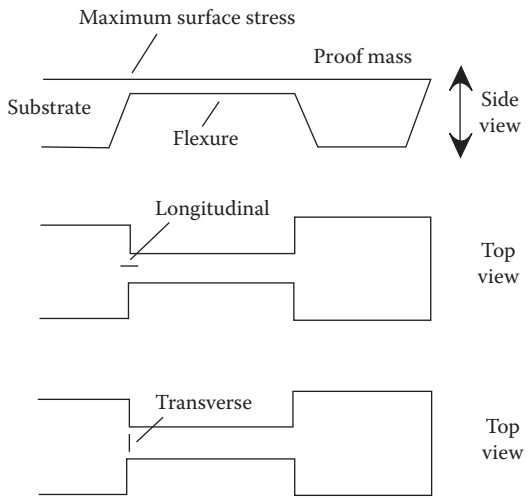
FIGURE 4.73 Piezoresistance coefficients π_1 and π_t for (100) silicon in the (001) plane in 10^{-12} cm^2 dyne^{-1} or 10^{-11} Pa^{-1}. (a) For p-type. (b) For n-type. (From Kanda, Y. 1982. *IEEE Trans Electr Dev* 29:64. With permission.)

p-type and Figure 4.73b for n-type.[32] For lightly doped silicon (n- or p-type < 10^{18} cm^{-3}), the temperature coefficient of resistance for π_1 and π_t is approximately 0.25%/°C. It decreases with dopant concentration to about 0.1%/°C at 8×10^{19} cm^{-3}. By maximizing the expression for the stress-induced resistance change in Equation 4.94, the achievable sensitivity in a piezoresistive silicon sensor is optimized. From Figure 4.73, the maximum longitudinal piezoresistive coefficient for p-type Si is $\pi_{l,[111]} = 93.5 \times 10^{-11}$ Pa^{-1} and for n-type Si it is $\pi_{l,[100]} = -102.2\ 10^{-11}$ Pa^{-1} (this latter value is also found in Table 4.14).

The orientation of a sensing membrane or beam (Figure 4.74) is determined by its anisotropic fabrication.

In p-type Si the piezoresistive coefficients are opposite in sign and comparable in magnitude, making them ideal for Wheatstone bridge detection (see Figure 4.75).

The surface of the silicon wafer is usually a (100) plane; the edges of the etched structures are intersections of (100) and (111) planes and are thus <110> directions. For pressure sensing, p-type piezoresistors are most commonly used. This is because the orientation of maximum piezoresistivity (<110>)



$\pi_1 = -31.2 \cdot 10^{-11}$ Pa^{-1}	$\pi_1 = -17.6 \cdot 10^{-11}$ Pa^{-1}
$\pi_1 = 71.8 \cdot 10^{-11}$ Pa^{-1}	$\pi_1 = -66.3 \cdot 10^{-11}$ Pa^{-1}

FIGURE 4.74 Bulk micromachined cantilever with longitudinal and transverse piezoresistors.

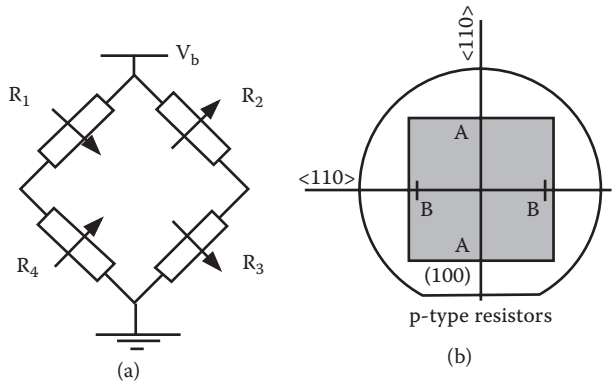


FIGURE 4.75 Measuring membrane resistance with piezoresistors. (a) Wheatstone bridge configuration of four in-diffused piezoresistors. The arrows indicate resistance changes when the membrane is bent downward (into the page). (b) Maximizing the piezoresistive effect with p-type resistors. The A resistors are stressed longitudinally and the B resistors are stressed transversally. (Based on Peeters, E. 1994. *Process development for 3D silicon microstructures, with application to mechanical sensor design.* PhD diss., Katholieke Universiteit Leuven, Leuven, Belgium. With permission;[2] and Maluf, N. 2000. *An introduction to microelectromechanical systems engineering.* Boston: Artech House.[34])

happens to coincide with the edge orientation of a conventionally etched diaphragm and because the longitudinal coefficient is roughly equal in magnitude but opposite in sign as compared with the transverse coefficient (see Figures 4.73a and 4.75b and Table 4.14).[2] Piezoresistors oriented at 45° with respect to the primary flat, that is, in the <100> direction, are insensitive to applied stress, which provides an inexpensive way to incorporate stress-independent diffused temperature sensors. Longitudinal and transverse piezoresistance coefficients are related to principal piezoresistance coefficients through coordinate transformation. With the values in Table 4.14, π_l and π_t can be calculated numerically for other primary orientations. The longitudinal piezoresistive coefficient in the <110> direction is $\pi_l = 1/2(\pi_{11} + \pi_{12} + \pi_{44})$. The corresponding transverse coefficient is $\pi_t = 1/2(\pi_{11} + \pi_{12} - \pi_{44})$. From Table 4.14, we know that for p-type resistors π_{44} is more important than the other two coefficients, and Equation 4.94 is approximated by:

$$\frac{\Delta R}{R} = \frac{\pi_{44}}{2}(\sigma_l - \sigma_t) \qquad (4.95)$$

For n-type resistors, π_{44} can be neglected, and we obtain:

$$\frac{\Delta R}{R} = \frac{\pi_{11} + \pi_{12}}{2}(\sigma_l + \sigma_t) \qquad (4.96)$$

Equations 4.95 and 4.96 are valid only for uniform stress fields or if the resistor dimensions are small compared with the membrane or beam size. When stresses vary over the resistors, they have to be integrated, which is conveniently done by computer simulation programs. More details on the underlying physics of piezoresistivity and its dependence on crystal orientation can be found in Kanda[32] and Middlehoek and Audet.[33]

To convert the piezoresistive effect into a measurable electrical signal, a Wheatstone bridge is often used. A balanced Wheatstone bridge configuration is constructed as in Figure 4.75a by locating four p-piezoresistors midway along the edges of a square diaphragm as in Figure 4.75b (location of maximum stress). Two resistors are oriented so that they sense stress in the direction of their current axes, and two are placed to sense stress perpendicular to their current flow. Two longitudinally stressed resistors (A) are balanced against two transversally stressed resistors (B); two of them increase in value, and the other two decrease in value on application of a stress. In this case, from Equation 4.95:

$$\frac{\Delta R}{R} \approx 70 \cdot 10^{-11}(\sigma_l - \sigma_t) \qquad (4.97)$$

with σ in Pa. For a realistic stress pattern where σ_l = 10 MPa and σ_t = 50 MPa, Equation 4.97 gives us $\Delta R/R \approx 2.8\%$.[2] Thus, resistance change can be calculated as a function of the membrane or cantilever beam stress.

By varying the diameter and thickness of the silicon diaphragms, piezoresistive sensors in the range of 0–200 MPa have been made. The bridge voltages are usually between 5 and 10 V, and the sensitivity may vary from 10 mV/kPa for low-pressure to 0.001 mV/kPa for high-pressure sensors.

A schematic illustration of a pressure sensor with diffused piezoresistive sense elements is shown in Figure 4.76. In this case, p-type piezoresistors are diffused into a thin epitaxial n-type Si layer on a p-type Si substrate.[34]

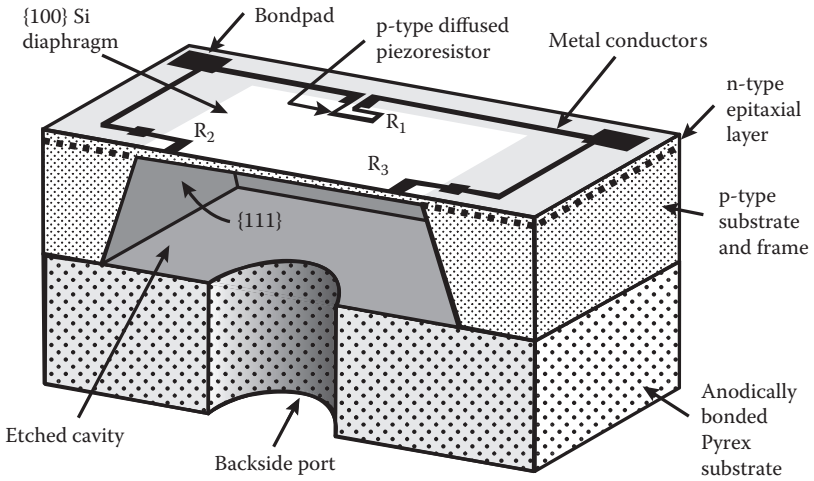


FIGURE 4.76 Schematic illustration of a pressure sensor with p-type diffused resistor in an n-type epitaxial layer. (Based on Maluf, N. 2000. *An introduction to microelectromechanical systems engineering.* Boston: Artech House.[34])

Peeters[2] shows how a more sensitive device may be based on n-type resistors when all the n-type resistors oriented along the <100> direction are subjected to a uniaxial stress pattern in the longitudinal axis, as shown in Figure 4.77. The overall maximum piezoresistivity coefficient (π_1 in the <100> direction) is substantially higher for n-type silicon than it is for p-type silicon in any direction (maximum π_t and π_l in the <100> direction, Figure 4.73b). Exploitation of these high piezoresistivity coefficients is less obvious, though, because the resistor orientation for maximum sensitivity (<100>) is rotated over 45° with respect to the <110> edges of an anisotropically etched diaphragm. Also evident from Figure 4.73b is a transversally stressed resistor cannot be balanced against a longitudinally stressed resistor. Peeters has circumvented these two objections with a uniaxial, longitudinal stress pattern in the rectangular diaphragm represented in Figure 4.77. With a (100) substrate and a <100> orientation (45° to wafer flat), we obtain:

$$\frac{\Delta R}{R} \approx 53.10^{-11} \times \sigma_t - 102.10^{-11} \times \sigma_l \qquad (4.98)$$

with σ in Pa. Based on Equation 4.98, with σ_t = 10 MPa and σ_l = 50 MPa, ΔR/R ≈ –4.6%. In the proposed stress pattern it is important to minimize the transverse stress by making the device truly uniaxial as the longitudinal and transverse stress components have opposite effects and can even cancel out one another. In practice, a pressure sensor with an estimated 65% gain in pressure sensitivity over the more traditional configurations could be made in the case of 80% uniaxiality.[2]

Silicon as a Mechanical MEMS Material: Summary

Here we want to refine decision criteria about the choice of an "active substrate" for mechanical micromachining applications in which the substrate is involved in determining the mechanical performance of the device crafted from/on it. Mechanical stability is crucial for mechanical sensing applications. Any sensing device must be free of drift to avoid recalibration at regular intervals. Part of the

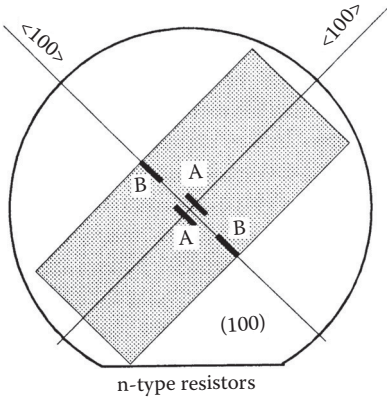


FIGURE 4.77 Higher pressure sensitivity by strategic placement of in-diffused piezoresistors proposed by Peeters.[2] The n-type resistors are stressed longitudinally with the A resistors under tensile stress and the B resistors under compressive stress. (From Peeters, E. 1994. *Process development for 3D silicon microstructures, with application to mechanical sensor design.* PhD diss., Katholieke Universiteit Leuven, Leuven, Belgium. With permission.[2])

drift in mechanical sensors may be associated with movement of crystal dislocations in the loaded mechanical part. In ductile materials, such as metals, dislocations move readily. In contrast, brittle materials, such as semiconductors, have dislocations that hardly move. Mechanical engineers often avoid using brittle materials and opt for ductile materials, even though these plastically deform, meaning that they are subject to mechanical hysteresis. The lack of plastic deformation coupled with extremely high yield strength—comparable with steel—makes Si a material superior to any metal in most applications. As a consequence, Si has been quite successful as a structural element in mechanical sensors, particularly during the past 25 years. Pressure and acceleration sensors, based on simple piezoresistive elements embedded in a Si movable mechanical member, have turned into major commercial applications. Besides the desirable mechanical and known electrical properties, this success can also be attributed to the available fabrication technology of ICs.

The significant properties that have made Si a successful material not only for electronic applications but also for mechanical applications are reviewed in Tables 4.15 and 4.16. From these tables we summarize some of the major reasons behind the success of Si as a mechanical sensor element:

- Silicon surpasses stainless steel in yield strength while displaying a density lower than that of aluminum. In fact, Si's specific strength, defined as the ratio of yield strength to density, is significantly higher than for most common engineering materials.
- The hardness of Si is slightly better than that of stainless steel; it approaches that of quartz and is higher than most common glasses.
- The Young's modulus of Si has a value approaching that of stainless steel and is well above that of quartz. From Tables 4.15 and 4.16 we also note that Si_3N_4, a coating often used on silicon, has a hardness topped only by a material such as diamond. Therefore, the combination of silicon with silicon nitride coatings can be used for highly wear-resistant components as required in micromechanisms such as micromotors (see Volume II, Chapter 7).

Table 4.15 Mechanical Properties of Single-Crystal Silicon (SCS) among Other Important Technological Materials

	Yield Strength (10^9 N/m^2= GPa)	Specific Strength (10^3 m^2 s^{22})	Knoop Hardness (kg/mm^2)	Young's Modulus (10^9 N/m^2= GPa)	Density (10^3 kg/m^3)	Thermal Conductivity at 300 K (W/cmK)	Thermal Expansion (10^{26}/°C)
Diamond (SC)	53	15,000	7,000	10.35	3.5	20	1.0
Si (SCS)	2.8–6.8	3,040	850–1,100	190 (111)	2.32	1.56	2.616
GaAs (SC)	2.0			0.75	5.3	0.81	6.0
Si_3N_4	14	4,510	3,486	323	3.1	0.19	2.8
SiO_2 (fibers)	8.4		820	73	2.5	0.014	0.4–0.55
SiC (6H-SiC)	21	6,560	2,480	448	3.2	5	4.2
Iron	12.6		400	196	7.8	0.803	12
Tungsten (W)	4	210	485	410	19.3	1.78	4.5
Al	0.17	75	130	70	2.7	2.36	25
AlN	16			340		1.60	4.0
Al_2O_3	15.4		2,100	275	4.0	0.5	5.4–8.7
Stainless steel	0.5–1.5		660	206–235	7.9–8.2	0.329	17.3
Quartz values // Z ⊥ Z	9		850	107	2.65	0.014	7.1 13.2 (increases with T)
Polysilicon	1.8 (annealed)			161			2.8

SC, single crystal; SCS, single-crystal silicon; T, temperature; // Z, parallel with *z*-axis; ⊥Z, perpendicular to *z*-axis.

Table 4.16 Silicon Single Crystal Material Characteristics

Si Parameter	Value and Comment
Atomic weight	28.1
Atoms/cm^3	5×10^{22}
Bandgap at 300 K	1.12 eV. Si has a high bandgap, making it useful electrically at high temperatures. Indirect bandgap in the near infrared. It is opaque to ultraviolet and transparent to IR.
Chemical resistance	High. Si is resistant to most acids, except combinations of HF/HNO_3 and certain bases.
Density (g/cm^3)	2.4. Si has a lower density than aluminum (2.7).
Dielectric constant	11.9 vs. 13.1 for GaAs
Dielectric strength (V/cm 10^6)	3
Dislocation density	<100 cm^2. IC-grade Si contains virtually no imperfections; thus, it is relatively insensitive to cycling and fatigue failure.
Electron mobility (cm^2/Vs)	1500
Hole mobility (cm^2/Vs)	400
Intrinsic carrier concentration (cm^{-3})	1.45×10^{10}
Intrinsic resistivity (Ωcm)	2.3×10^5 vs. 10^8 for GaAs
Knoop hardness (kg/mm^2)	850 (stainless steel is 820). Si is harder than steel and can readily be coated with silicon nitride, providing high abrasion resistance.
Lattice constant (Å)	5.43
Linear coefficient of thermal expansion at 300 K (10^{-6}/°C)	2.6. The low expansion coefficient of Si is closer to quartz than to metal, making it insensitive to thermal shock.
Melting point	1415°C. Silicon is a high-melting material, making it suitable for high-temperature applications.
Minority carrier lifetime(s)	2.5×10^{-3}
Oxide growth	Si grows a dense, strong, chemically resistant, passivating layer of SiO_2. This oxide is an excellent thermal insulator with a low expansion coefficient.
Poisson ratio	0.22
Relative permittivity	11.8
Silicon nitride	A typical coating for Si with a hardness and wear resistance only topped by diamond.
Specific heat at 300 K (J/gK)	0.713
Thermal conductivity at 300 K (W/cmK)	1.56. Si has a high thermal conductivity comparable to metals such as carbon steel (0.97) and Al (2.36).
Temperature coefficient of Young's modulus (10^{-6} K^{-1}) at 300 K	–90
Temperature coefficient of piezoresistance (10^{-6} K^{-1}) at 300 K (doping < 10^{18} cm^{-3})	–2500
Temperature coefficient of permittivity (10^{-6} K^{-1}) at 300 K	1000
Thermal diffusivity (cm^2/s)	0.9
Yield strength (GPa)	7 (steel is 2.1). IC-grade Si is stronger than steel.
Young's modulus E (GPa)	190 [111] direction. The elastic modulus is similar to that of steel (steel is 200).

Other Si Sensor Properties

Thermal Properties of Silicon

In Figure 4.78, the thermal expansion coefficient of Si, W, SiO_2, Ni-Co-Fe alloy, and Pyrex is plotted versus absolute temperature. Single-crystal Si (SCS) has a high thermal conductivity (comparable with metals such as steel and aluminum and approximately 100 times larger than that of glass) and a low thermal expansion coefficient. Because of its thermal conductivity, Si is used in some devices as an efficient heat sink. Its thermal expansion coefficient is closely matched to Pyrex glass but exhibits considerable temperature dependence. A good match in thermal expansion coefficient between the device wafer and a support substrate is often required in sensor manufacture. A poor match introduces stress, which degrades the device performance. This makes it difficult to fabricate composite structures of Pyrex and Si that are stress-free over a wide range of temperatures. Drift in silicon sensors often stems from packaging. In this respect, several types of stress relief, subassemblies for stress-free mounting of the active silicon parts, play a major role; using silicon as the support for silicon sensors is highly desirable. The latter aspect will be addressed when discussing anodic bonding of Pyrex glass to Si and fusion bonding of Si to Si in Volume III, Chapter 4 on packaging, assembly, and self-assembly.

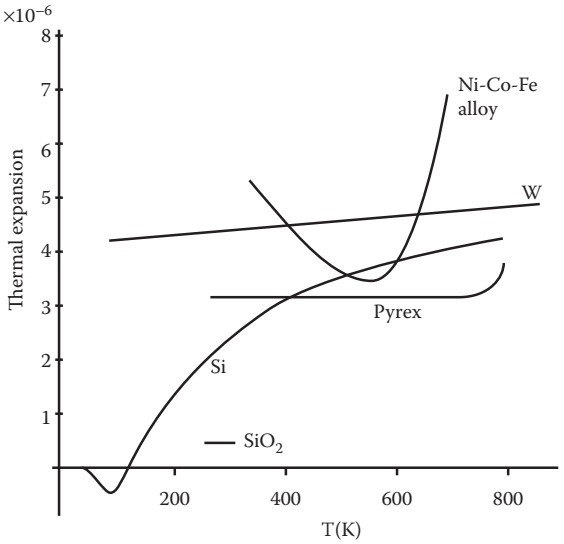


FIGURE 4.78 Thermal expansion coefficient versus absolute temperature. (From Greenwood, J.C. 1988. *J Phys E Sci Instrum* 21:1114. With permission.[26])

Although the Si bandgap is relatively narrow, by using SOI wafers, high-temperature sensors can be fashioned. For the high-temperature applications, highly doped Si, which is relatively linear in its temperature coefficient of resistance and sensitivity over a wide range, is typically used.

When fabricating thermally isolated structures on Si, such as the miniature heater element in Volume II, Figure 9.11, the large thermal conductivity of Si poses a considerable problem. For thermally isolated structures, machining in glass, quartz, or ceramics with their lower thermal conductivity represents an important alternative.

Silicon Optical Properties

Silicon features an indirect bandgap and is not an active optical material (also see Chapters 3 and 5). As a consequence, silicon-based lasers do not exist. Silicon is effective only in detecting light because the indirect bandgap makes emission of light difficult. Above 1.1 μm, silicon is transparent, but at wavelengths shorter than 0.4 μm (in the blue and ultraviolet portions of the visible spectrum), it reflects more than 60% of the incident light. One of the most established applications for silicon sensors, although not often typified as MEMS, is visual imaging with charge-coupled devices (CCDs). In a CCD imager, each element in a 2D array generates an electrical charge in proportion to the amount of light it receives. The charge, stored by CCD elements along a row, is subsequently transferred to the next element in a bucket-brigade fashion as the light input is read out line by line. The number of picture elements (pixels) on the CCD determines the resolution. In top-of-the-line CCD cameras (http://www.pctechguide.com/19digcam.htm) a 3.34-megapixel CCD is capable of delivering a maximum image size of 2048 × 1536 pixels. Si as the pixel semiconductor can be used for a wide variety of electromagnetic radiation wavelengths, from gamma rays to infrared. There is a trend now to configure pixels in clever ways to make novel optical sensors feasible.

One example of a smart pixel configuration is embodied in the retina chip shown in Figure 4.79.[35] This is an IC chip that works like the human retina, selecting only the necessary information from

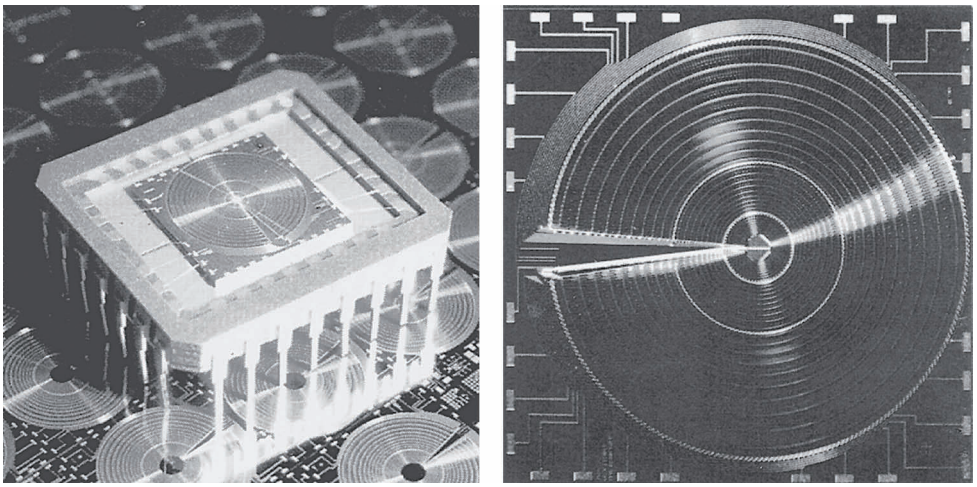


FIGURE 4.79 Photos of the retina sensor. (Courtesy of Dr. Lou Hermans, IMEC, Belgium.)

a presented image to greatly speed up image processing. The chip features 30 concentric circles with 64 pixels each. The pixels increase in size from 30 × 30 μm for the inner circle to 412 × 412 μm for the outer circle. The circle radius increases exponentially with eccentricity. The center of the chip, called the *fovea*, is filled with 104 pixels measuring 30 × 30 μm placed in an orthogonal pattern. The total chip area is 11 × 11 mm. The chip is designed for those applications where real-time coordination between the sensory perception and the system control is of the primary concern. The main application is active vision, and its potential application is expected in robot navigation, surveillance, recognition, and tracking. The system can cover a wide field of view with a relatively low number of pixels without sacrificing the overall resolution and leads to a significant reduction in required image-processing hardware and calculation time. The fast, but rather insensitive, large pixels on the rim of the retina chip pick up a sudden movement in the scenery (peripheral vision), prompting the robot equipped with this "eye" to redirect itself in the direction of the movement to better focus on the moving object with the more sensitive fovea pixels.

The area where MEMS saw one of its biggest commercial breakthroughs is in micromachined mirrors for optical switching in both fiberoptic communications and in data storage applications. Optical switches are to optical communications what transistors are to electronic signaling. What makes single-crystal Si attractive in this case is the optical quality of the Si surface. The quality of the mirror surface is of primary importance to obtain very low insertion loss even after multiple reflections. Wet bulk micromachining has an advantage here over deep reactive-ion etching (DRIE), as the latter leaves Si mirror surfaces with more losses as a result of the inevitable ripples (Volume II, Chapter 3 on dry etching). Important commercial developments in this field by companies such as Quinta Corporation, MEMX, Corning's IntelliSense, Nortel (Xros), and Lucent are summarized in Volume III, Chapter 10. We will see that, after causing a huge stock market technology bubble in 2001, this technology is slowly re-entering the market with more realistic expectations.

Biocompatibility and Si-Based MEMS

Mechanical and Chemical Components of Biocompatibility The definition and history of biocompatibility and the biocompatibility of selected materials are reviewed in Volume III, Chapter 5. Here we cover biocompatibility of Si only. Biocompatibility has both a mechanical and a chemical component. For example, it is well known that, for better biocompatibility, a smooth surface and rounded corners are preferable to a rough surface, sharp corners, and small crevices but also that for a given surface finish and shape, carbon is preferable to copper. The research community remains somewhat divided on the relative contribution of mechanical effects (e.g., surface roughness and shape) versus chemical effects (e.g., Cu vs. C) to the overall bioresponse

of an implant. von Recum's work demonstrates that micromachining (microtexturing) of surfaces of different materials may help deconvolute these relative contributions.[36] Some specific examples of substrate microtexturing effects follow. An increase in biosynthetic activity and mobility of bone cells on polymer films cast from micromachined molds has been identified (this might have applications for bone regeneration and bioengineering of prosthetics). The regeneration of severed tendons only on textured surfaces has been demonstrated and might lead to the use of textured bandages to accelerate healing of tissue. Many examples in the MEMS literature also illustrate directed growth of nerve cells on chemically patterned surfaces. In a typical procedure, a hydrophobic silane pattern on a glass slide is produced by "lift-off" where the hydrophobic silane (dimethyltrichlorosilane) replaces the more familiar metallization step (see Volume II, Figure 1.37). Hydrophobic patterning is followed by the deposition of an amino silane (3-aminoethylaminopropyltrimethoxysilane) in the remaining areas, the amino group forming an attachment site for proteins.[37,38] Successful chemical patterning of glass slide surfaces is demonstrated by the growth pattern of cells. Some cells, such as fibroblasts, grow in hydrophilic areas but have difficulty adhering and spreading in hydrophobic areas. Thus, sharp edge definition is obtained along edges of hydrophilic/hydrophobic interfaces. Using this technique, neuronal growth can be controlled, and neuronal processes may be induced to follow the straight lines of the original mask pattern. It has been suggested that this type of approach could be applied to nerve repair techniques. The method and its range of applications are described in Britland et al.[39] Surface characteristics of bio-MEMS materials that should be considered are not only chemistry and surface texture but also hydrophilicity, charge, polarity and energy, heterogeneous distribution of functional groups, water absorption, and chain mobility (for a polymer).

Biocompatibility of Si in MEMS

The above background on the many aspects determining biocompatibility may explain why literature on biocompatibility of Si remains somewhat of an enigma; it is often hard to understand what is meant when a Si device is proclaimed to be biocompatible. Maluf, without providing references, claims that preliminary medical evidence indicates that silicon, even without protein adsorption control, is benign in the body.[34] In reality, it remains of considerable interest to improve the biocompatibility of silicon microdevices and to engineer surfaces with reduced protein and cell adherence for the prevention of fouling and fibrotic response—especially when dealing with active silicon devices in contact with blood (e.g., an in vivo Si-based glucose sensor).[40] Even if significant control over protein adsorption and cell adhesion is achieved, Si remains a brittle material, which may shard in the body, and, just like glass, is to be avoided in any load-bearing application in living subjects or foodstuffs (e.g., as a micromachined needle). Although the motivation for most MEMS devices (especially active ones) is usually to eliminate any adhesion between proteins and surfaces, in some cases the ability to enable protein adhesion locally and not in other areas to provide for a background signal (e.g., for cell-based assays) clearly requires two different surface chemistries. Studies at the University of Michigan and MIT, according to Issys' web site, demonstrated that boron-doped SCS has a superior biocompatibility when compared with standard SCS or polysilicon-based devices (http://www.memsissys.com/new_htm/memsmedical/memsmedicalbiocompatibility.htm). Some definitive work has been done with Si for cortical implants. Si electrode arrays were proposed as permanent cortical implants with the purpose of restoring useful vision, and their biocompatibility with neural tissue was investigated. In this context, silicon electrode arrays were left in place within a rabbit cortex for six months. Afterward, the neuron density as a function of radius from the center of the electrode shafts was used as the means of assessing damage. The researchers concluded that the electrode array showed long-term biocompatibility because normal brain tissue was observed within micrometers of the surface of the shafts.[41,42] Initial results on porous silicon from DeMonfort University (Leicester, UK) indicate that it is biocompatible with living tissue and offers the potential to bridge the gap between a mechanical device and human tissue.

Example 4.1: Design of an SOI-based high-sensitivity static piezoresistive cantilever for label-less sensing

In this example, we want to implement an integrated piezoresistor for position sensing of a static cantilever. Most cantilever design considerations involve a beam in resonance mode. Here we consider the optimal design rules for measuring bending induced in a static cantilever by differential surface stresses. This is especially important for measurements in liquids; immersion of a cantilever in water damps the resonance response to a value approximately an order of magnitude less than in air, whereas the bending response remains unaffected by the presence of water. Thundat et al. claim that static, adsorption-induced stress sensors have a sensitivity three orders of magnitude higher than dynamic based ones.[43] In practice, using optical techniques to measure deflection, surface stresses as small as ~10^{-4} Nm^{-1} have been measured.[44] Here, we want to implement a simpler piezoresistive measurement. Unfortunately this usually comes at the expense of resolution. The aim is to design the cantilever and the piezoresistor such that the same sensitivity that is obtained with the optical methods can be maintained in this case as well.

For the analysis of the degree of bending of a cantilever caused by differential stress, we examine Stoney's formula. In 1905, Stoney derived a relation relating thin-film stress to the radius of curvature of a thin substrate. In the case where the substrate is a thin cantilever, the Stoney equation is given by:

$$\frac{1}{R} = \frac{6(1-\nu)}{Et^2}\Delta\sigma \qquad (4.99)$$

where R corresponds to the radius of curvature of the cantilever, t its thickness, σ the differential surface stress ($\Delta\sigma = \sigma_1 - \sigma_2$, i.e., the difference in surface stress between the top and bottom surfaces, in units of N/m), and ν and E are Poisson's ratio and Young's modulus for the substrate, respectively (see also Equation 6.23 in Volume III, Chapter 6, "Disk Method: Biaxial Measurements of Mechanical Properties of Thin Films" for bending of a thin disk).

From geometric considerations, the radius of curvature is related to the displacement of the free end of the cantilever, δ, and its length, L, as $1/R = 2\delta/L^2$. Combining this last expression with Equation 4.99, we derive the cantilever displacement as a function of the differential surface stress as:

$$\delta = \frac{3L^2(1-\nu)}{Et^2}\left(\sigma_1 - \sigma_2\right) \qquad (4.100)$$

From this equation, if we assume a constant differential surface stress, maximum deflection occurs with a low t and E, a large L, and a low ν. To maximize sensitivity, the piezoresistor should be placed in the zone of maximal stress. The stress in the beam is zero at the centerline and increases linearly with the distance away from the centerline. The stress also increases toward the base of the cantilever so that the highest sensitivity will be achieved with a piezoresistor placed at the surface of the cantilever beam near the base. The stress at that point can be calculated to be:

$$\sigma_{max} = \frac{6L}{Wt^2}F = \frac{3Et}{2L^2}\delta \qquad (4.101)$$

where F is the applied force (i.e., $K \times \delta$). The resulting fractional resistance change is given by the piezoresistive relation:

$$\frac{\Delta R}{R} = \pi_L\sigma_{max} = \beta\frac{6L\pi_L}{Wt^2}F = \beta\frac{3Et\pi_L}{2L^2}\delta \qquad (4.102)$$

where π_L is the longitudinal piezoresistive coef) ficient of silicon at the operating temperature and at a given doping. The cantilever is oriented along the <110> crystallographic axis of silicon where the piezoresistive coefficient, π_L, is maximum. The coefficient β is a correction factor between 0 and 1 and accounts for the fact that the resistor is not limited to the surface of the cantilever but has a certain depth. Its value depends on the silicon doping depth profile and the thickness of the beam. The doping depth is determined by the implantation parameters and subsequent thermal processes. It is important that the resistor be made shallow, so that the current flows as close as possible to the surface of the cantilever. In an extreme case, a uniformly doped cantilever would exhibit no net piezoresistive response because opposite sign stresses on the top and the bottom of the cantilever would give an equal but opposite contribution to the change

in resistance. Substituting Equation 4.100 in Equation 4.102, we obtain the expression for fractional resistance change in terms of surface stress:

$$\frac{\Delta R}{R} = \beta \frac{3\pi_L (1-\nu)}{t} (\sigma_1 - \sigma_2) \qquad (4.103)$$

Thinning the cantilever is the principal means to obtain the largest resistivity change. Making the overall device smaller (more specifically *W*, *L*, and *t*) is helpful for making denser arrays and is simpler to achieve with piezoresistive cantilevers, as they are easier to make smaller than the ones that are used in conjunction with an optical detection technique, principally because their surface area is not limited by the laser spot size (typically about 30 μm).

We now address the need to design in a reference cantilever. Immersion of cantilevers in a liquid results in two reported long-term drift phenomena. First, thermal effects (especially where a gold layer on one surface creates a sensitive bimetallic structure) occur slowly over a period of hours. A second slow effect, which takes up to 10 h to stabilize, seems associated with slow rearrangement of surface adsorbates.[45] Both drift phenomena imply that a differential-type measurement will be beneficial to extract the data from adjoining cantilevers and to exclude common mode drift phenomena. The latter approach was implemented by Fritz et al., who demonstrated that in doing so, a single mismatch between two DNA sequences could be detected.[46] From the above, it would also be preferable to derivatize the cantilever pairs with the required chemical coatings without using a metal deposit. This way, the temperature effect can be reduced, and the mass of the sensor will be less.

Thus, the fabrication challenge is to make an array with the thinnest possible cantilevers having an extremely shallow doped layer for the piezoresistor at its base to keep β in Equation 4.103 as close to 1 as possible. Moreover, a symmetrical Wheatstone bridge design with a built-in reference cantilever is required to reduce the described background drift and common mode vibrations. For the fabrication of large arrays of cantilever pairs, one would like to have them all as identical as possible, and, if possible, the chemical coatings should be deposited directly on the beam rather than on a deposited metal layer. Our proposed design incorporates features from various MEMS research groups around the world.

The first concerns the depth of the doped region in the piezoresistors. For very thin beams, it is difficult to confine the doped region to the surface of the beam. Because activation of dopants is achieved by annealing, dopant diffusion is unavoidable (see above section "Doping of Si"). To maximize β we rely on Harley et al., who made cantilevers under 1000 Å thick (870–900 Å) with lengths ranging from 10 to 350 μm and widths from 2 to 44 μm.[47] To reduce the depth below the capabilities of conventional implantation, this team used vapor-phase epitaxial growth to deposit the boron-doped layer. The boron atoms are incorporated into the lattice during the epitaxy, so an activating anneal is not required. Furthermore, because there is no damage-enhanced mobility from the ion-implantation step, some high-temperature steps can be tolerated. Their fabrication starts by growing and removing a 2000-Å thermal oxide from the epi Si layer of a 10-Ω cm p-type SIMOX SOI wafer, thinning the Si layer to 800 Å. A 30-s HCl clean in the epichamber removes another 100 Å before a 300-Å layer of 4×10^{19} cm^{-3} boron doped silicon is grown over the entire wafer. The boron layer defines the thickness of the resistors. The intermediate oxide layer is used as an etch stop during the process of etching the bulk silicon substrate in one of the last fabrication steps. In a first step, the cantilevers are patterned and plasma etched. Boron contacts are implanted at 1×10^{15} cm^{-2}, 30 keV, followed by a 200-Å growth of passivating thermal oxide during a 3-h anneal at 700°C. For contacting the piezoresistors, aluminum is deposited and annealed in a forming gas at 400°C for 1 h. Subsequently, a release mask is sputtered on the back of the wafer, and a Bosch DRIE (see Volume II, Chapter 3 under "Inductively Coupled Plasma") is used to release the cantilevers. The DRIE stops at the buried oxide, which is removed with a 6:1 buffered oxide etch. For implementing a reference probe, one could take a cue from Thaysen et al.,[48] who fabricated a thermally symmetrical Wheatstone bridge as shown in Figure 4.80.

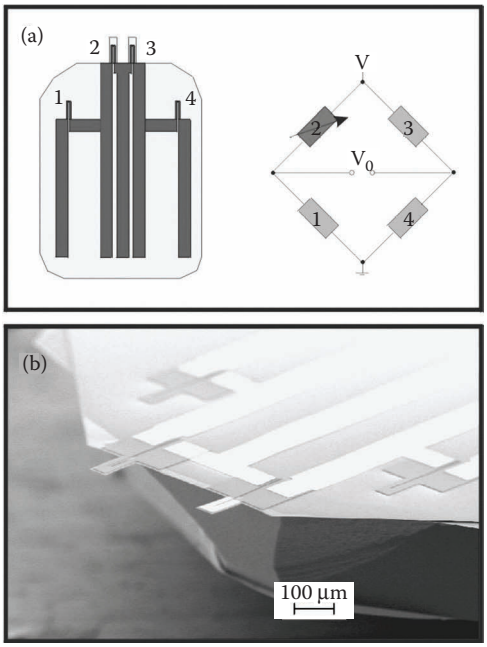


FIGURE 4.80 (a) Thermally balanced Wheatstone bridge. (b) SEM of a cantilever pair. (Sensor and reference courtesy of Dr. Anja Boisen, Technical University Denmark.)

Appendix 4A: Some Properties of Error Functions and Complementary Error Functions

$$\text{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-dy}$$

$$\text{erf}(x) \equiv 1 - \text{erf}(x)$$

$$\text{erf}(0) = 0$$

$$\text{erf}(\infty) = 1$$

$$\text{erf}(x) \cong \frac{2}{\sqrt{\pi}} x \quad \text{for} \quad x << 1$$

$$\text{erfc}(x) \cong \frac{1}{\sqrt{\pi}} \frac{e^{-x^2}}{x} \quad \text{for} \quad x >> 1$$

$$\frac{d}{dx}\text{erf}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2}$$

$$\frac{d^2}{dx^2}\text{erf}(x) = -\frac{4}{\sqrt{\pi}} x e^{-x^2}$$

$$\int_0^x \text{erfc}(y')dy' = x\,\text{erfc}(x) + \frac{1}{\sqrt{\pi}}(1 - e^{-x^2})$$

$$\int_0^\infty \text{erfc}(x)dx = \frac{1}{\sqrt{\pi}}$$

Questions

4.1: Describe a fabrication process for a planar npn transistor.
Thanks to Andrew Hatch,UC Irvine

4.2: Consider the growth of a thermal oxide on a 4-in. <111> silicon wafer that is initially 525 µm thick and comes with a native oxide layer of 25 Å. If the oxidation process is carried out dry for 1 h at 1000°C and at 1 atm of pressure, what is the final wafer thickness? How much of this will be Si and how much will be SiO_2? What color will the resulting oxide be? Repeat these calculations for 1 hr of wet oxide growth at 1000°C. What is the ratio of the rates of wet to dry oxide growth? To solve the latter, assume that B (H_20) = 0.3 µm²/hr, B(O_2) = 0.01 µm²/hr; and B/A (H_2O) = .55 µm/hr, B/A (O_2) = 0.05 µm/hr for <111> Si: how do your results compare if you use a <100> wafer instead of a <111> wafer? (Experimental from http://www.eng.tau.ac.il/~yosish/Courses/vlsi1/I-4-1-Oxidation.pdf.)
Thanks to Andrew Hatch, UC Irvine

4.3: Great care is taken to ensure that silicon boules are single crystalline with the highest possible purity. Describe how some adverse effects might manifest themselves if the processing was not handled with enough care; i.e., what makes single crystallinity and high purity so important?
Thanks to Matt Sullivan, UC Irvine

4.4: We are trying to etch a 20-µm × 20-µm through hole in a Si wafer. What mask dimensions would you need to achieve such a through hole in a typical 6-in [100]-oriented silicon wafer (600 µm thick)? Assume that the 20-µm × 20-µm through hole is the exit hole in the bottom plane of the wafer.
Thanks to Rajtarun Madangopal, UC Irvine

4.5: How would you obtain vertical sidewalls using a [100]-oriented Si wafer?
Thanks to Rajtarun Madangopal, UC Irvine

4.6: In an experiment, boron is used to dope a silicon substrate by diffusion. The temperature

for diffusion is 1100°C. The activation energy is 3.69 eV. D_0 is 10.5 cm^2/s.

(a) Determine the type of resulting semiconductor.

(b) If the surface concentration of boron is a constant: $N_s = 10$ /cm^3, calculate the total amount of dopant atoms per unit area of the semiconductor after diffusion 10 min later.

(c) If the diffusing amount Q is constant, which is the same as the value calculated in step (b), derive the diffusion equation and calculate the surface dopant concentration after diffusing for 1 h.

Thanks to Xi Wang, UC Irvine

4.7: A pn junction is made by doping Si at room temperature. At the *n* edge of the depletion region, N_A was measured to be 10^{11}/cm^3. Calculate the barrier voltage at the *n* edge of the p-n junction. For Si at room temperature, $N_D = 10^{15}$/cm^3 and $N_i = 10^{10}$/cm^3.
Thanks to Chris Hoo, UC Irvine

4.8: Draw the energy band diagram in **k**-space for an intrinsic semiconductor. Why can we locate the Fermi level in the middle of the band gap?
Thanks to Anna Torrents Cabestany, UC Irvine

4.9: Give an example of a (111) plane and a (110) plane intersecting at 35.26°, 90°, and 144.74°.
Thanks to Yuka Okabe, UC Irvine

4.10: Calculate the values for Young's modulus, shear modulus, and Poisson's ratio for [100]-directed Si. Given are the values of the following compliance constants: S_{11} (7.68×10^{-12} m^2/N), S_{44} (12.6×10^{-12} m^2/N), and S_{12} (-2.14×10^{-12} m^2/N).
Thanks to Jessica Lawson, UC Irvine

4.11: The identity of a semiconductor can be determined by measuring its intrinsic carrier concentrations at various temperatures. An unknown semiconductor is submerged in ice water (273 K) and boiling water (373 K). Its intrinsic carrier concentrations, n_i, were found to be 4.6×10^{18} m^{-3} and 3.1×10^{20} m^{-3} at 273 K and 373 K, respectively. What is the bandgap and identity of the semiconductor?
Thanks to William Sin, UC Irvine

4.12: At what temperature does the concentration of intrinsic carriers in intrinsic α-quartz (SiO_2) match the concentration in Si at T = 300 K? Comment on the semiconductor potential of SiO_2. How does the oxidation of a Si melt during single crystal Si growth impact the construction of ICs?
Thanks to Jungyun Kim, UC Irvine

4.13: Explain the electron and the hole flow in a semiconductor substrate using a daily life phenomena.
Thanks to Kivanc Azgin, UC Irvine

4.14: (a) Describe the physical differences between nMOS and pMOS. (b) Which is faster and why?
Thanks to John Clanton, UC Irvine

4.15: (a) Why does electron mobility decrease as gate length decreases? (b) What are some ways to counteract this problem?
Thanks to Gordon Pfeiffer, UC Irvine

4.16: Explain the relevance of plastic deformation in Si and metals. Include any role that temperature may play in plastic deformation.
Thanks to Daniel Almendarez, UC Irvine

4.17: Aluminum is to be diffused into a silicon single crystal. At what temperature will the diffusion constant (D) be 5×10^{-11} cm^2/s? (Given: $E_A = 334$ kJ/mole and $D_0 = 8$ cm^2/s.)

4.18: Calculate the number of silicon atoms per unit volume of silicon. During the growth of silicon single crystals it is often desirable to deliberately introduce atoms of other elements (known as dopants) to control and change the electrical conductivity and other electrical properties of silicon. Phosphorus (P) is one such dopant that is added to make silicon crystals n-type semiconductors. Assume that the concentration of P atoms required in a silicon crystal is 10^{17} atoms/cm^3. Compare the concentrations of atoms in silicon and the concentration of P atoms. What is the significance of these numbers from a technological viewpoint? Assume that density of silicon is 2.33 g/cm^3.

4.19: Describe briefly the crystal structure of silicon: include such details as structure type (how does it differ from FCC), lattice parameters, packing density, and highest packing density plane.

4.20: Write a paragraph on the mechanical properties of single-crystal silicon (SCS), with special emphasis on strength, plastic/elastic behavior, and hardness.

4.21: Ionization energy of dopants using the Bohr model: Consider an N-type silicon crystal, doped with phosphorus atoms. Each dopant atom replaces a silicon atom from its lattice site. It hence donates one extra electron to the system. Calculate the ionization energy of this extra electron, assuming the other 4 valence electrons, the inner electrons, and the nucleus of the P atom as a core of net charge +1. Assume the effective mass of electrons in silicon $m_e^+ = 0.2\ m_0$, and a relative permittivity $\varepsilon_r = 12$.
Thanks to Arvind Ajoy, Indian Institute of Technology Madras

4.22: In a silicon crystal, there are 8 equivalent (111) planes and 12 equivalent (110) planes that can be drawn into the structure, making for 96 intersection combinations.
(a) Sketch the 8 (111) planes, including the Miller indices for each.
(b) Sketch the orientation of Si atoms on the (111) plane.
(c) Sketch the 12 (110) planes, including the Miller indices for each.
(d) Sketch the orientation of Si atoms on the (110) plane.
(e) Perform the calculations that show that the {111} and {110} families of planes can intersect at only three defined angles, 35.26°, 90° or 144.74°.
(f) Assign one of these angles to each of the 96 combinations of planes. How many combinations yield 35.26°? 90°? 144.74°? Comment on the significance of these angles.
Thanks to Matt Sullivan, UC Irvine

4.23: Differentiate between intrinsic and extrinsic semiconductors.
Thanks to Rajtarun Madangopal, UC Irvine

4.25: What wafer orientation would you select if you had to etch under a microbridge a shallow V-groove crossing it at a 90° angle?
Thanks to Rajtarun Madangopal, UC Irvine

4.26: List three uses of wafer flats. Also justify why we need to round wafer edges.
Thanks to Rajtarun Madangopal, UC Irvine

4.27: Silicon dioxide growth involves the heating of a silicon wafer in a stream of steam or in wet or dry oxygen/nitrogen mixture at elevated temperatures.
(a) Write the oxidation reaction for both wet and dry oxidation of silicon.
(b) Calculate the silicon dioxide thickness after diffusing for 10 h for both wet and dry oxidation. [Assume that it is a long oxidation time. $C^*(O_2) = 5.2 \times 10^{16}/cm^3$, $C^*(H_2O) = 3.0 \times 10^{19}/cm^3$; $N(O_2) = 2.2 \times 10^{22}/cm^{-3}$, $N(H_2O) = 4.6 \times 10^{22}/cm^{-3}$.]
Thanks to Xi Wang, UC Irvine

4.28: There is an n-type silicon semiconductor with electron concentration of $10^{15}\ cm^{-3}$, electron mobility of $1000\ cm^2/v{\cdot}s$.
(a) Calculate the resistance of this semiconductor.
(b) If the temperature is 40 K, $N_c = 2.8{\times}10^{17}\ / cm^3$, calculate the Femi level energy.
Thanks to Xi Wang, UC Irvine

4.29: In a doped piece of Si at 300 K, the mobility of minority carriers was found to be $476\ cm^2/V\text{-}s$ with a carrier lifetime of 10^{-7} s. Determine the diffusion length of the minority carrier.
Thanks to Chris Hoo, UC Irvine

4.30: Tunneling in a reversed-biased diode is called the Zener process. What are two requirements need to be fulfilled before tunneling occurs?
Thanks to Chris Hoo, UC Irvine

4.31: Why is the <110> direction the only orientation in a [100]-silicon wafer that does cleave with exactly perpendicular edges?
Thanks to Anna Torrents Cabestany, UC Irvine

4.32: Explain why single-crystal silicon (SCS) exhibits a linear-elastic behavior at lower strain values and yields catastrophically when stressed beyond the limit? What happens to SCS at elevated temperatures?
Thanks to Jessica Lawson, UC Irvine

4.33: Describe some of the optical properties of silicon and some of the current applications in which these properties are utilized.
Thanks to Jessica Lawson, UC Irvine

4.34: Given the values below for GaAs, calculate the Fermi energy of intrinsic GaAs at room temperature (300 K). Does the difference between effective mass of electrons to holes have an impact on the Fermi energy?
Thanks to Jungyun Kim, UC Irvine

4.35: What are the three major categories of high-*k* materials, name a material from each category, and explain why engineers are searching for higher-*k* materials despite the immense benefits of SiO_2 (k = 3.9).
Thanks to Gordon Pfeiffer, UC Irvine

4.36: What is a tensor of second rank? Name one example. How many components does a tensor of second rank have? What about a tensor of fourth rank?
Thanks to Wei Xiong, UC Irvine

4.37: List some possible difficulties arising due to MOSFET size reduction.
Thanks to Yanan Liu, UC Irvine

4.38: For silicon dioxide growth, compare dry oxide growth and wet oxide growth.
Thanks to Yiying Yao, UC Irvine

Further Reading

Campbell, S. A. 1966. *The science and engineering of microelectronic fabrication*. Oxford, UK: Oxford University Press.

Dorf, R. C. ed. 2006. *Electronics, power electronics, optoelectronics, microwaves, electromagnetics, and radar* (*the electrical engineering handbook*) 3rd ed. Boca Raton, FL: CRC Taylor &Francis Group.

Glaser, A. B., and G. E. Subak-Sharpe. 1979. *Integrated circuit engineering*. Reading, MA: Addison-Wesley Publishing Company.

Hess, D. W., and K. F. Jensen. 1989. *Microelectronic processing, chemical engineering aspects, advances in chemistry series 221*. Washington, DC: American Chemical Society.

Maluf, N. 2000. *An introduction to microelectromechanical systems engineering*. Boston, Artech House.

Mayer, J. W., and S. S. Lau. 1990. *Electronic materials science: for integrated circuits in Si and GaAs*. New York: Macmillan Publishing Company.

Pillai, S. O. 2005. *Solid state physics*. 6th ed. New Delhi: New Age International (P) Limited.

Sze, S. M. 1988. *VLSI technology*, 2nd ed. New York: McGraw-Hill Book Company.

References

1. Schumacher, A., H.-J. Wagner, and M. Alavi. 1994. Mit Laser und Kalilauge. *Technische Rundschau* 86:20–23.
2. Peeters, E. 1994. *Process development for 3D silicon microstructures, with application to mechanical sensor design*. PhD diss., Katholieke Universiteit Leuven, Leuven, Belgium.
3. Elwenspoek, M., H. Gardeniers, M. de Boer, and A. Prak. 1994. Report no. 122830. University of Twente, Twente, the Netherlands.
4. Katz, L. E. 1988. *VLSI technology*. Ed. S. M. Sze. New York: McGraw-Hill.
5. Kendall, D. L. 1979. Vertical etching of silicon at very high aspect ratios. *Ann Rev Mater Sci* 9:373–403.
6. Seidel, H. 1990. *Technical digest: 1990 solid state sensor and actuator workshop*. Hilton Head Island, SC.
7. Ho, C. P., and J. D. Plummer. 1979. Si/SiO2 interface oxidation kinetics: A physical model for the influence of high substrate doping levels, I. Theory. *J Electrochem Soc* 126:1516–22.
8. Ho, C. P., and J. D. Plummer. 1979. Si/SiO2 interface oxidation kinetics: A physical model for the influence of high substrate doping levels, II. Comparison with experiment and discussion. *J Electrochem Soc* 126:1523–30.
9. Ho, C. P., J. D. Plummer, and J. D. Meindl. 1978. Thermal oxidation of heavily phosphorus-doped silicon. *J Electrochem Soc* 125:665–71.
10. Schmidt, P. F., and W. Michel. 1957. Anodic formation of oxide films on silicon. *J Electrochem Soc* 104:230–36.
11. Lewerenz, H. J. 1992. Anodic oxides on silicon. *Electrochimica Acta* 37:847–64.
12. Madou, M. J., W. P. Gomes, F. Fransen, and F. Cardon. 1982. Anodic oxidation of p-type silicon in methanol as compared to glycol. *J Electrochem Soc* 129:2749–52.
13. Fair, R. B. 1989. *Microelectronics processing: chemical engineering aspects*. Eds. Hess, D. W., and K. F. Jensen. Washington, DC: American Chemical Society.
14. Madou, M. J., and S. R. Morrison. 1989. *Chemical sensing with solid state devices*. New York: Academic Press.
15. Moreau, W. M. 1988. *Semiconductor lithography*. New York: Plenum Press.
16. Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication*. New York: Plenum Press.
17. Landauer, R. 1961. Irreversibility and heat generation in the computing process. *IBM J Res Dev* 5:183–91.
18. Thompson, S. E., S. Suthram, Y. Sun, G. Sun, S. Parthasarathy, M. Chu, and T. Nishida. 2006. Future of Strained Si/Semiconductors in Nanoscale MOSFETs. *IEDM Tech Digest*.
19. Chou, P. C., and N. J. Pagano. 1967. *Elasticity: tensor, dyadic, and engineering approaches*. New York: Dover Publications.
20. Kittel, C. 1976. *Introduction to solid state physics*. New York: Wiley.

21. Brantley, W. A. 1973. Calculated elastic constants for stress problems associated with semiconductor devices. *J Appl Phys* 44:534–35.
22. Nikanorov, S. P., Y. A. Burenkov, and A. V. Stepanov. 1972. Elastic properties of silicon. *Sov Phys-Solid State* 13:2516–18.
23. Khazan, A. D. 1994. *Transducers and their elements.* Englewood Cliffs, NJ: PTR Prentice Hall.
24. Worthman, J. J., and R. A. Evans. 1965. Young's modulus, shear modulus, and Poisson's ratio in silicon and germanium. *J Appl Phys* 36:153–56.
25. Metzger, H., and F. R. Kessler. 1970. Der Debye-Sears Effect zur Bestimmung der Elastischen Konstanten von Silicium. *Z Naturf A* 25:904–06.
26. Greenwood, J. C. 1988. Silicon in mechanical sensors. *J Phys E Sci Instrum* 21:1114–28.
27. Huff, M. A., and M. A. Schmidt. 1992. *Technical digest: 1992 solid state sensor and actuator workshop.* Hilton Head Island, SC.
28. Maseeh, F., and S. D. Senturia. 1990. Plastic deformation of highly doped silicon. *Sensors Actuators* 23:861–65.
29. Ding, X., and W. Ko. 1991. Sixth International Conference on Solid-State Sensors and Actuators (Transducers '91), San Francisco, CA.
30. Callister, D. W. 1985. *Materials science and engineering,* New York: Wiley.
31. Smith, C. S. 1954. Piezoresistance effect in germanium and silicon. *Phys Rev* 94:42–49.
32. Kanda, Y. 1982. A graphical representation of the piezoresistance coefficients in silicon. *IEEE Tran Electron Devices* 29:64–70.
33. Middlehoek, S., and S. A. Audet. 1989. *Silicon sensors*. San Diego: Academic Press.
34. Maluf, N. 2000. *An introduction to microelectromechanical systems engineering.* Boston: Artech House.
35. IMEC. 1994. *Silicone detectors.* IMEC Brochure. Leuven, Belgium: Interuniversity Microelectronics Centre.
36. von Recum, A. ed. 1998. *Handbook of biomaterials evaluation: scientific, technical, and clinical testing of implant materials.* New York: Hemisphere.
37. Cooper, J. M., J. R. Barker, J. V. Magill, W. Monaghan, M. Robertson, C. D. W. Wilkinson, A. S. G. Curties, and G. R. Moores. 1993. A review of research in bioelectronics at Glasgow University. *Biosens Bioelectron* 8:R22–R30.
38. Connolly, P., G. R. Moores, W. Monaghan, J. Shen, S. Britland, and P. Clark. 1992. Microelectronic and nanoelectronic interfacing techniques for biological systems. *Sensors Actuators* 6:113–21.
39. Britland, S., G. R. Moores, P. Clark, and P. Connolly. 1990. Patterning and cell adhesion and movement on artificial substrate: a simple method. *J Anat* 170:235–36.
40. Zhang, M., T. Desai, and M. Ferrari. 1998. Proteins and cells on PEG immobilized silicon surfaces. *Biomaterials* 19:953–60.
41. Clark, L. D., D. J. Edell, V. M. McNeil, and V. V. Toi. 1992. Factors influencing the biocompatibility in insertable silicon microshafts in cerebral cortex. *IEEE Trans Biomed Eng* 39:635–43.
42. Horch, K., R. A. Normann, and S. Schmidt. 1993. Biocompatibility of silicon-based electrode arrays implanted in feline cortical tissue. *J Biomed Mater Res* 27:1393–99.
43. Thundat, T., L. A. Bottomley, S. Meller, W. H. Velander, and R. Van Tassell. 2001. *Immunoassays: methods and protocols*. Eds. Ghindilis, A. L., A. R. Pavlov, and P. B. Atanajov. Totawa, NJ: Humana Press.
44. O'Shea, S. J., M. E. Welland, T. A. Brunt, A. R. Ramadan, and T. Rayment. 1996. Atomic force microscopy stress sensors for studies in liquids. *J Vac Sci Technol* 14:1383–85.
45. Moulin, A. M. 2000. Microcantilever-based biosensors. *Ultramicroscopy* 82:23–31.
46. Fritz, J., M. K. Baller, H. P. Lang, H. Rothuizen, P. Vettinger, E. Meyer, H.-J. Güntherodt, C. Gerber, and J. K. Gimzewski. 2000. Translating biomolecular recognition into nanomechanics. *Science* 288:316–18.
47. Harley, J. A., and T. W. Kenny. 1999. High-sensitivity cantilevers under 1000 Å thick. *Appl Phys Lett* 75:289–91.
48. Thaysen, J., A. Boisen, O. Hansen, and S. Bouwstra. 2000. Atomic force microscopy probe with piezoresistive readout and a highly symmetrical Wheatstone bridge arrangement. *Sensors Actuators* 83:47–53.
49. Thompson, S. E., M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, et al. 2004. A 90-nm logic technology featuring strained-silicon. *IEEE Trans Electron Devices* 51:1790–7.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

Photonics

If it only were possible to make materials in which electromagnetically waves cannot propagate at certain frequencies, all kinds of almost-magical things would happen.

Sir John Maddox, *Nature* (1990)

Outline

Introduction

The Nature of Light

Diffraction and Image Resolution

Refraction

Reflectance and Total Internal Reflectance

Light Polarization

Maxwell's Equations

Beyond Maxwell

Optical Properties of Materials

Light Interaction with Small Particles

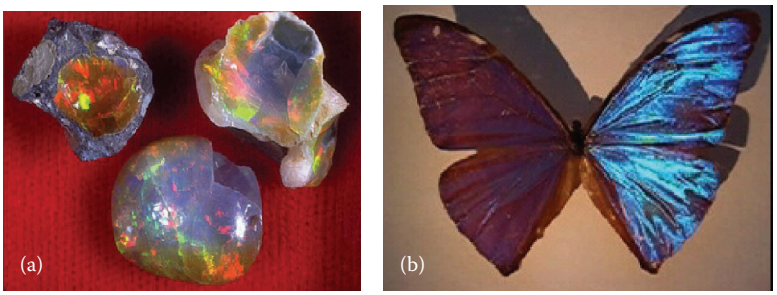
Comparing Photons with Electrons—Photonic Crystals

The μ-ε Quadrant and Metamaterials

Lasers

Questions

References



Natural opals (a) and *P. nireus* butterfly (b). Periodic structure leads to striking iridescent color effects even in the absence of pigments as a result of photonic crystals, i.e., stacks of nanoscale gratings.

Introduction

Photonics involves the use of radiant energy and uses photons the same way that electronic applications use electrons. We start this chapter with a discussion of the nature of light, its near- and far-field diffraction, and how this relates to image formation and resolution limits. Next we introduce refraction, reflection, and polarization of light and launch into the Maxwell's equations for electromagnetic (EM) waves propagating in a vacuum, in matter, and at interfaces. We solve the Maxwell's equations for EM waves propagating in a vacuum, in media (without charges and with charges and currents), in right-handed and left-handed materials, and calculate their energy and momentum. From the considerations of EM waves at boundaries we derive the Fresnel equations linking reflectance and refractive index of a material and get our first exposure to bulk, surface, and localized plasmons in plasmonics. Plasmonics is of growing importance for use in submicron lithography, near-field optical microscopy, enhancement of light/matter interaction in sensors, high-density data storage, and highly integrated optic chips.

Newton's laws were found to remain the same under a Galilean transformation, but a Galilean transformation does not apply to Maxwell's

equations. We will see that, in his special theory of relativity, Einstein convinced the world that whenever the speed of light is involved, the Galilean transformation has to be replaced by the Lorentz transformation.

To explain the interaction of EM radiation with materials at a level of the atomic structure of the material, we launch the Drude-Lorentz model, an upgrade of the Drude model from Chapter 3 on quantum mechanics. Using the Drude-Lorentz model, we review the distinctive optical properties of metals, insulators, and semiconductors. Besides reviewing the optoelectronic behavior of these bulk materials, we also deal with the interaction of electrons and photons with lower-dimensional materials. In the case of metals this brings us back to plasmonics, which we detail a bit further in this section. Then we review scaling of electronic and optical properties of semiconductors and compare semiconductor nanoparticles with metal nanoparticles.

We follow this treatise on optoelectronic properties of materials with a study of light scattering on nano- and microstructures (with sizes smaller, comparable, and larger than λ). We cover elastic Rayleigh and Mie scattering, as well as inelastic Raman scattering of light from these small particles. We learn how surface plasmons enhance the Raman effect in surface-enhanced Raman spectroscopy (SERS) by up to 10^{12}–10^{14} times compared with nonresonant Raman spectroscopy, and we discuss nanoscale confinement of radiation in near-field scanning optical microscopy (NSOM) for subwavelength resolution.

We proceed with a study of EM effects in periodic photonic crystals [three-dimensional (3D)], planar waveguides [two-dimensional (2D)], optical fibers [one-dimensional (1D)], and microsphere optical cavities [zero dimension (0D)]. This study includes a comparison of the behavior of photons with electrons in a vacuum and in periodic media, and we find that the analogous electron confinement structures to the aforementioned photonic structures are bulk crystals (3D), quantum wells (2D), quantum wires (1D), and quantum dots (OD) and that the tunneling of electrons through a forbidden zone has its equivalent in light leaking out of a waveguide to create an evanescent field.

Then we delve into the fascinating new topic of metamaterials, human-made structures with a negative refractive index, and explain how these could make for perfect lenses and might change the photonics field forever.

An introduction to diode lasers, quantum well lasers, and quantum cascade lasers closes the chapter.

The Nature of Light

In his theory, Maxwell described light as electric and magnetic fields, oscillating perpendicular to each other and propagating perpendicular to these oscillations as shown in Figure 5.1. Fusing two fields into one, he calculated that waves in an electromagnetic (EM) field traveled at the speed of light and conjectured that light was nothing but an EM vibration at a particular frequency. In this way Maxwell was the first physicist to show that the task of field unification is a very worthwhile exercise. Before Maxwell, there was no connection between light and electricity and magnetism: it was one of the great triumphs of science—to show that two things that appear so different are in fact described by the same simple laws!

EM waves exist in an enormous range of wavelengths/frequencies. This continuous range of wavelengths/frequencies is known as the *EM spectrum*. The entire range of the spectrum is often broken up into specific bands. The subdividing of the entire spectrum into smaller spectra is done mostly based on how each region of EM waves interacts with matter. In Figure 5.2, we show a typical EM spectrum representation. The color our eyes perceive is not a frequency measurement. The eye has three color-sensitive cell types ("cone cells") that respond to red, green, and blue, and the apparent color of light is the result of the added response from the different cone types. For example, the color "purple" is not a frequency but

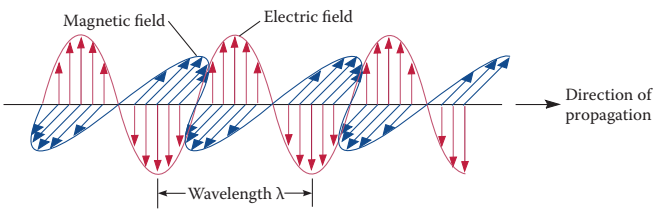


FIGURE 5.1 Maxwell's description of a light wave.

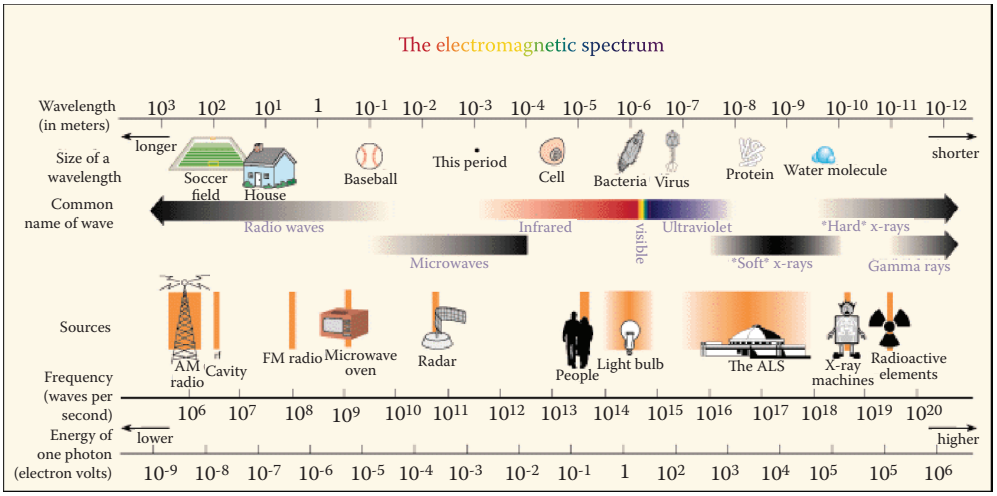


FIGURE 5.2 The electromagnetic (EM) spectrum, the source of the EM radiation, frequency, energy, wavelength, size comparison, and common name.

rather a mix of red and blue. This is great news for all types of devices: we can trick the eye by mixing three different colors (RGB) in monitors, copiers, cameras, and so on. Visible light goes from violet-blue, with a wavelength, λ, of about 400 nm and energy of about 3 eV (and since E = hν = 1.24/0.4 = 3 eV), to red, with a wavelength, λ, of about 700 nm and energy of about 1.77 eV (=1.24/0.7 = 1.77 eV). In the deep UV, and even more so at yet shorter wavelengths, materials turn transparent. Typical thermal energies at room temperature are given by kT or about 24 meV, with a radiation wavelength of 50 μm. In the microwave region, EM waves couple excellently to water molecule vibrations, making your microwave work. The radio wave regions the government assigns as "bands" to be used by ships, airplanes, police, military, amateurs, space, satellites, and radar. Cellular phones (a teenage ear infection) occupy a global system for mobile (GSM) communications band from GSM-850 MHz to GSM-1900 MHz (in the United States, Canada, and many other countries in the Americas).

The formation of EM waves is illustrated in Figure 5.3, where we drew a switch in the middle of an "antenna" wire. According to Maxwell, magnetic fields are produced in space if there is a changing electric field. In Figure 5.3A we use a battery for energy supply and in Figure 5.3B an AC source. In Figure 5.3A, when closing the switch, a current flows briefly, setting up an electric field. The current flow in turn sets up a magnetic field that circles around the wire. The + and − signs indicate the charge on each wire segment, and the black arrow is in the direction of the current. Electrical field lines are in the plane of the page and shown in red. The magnetic field lines, in accordance with the right-hand rule, are perpendicular to the plane of the page and marked by ⊗ (into the plane) and ⊙ (out of the plane). The magnetic and electrical fields do not arise instantaneously but form in time: energy is stored in these fields, and energy cannot move infinitely fast. In Figure 5.3Ba, we show a first half cycle of the AC current cycle. When the current reverses (Figure 5.3Bb), the fields reverse and we see the first disturbance moving outward. Oscillations of current and voltage at the frequency of radiation, along the simple wire antenna in Figure 5.3, cause a portion of the stored energy to be "radiated" from the antenna. These disturbances are the EM waves. The magnitudes of the electric field **E** and the magnetic field **B** decrease as $1/r$ with distance r away from the antenna. The energy density of the radiated fields falls off as $1/r^2$ because the energy carried by the EM waves is proportional to the square of the amplitude ($\mathbf{E}^2$ or $\mathbf{B}^2$). The complicated field distribution near the antenna is called *near field*. Far from the antenna, *far field*, the field lines become quite flat, and these waves are called *plane waves*. We will come back to the concept of near field and far field in the next section.

EM waves are produced whenever electric charges are accelerated, so we can produce EM waves by

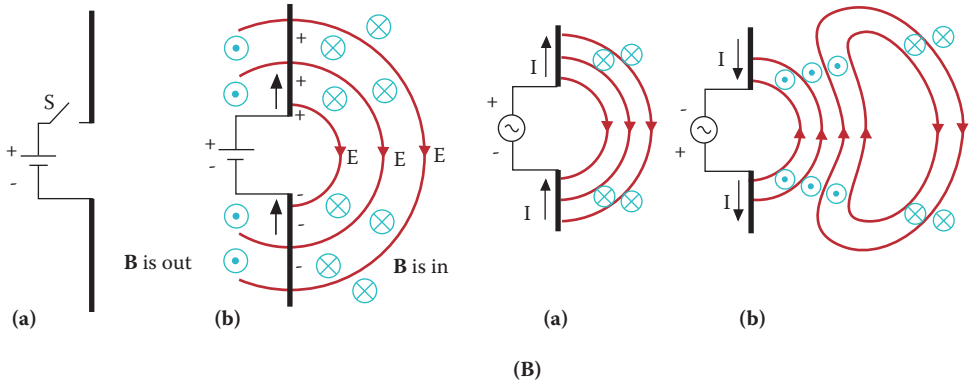


FIGURE 5.3 Formation of electromagnetic waves. (A) When closing a switch, a current flows briefly, setting up an electric field. The current flow sets up a magnetic field as circles around the wires. These fields do not arise instantaneously but form in time. Energy is stored in these fields, and energy cannot move infinitely fast. (B) When current reverses in picture (b), the fields reverse, and we see the first disturbance moving outward. These are the electromagnetic waves "…and there was light" (Genesis 1:3).

letting an alternating current flow through a wire, an antenna. The frequency of the waves created this way equals the frequency of the alternating current in the antenna. At much shorter wavelengths, the light emitted by an incandescent light bulb is caused by thermal motion that accelerates the electrons in the hot filament sufficiently to produce visible light. The inverse effect also happens: if an EM wave strikes a conductor, its oscillating electric field induces an oscillating electric current of the same frequency in the conductor. This is how the receiving antennas of radios and televisions work.

Thus, EM waves, in general, are produced by electric charges that oscillate in an antenna or perhaps in a molecule, in the case of very short wavelengths. Almost all optical effects are caused by light interacting with electric charges. An interesting exception to this is gravitational lensing. The latter occurs when the gravity of a massive celestial object, such as a galaxy, bends and focuses the light of a more distant object, resulting in a magnified, distorted, or multiple image of the original light source for a distant observer. The process was one of the predictions made in Einstein's general relativity theory.

Diffraction and Image Resolution

Far Field (Fraunhofer) and Near Field (Fresnel)

Diffraction is the apparent bending and spreading of waves when they encounter obstacles whose size is about the same order of magnitude as their wavelength. It takes place with any type of waves, including mechanical sound and water waves (Figure 5.4), as well as electromagnetic (EM) waves such as light and radio waves.

Waves either appear as plane waves, which we can envision as waves that come rolling in on a long stretch of beach, or closer to an obstruction, say, a dock, where they come as spherical waves. The latter is analogous to the circular waves expanding from a pebble dropped in a lake. Light waves, a large distance away from the obstruction (compared with the size of the obstruction), will generate an illumination pattern of light and dark. This pattern is a Fraunhofer diffraction pattern, also called *far-field diffraction*. The distances away from the obstruction must be large enough so that the light reaching the

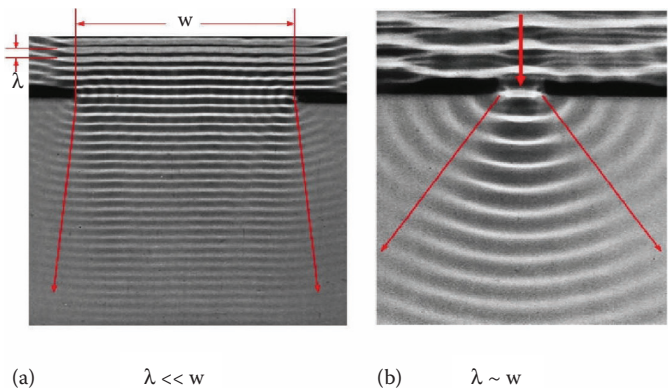


FIGURE 5.4 Diffraction in a water tank with a wavelength small compared to the aperture (a) and with a wavelength λ similar to the aperture (b).

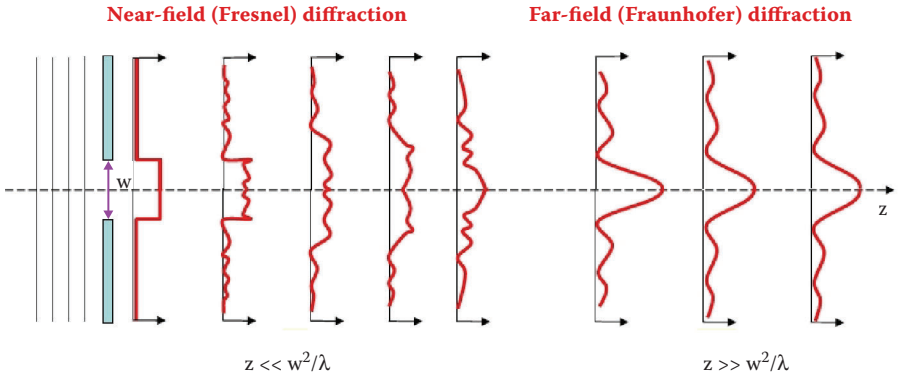


FIGURE 5.5 Fraunhofer and Fresnel diffraction regions. Aperture diameter is *w*, and *z* is the distance away from the aperture.

monitoring point, e.g., a photographic emulsion, is nearly plane waves. Spherical waves, on the other hand, produce a Fresnel diffraction pattern at a point not too far beyond the obstruction. Because these distances are close to the diffracting object, this phenomenon is also called *near-field diffraction*. Fresnel diffraction is more complex to treat mathematically than Fraunhofer diffraction. For Fraunhofer diffraction one needs distances greater than w^2/λ, where w is the size of the object (Figure 5.4). Fraunhofer diffraction works by summing the amplitudes of light waves coming from all sources. These interfere, and the intensity is determined by the square of the total amplitude.

The operational regimes for Fresnel and Fraunhofer diffraction are compared in Figure 5.5. One can calculate that Fraunhofer diffraction takes over from Fresnel diffraction if the observation point z is far enough away so that $z >> w^2/\lambda$, with w the diameter of the aperture and λ the wavelength of the light.

What's an Image?

A major breakthrough in the understanding of image formation came in 1872, when Ernst Abbe (Figure 5.6) applied Fourier transforms to explain the phenomenon (Bragg, in 1939, applied Fourier analysis to x-ray crystallography; see Chapter 2). Image formation, according to Abbe's theory, is a two-stage, double-diffraction process: an image is the diffraction pattern of the diffraction pattern of an object.

Consider an "ideal" lens system, in which an image depicts every detail of an object as shown in Figure 5.7. In the first stage of the image formation, light rays incident on the object are scattered, and an interference pattern—a Fraunhofer diffraction pattern—is created (see also Figure 5.5).

This diffraction pattern corresponds to a forward Fourier transformation. A lens, say, a microscope objective, is essential for image formation and acts to focus light from infinity to the "focal point" at a distance from the lens known as the *focal length*, *f*. Located at the focal point is the back focal plane of the lens where the diffraction pattern can be made visible on a screen, as shown in Figure 5.8. In Figure 5.8 we show how a grating-like object is imaged and that a diffraction pattern can be seen in the back focal plane.

The second stage of image formation takes place when the scattered radiation passes beyond the back focal plane of the lens and recombines to form an

FIGURE 5.6 Ernst Abbe (1840–1905). The formula from Ernst Abbe expressing the maximum possible resolution from a light microscope on a monument erected by the University of Jena in Abbe's memory.

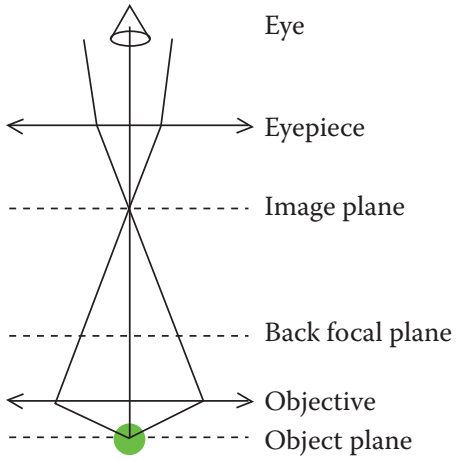


FIGURE 5.7 A lens is essential in image formation.

image. This is the back or inverse Fourier transform illustrated in Figure 5.9.

The image formed by the lens does not exactly represent the real object because not all scattered radiation enters the lens and therefore cannot all be focused on the image plane. The more of the higher orders of the fringes in Figure 5.8 can be picked up, the higher the fidelity of the image. Thus, the basic description of image formation according to Fourier optics is that a Fourier transform of the object in the back focal plane of the lens is generated, and that propagation to the image plane produces another Fourier transformation—an inverse Fourier transform—i.e., the image. Here it may also be understood why Fourier image analysis is such a powerful method for analyzing a wide variety of periodic phenomena. The Fourier transform process separates the image processing into two stages: the information contained in the diffraction pattern in the first stage reveals structural information in a straightforward manner and conveniently separates most of the signal and noise components and thus the transform may be manipulated and subsequently back-transformed in the second stage to produce a noise-filtered, better reconstructed image.

Remember from our discussion on x-ray diffraction of crystals in Chapter 2 that because there is no lens in this case, no image of the crystal lattice is generated; instead, Fresnel diffraction patterns form at finite distances and Fraunhofer diffraction patterns at infinity. Thus, x-ray diffraction methods provide a direct way to display the decomposition of x-rays in component waves (frequencies) in a forward Fourier transform. To get an image of the crystal one would need an x-ray focusing lens, which does not exist, or, as we do in practice, Fourier transform the diffraction pattern to get the real image of the crystal.

Image Resolution and Diffraction Limit

Abbe's Limit

Besides recognizing how an image is formed, Ernst Abbe was also the first to characterize the resolution of an optical system. When an object illuminated in a microscope with coherent (in-phase) light is an optical grating, the diffraction orders (N) of the grating are observed in the back focal plane of the objective, and these grating orders are then combined to form an image in the image plane (see Figures 5.7 and 5.8). The resolution of the resulting image depends on the number of grating orders that are transmitted through the optical system. Based

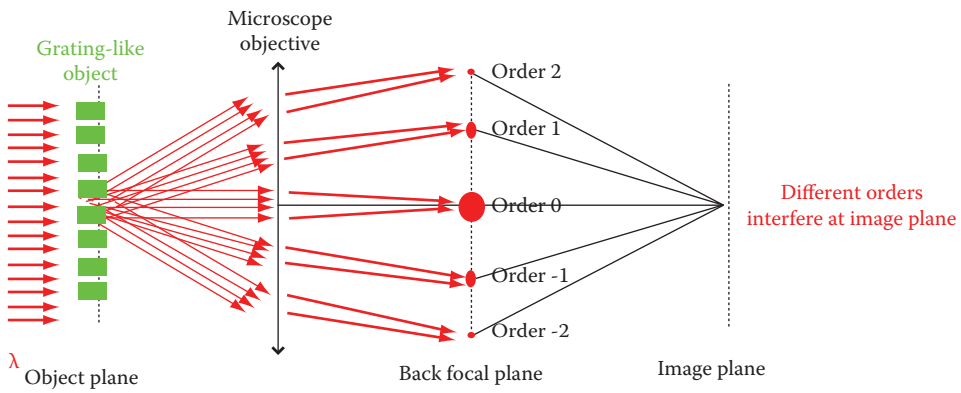


FIGURE 5.8 A grating-like object is imaged, and in the back focal plane an interference pattern—a Fraunhofer diffraction pattern—is formed that corresponds to the forward Fourier transform of the object.

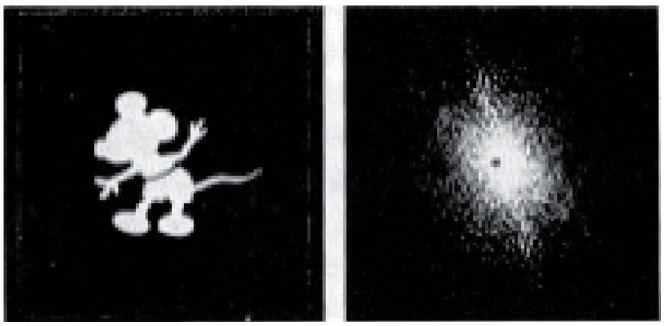


FIGURE 5.9 An inverse Fourier transform gives back the original image of Mickey Mouse on the left.

on imaging of gratings, Abbe, together with Carl Zeiss, published a paper in 1877 defining the physical laws that determined the resolving distance of an optical system. When imaging a grating with coherent light, the direction of diffracted rays is given by the grating formula:

$$d = \frac{N\lambda}{2n\sin\theta} = \frac{N\lambda}{2NA} \tag{5.1}$$

where d = grating spacing
n = refractive index in image space
λ = wavelength of the light
N = diffraction order
θ = angle of the ray of order *N* emerging from the grating
NA = numeric aperture

The grating spatial frequency (ν = 1/2d) corresponding to the first-order (N = 1) diffracted peak is then given as:

$$\nu = \frac{1}{2d} = \frac{n\sin\theta}{\lambda} \tag{5.2}$$

For imaging with a lens, the lens needs to be able to capture the light, so it is required that $\theta = \theta_{max}$, where θ_{max} is defined by the numerical aperture NA of the lens. The latter is given by:

$$NA = n\sin\theta_{max} \tag{5.3}$$

This is illustrated in Figure 5.10, where the aim is to image a fine mask pattern onto a thin layer of photosensitive photoresist. The NA of the lens in a medium of refractive index *n* (1.0 in air) defines the angle of acceptance, $2\theta_{max}$, of the cone of diffracted light from an object that the lens can accept. It lies between 0 and 1. An NA of 0 means that the lens gathers no

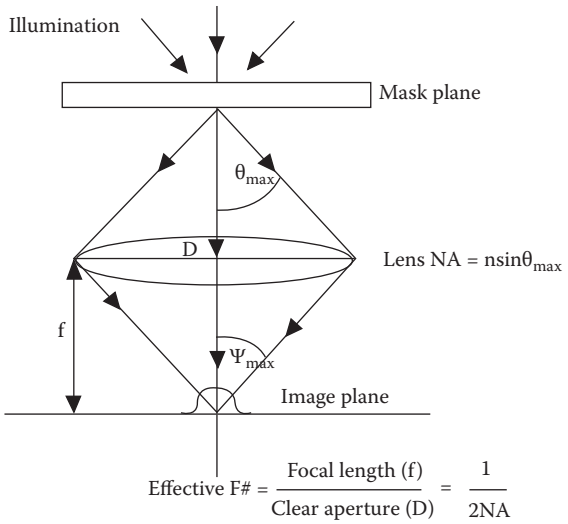


FIGURE 5.10 Relationship between the object, image, focal length, and diameter of a lens to define the numerical aperture. An important application is projection lithography, where the image on the mask is imaged onto a photoresist (see Volume II, Chapter 1).

light; an NA of 1 means that the lens gathers all the light that falls onto it. Hence the numerical aperture characterizes the ability of a lens to transmit light. It is proportional to the size of the lens, lens diameter *D*, and inversely proportional to effective *F* number (*F#*), defined as *f*/*D*, or the focal length *f* divided by the lens diameter *D*:

$$NA = n\sin\theta_{max} = \frac{D}{2f} = \frac{1}{2F\#} \tag{5.4}$$

The assumptions here are a large *F#*, which implies that θ is small so that NA = sin θ ~ tan θ ~ θ, a situation called the *paraxial approximation*.*

At angles larger than θ_{max}, the imaging lens no longer captures light. Therefore, the highest grating spatial frequency that can be imaged by a coherent illumination system is given by:

$$\nu_{max} = \frac{1}{2d_{min}} = \frac{n\sin\theta_{max}}{\lambda} = \frac{NA}{\lambda} \tag{5.5}$$

Consequently, resolution *R* (defined here as d_{min}, i.e., slit or a line) in this case is given by:

$$R = d_{min} = \frac{\lambda}{2NA} \tag{5.6}$$

* To understand all the simplifications involved in this result consult Meinhart and Wereley (2003).[53]

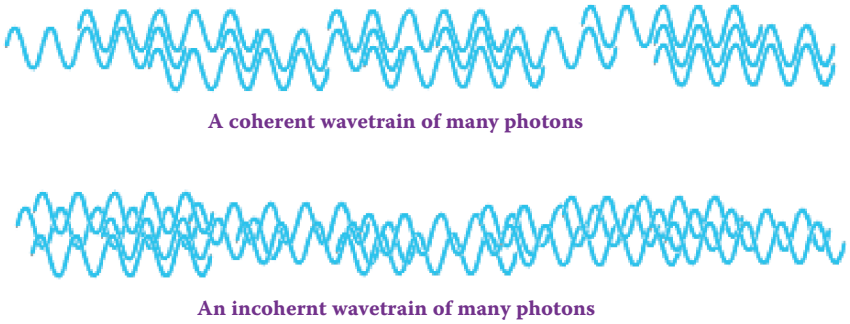


FIGURE 5.11 Spatial coherence of light.

In the case of incoherent illumination (see spatial coherence in Figure 5.11) of the same grating, each ray is diffracted by the grating and forms its own image in the wafer plane. The direction of the first diffraction peak (N = 1) for incoherent rays, incident at an angle *i*, is given by a more general grating equation:

$$d = \frac{\lambda}{2n\,(\sin\theta + \sin i)} = \frac{\lambda}{2NA} \qquad (5.7)$$

representing the path difference for light passing through adjacent slits. For light incident normal to the grating (sin i = 0), Equation 5.7 reduces to Equation 5.1. For image formation it is now required that both *i* and $\theta = \theta_{max}$. Therefore:

$$\nu_{max} = \frac{1}{2d_{min}} = \frac{2n\sin\theta_{max}}{\lambda} = \frac{2NA}{\lambda} \qquad (5.8)$$

or:

$$R = d_{min} = \frac{\lambda}{4NA} \qquad (5.9)$$

so that the resolution for incoherent light is twice better (smaller features can be resolved) than the resolution of coherent light.

Rayleigh's Limit

The above imaging resolutions for coherent and incoherent light are theoretical limits. Let us calculate the practical limit to the resolution of a lens. Light passing through a circular aperture interferes with itself, creating ring-shaped diffraction patterns known as an *Airy pattern*, which blurs the image (see Figures 5.13 and 5.14). It is named for George Biddell Airy, a British astronomer, who first explained the phenomenon. In 1872, Rayleigh proposed his resolution criterion based on this Airy pattern. This criterion defines two identical incoherent light sources, separated by a distance *d*, as resolved when the maximum of the Airy disc from one point source falls on the first zero of the intensity distribution of the Airy disc from a second point source. The diffraction for a circular aperture* is given as:

$$E = \frac{2J_1\left(\frac{\pi D\sin\theta}{\lambda}\right)}{\left(\frac{\pi D\sin\theta}{\lambda}\right)} \text{ and } I = E^2 = \left[\frac{2J_1\left(\frac{\pi D\sin\theta}{\lambda}\right)}{\left(\frac{\pi D\sin\theta}{\lambda}\right)}\right]^2 \qquad (5.10)$$

where E = electrical field
I = intensity
J_1 = a Bessel function of the first kind of order 1 in *x* with $x = \pi D\sin\theta/\lambda$ (see Figure 5.12)
D = diameter of the aperture
λ = wavelength of light

The expression of $I(r)$ is called the *point spread function* (PSF), and the radius of the Airy disc is derived from it.

The PSF extends to infinity, but the "width" of the central intensity peak can be described from its first minimum where J_1 is zero, i.e., at x = 3.8 (see Figure 5.12). The radius of the PSF(r), at a distance *s* of the aperture, is defined by the radius of the first dark ring in the image plane, corresponds to the radius of the Airy disc, and is calculated as:

* Exercise: In Chapter 2 we calculated the diffraction of x-rays from a crystal (see Equation 2.39). Demonstrate that Equations 2.39 and 5.10 are identical expressions.

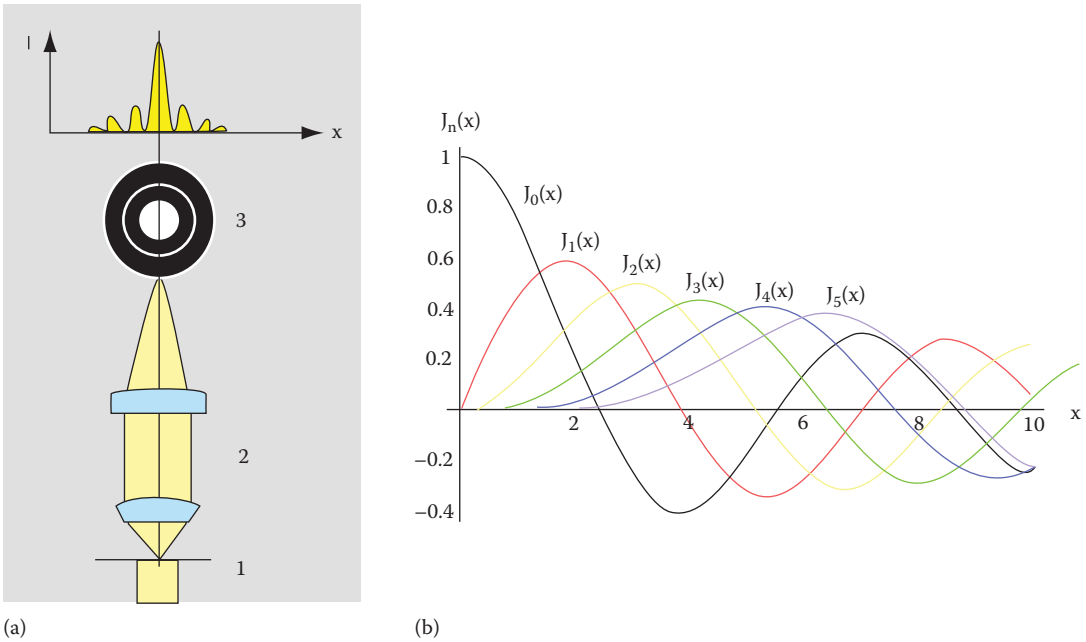


FIGURE 5.12 (a) 1) pinhole, 2) lens system, and 3) point spread function (PSF). (b) Bessel functions.

$$J_1 = 0 \text{ when } x = \frac{\pi D \sin\theta}{\lambda} = 3.8 \text{ or, for small } \theta, \tag{5.11}$$

$$D = \frac{1.22\lambda}{\theta} \text{ and since } \theta \cong \frac{r}{s}$$

$$r = \frac{1.22\lambda s}{D}$$

Fraunhofer diffraction patterns are obtained by moving the screen far enough away from the aperture. To obtain the Fraunhofer diffraction pattern at the focal point of a lens, we set the lens against the aperture and use the focal length (f) of the lens as the aperture to screen distance (i.e., replace s with f in Equation 5.11), or:

$$r = \frac{1.22\lambda f}{D} \tag{5.12}$$

This is illustrated in Figure 5.13. Or, based on Equation 5.4, for an ideal lens system of focal length f, the Rayleigh criterion yields a minimum spatial resolution as:

$$d = \frac{1.22\lambda}{2NA} \tag{5.13}$$

In this expression d is the radius of the smallest object that the lens can resolve and also the radius of the smallest spot on which a collimated beam of light can be focused. The Rayleigh criterion is illustrated in Figure 5.14.

In Figure 5.15 we show a schematic of a single lens with diameter D imaging a small circular aperture. By definition, the magnification M of a lens is given by:

$$M = \frac{S_i}{S_0} \tag{5.14}$$

where S_0 is the object distance and S_i is the image distance. Combining this expression for magnification with the Gaussian lens formula we obtain:

$$\frac{1}{S_i} + \frac{1}{S_0} = \frac{1}{f} \tag{5.15}$$

which leads to:

$$S_i = (M + 1)f \tag{5.16}$$

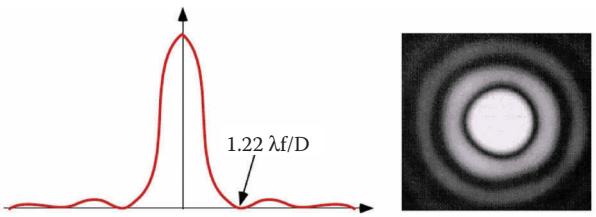


FIGURE 5.13 Image by a point source forms a circle with diameter of 2.44 $\lambda f/D$ (=2 × r) surrounded by diffraction rings (Airy pattern).

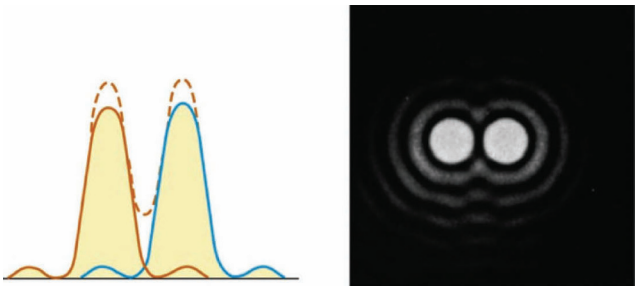


FIGURE 5.14 Rayleigh criterion: two light spots are considered resolved if the intensity minimum of one coincides with the maximum of the other $\left(d = \frac{1.22\lambda}{2NA}\right)$.

where f is the focal length of the lens. If the object distance (S_0) and the image distance (S_i) are set equal to the focal lengths, the magnification can be written in terms of the focal lengths, or:

$$M = \frac{S_i}{S_0} = \frac{f_i}{f_0} \tag{5.17}$$

In the case $S_i = S_0$ and $f_i = f_0$, the magnification M obviously equals 1. The F number ($F\#$) of such a setup is then by definition given as $F\# = f_0/D = f_i/D$ (see Equation 5.4); its numerical aperture is $NA_0 = NA_i = NA$, and these factors are linked with the magnification M via:

$$F\# = \frac{f_0}{D} = \frac{f_i}{D} = \frac{1}{2NA_0} = \frac{1}{2NA_i} = \frac{(1+M)f}{D} \tag{5.18}$$

D is the diameter of the lens or the diameter of the light beam, whichever is the smaller of the two.

The larger the NA of the projection lens in the lithography system of Figure 5.10, the greater the amount of light (containing diffraction information of the mask) collected and subsequently imaged. Because the image is constructed from diffracted light, and the collection of higher orders of diffracted light enhances the resolution of the image, a larger NA, allowing a larger acceptance angle, results in a better lithography resolution. Therefore, for a set wavelength, the resolution in a lithography system has traditionally been improved by increasing the NA of the optical system. This is usually achieved by increasing the lens diameter D. Lately, work is also directed at increasing NA by increasing the refractive index n of the medium surrounding the imaging lens (Equation 5.3). The latter approach is exploited in immersion lithography, an advanced lithography approach with the potential to keep the semiconductor industry on track with Moore's law. The physical limit to NA for lithography exposure systems using air as a medium between the lens and the wafer is 1 because the sine of any angle is always ≤1 and n = 1 for air. What happens, though, if a medium with a higher index of refraction is substituted for air, as is often done in microscopy, where oil is placed between the lens and the sample being viewed for resolution? Abbe and Zeiss developed these oil immersion systems using oils that matched the refractive index of glass. Thus, they were able to reach a maximum NA of 1.4, allowing light microscopes to resolve two points distanced only 0.2 μm apart. In Volume II, Chapter 1, we will further detail immersion lithography.

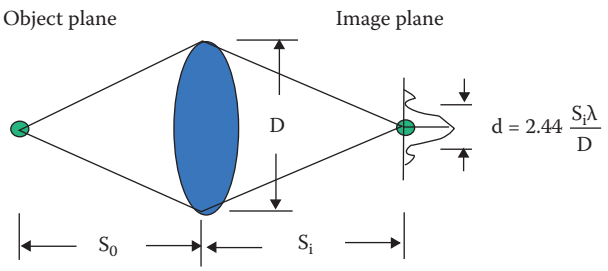


FIGURE 5.15 Schematic of a single thin lens with aperture diameter, D. The object distance is S_0, and the image distance is S_i. The point spread function through a circular aperture is the well-known Airy function of width or diameter, $d = 2.44S_i\lambda/D$ (see Equation 5.13 with $D = 2r$).

Refraction

Snellius Law

Geometric optics, also ray optics, has light propagate as "rays." These rays are bent at the interface between two dissimilar media. The "ray" is an abstract object, perpendicular to the wavefronts of the electromagnetic (EM) waves. Geometric optics provides rules for propagating these rays through an optical system, revealing how actual wavefronts propagate. Ray optics is a significant simplification of optics and fails to account for many important optical effects, such as diffraction and polarization. The latter phenomenon will only be understood properly after we introduce Maxwell's equations.

Light refracts, which means that it bends when passing from one medium to another. Thus, refraction is the change in direction of a ray of light passing from one transparent medium to another one with a different optical density. When light enters a denser medium (say, water) coming from one that is less dense (say, air), it bends toward a line normal to the boundary between the two media as illustrated in Figure 5.16. The deviation is greater for shorter wavelengths.

The ratio of the speed of the light going from one material into another material is proportional to the ratio of the sine of the angle of incidence and the sine of the angle of refraction as illustrated in Figure 5.16. This is the so-called *Snellius law* [1621; also Willebrord Snell (1580–1626)]:

$$\frac{\sin(i)}{\sin(r)} = \frac{n_1}{n_2} \tag{5.19}$$

with i being the angle of incidence and r the angle of refraction, n_1 is the index of refraction of material 1, and n_2 is the index of refraction of material 2. Two familiar examples to demonstrate refraction are a pencil in a glass of water (Figure 5.17) and a rainbow (Figure 5.18). In Figure 5.17 we show a pencil in a glass: first in an empty glass, then in one with water (positive refractive index), and then in a glass filled with a negative refractive index material.

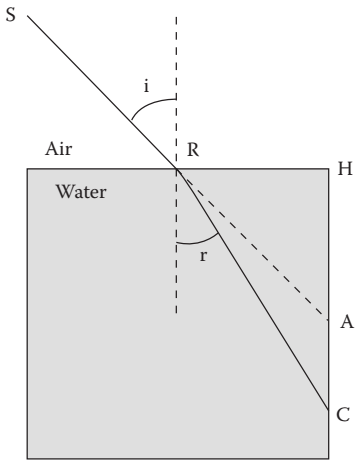


FIGURE 5.16 Snellius law illustrated: the ratio of the horizontal length, *RH*, to the diagonal, *RA*, is just the sine of the angle *RAH*, which is equal to the angle of incidence *i* measured from the line drawn perpendicular to the refracting surface at *R*. Likewise, the ratio *RH*/*RC* is just the sine of the angle of refraction *RCH*, or *r*. Thus, Snell's ratio *RC*/*RA* is the same as (sin *i*)/(sin *r*).

At the boundary between air and water, the light wave changes direction; its wavelength increases or decreases, but its frequency remains constant. We will see in the next section that the refractive index of a medium is related to the speed of light and that in a denser medium a "slow-down" of the light occurs relative to the speed of light in free space. When light slows down it appears bent. The case of negative refractive index water (Figure 5.17c) will be discussed further below under "The μ-ε Quadrant and Metamaterials." A bright vividly colored rainbow as shown in Figure 5.18 is caused by dispersion of sunlight refracted by raindrops between 0.1 and 2 mm in diameter where geometric optics applies. Complex Mie scattering effects are important for smaller drops, which do produce white fog (see further below). The light is first refracted as it enters the surface of the raindrop, reflected off the back of the drop, and again refracted as it leaves the drop. This angle is independent of the size of the drop but depends a lot on its shape and refractive index.

Definition of Refractive Index

Maxwell demonstrated that for electromagnetic (EM) waves the wave velocity v in a homogeneous, isotropic dielectric medium is determined by the magnetic permeability μ, and the dielectric constant (or permittivity), ε (for the detailed derivation see further below):

$$v = \frac{1}{\sqrt{\varepsilon\mu}} \tag{5.20}$$

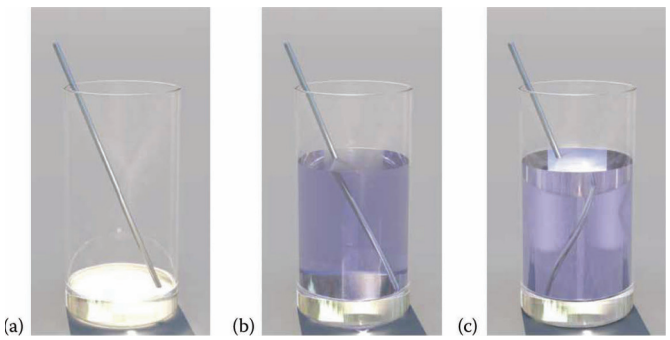


FIGURE 5.17 Refraction illustrated. (a) Empty glass: no refraction; (b) typical refraction with pencil in water with n = 1.3; (c) what would happen if the refractive index were negative with n = –1.3 (see "The μ-ε Quadrant and Metamaterials"). (From Gennady Shvets, The University of Texas at Austin.)

FIGURE 5.18 A rainbow is caused by dispersion of sunlight refracted by raindrops.

For a wave traveling in free space (where we use the permeability and dielectric constant for vacuum μ_0 and ε_0 instead of μ and ε) $v = c = 2.998 \times 10^8$ m/s. The "slow-down" factor, relative to free space, is by definition the refractive index n:

$$n = \frac{c}{v} \tag{5.21}$$

or also:

$$n^2 = \varepsilon\mu \tag{5.22}$$

For example, because the speed of light in air is $v = 1/\sqrt{\varepsilon_{air}\mu_{air}}$ (with the permeability and dielectric constant for air given by μ_{air} and ε_{air}, respectively), the refractive index of air is given by:

$$n = \frac{c}{v} = \frac{\sqrt{\varepsilon_{air}\mu_{air}}}{\sqrt{\varepsilon_0\mu_0}} = 1.003 \tag{5.23}$$

For most materials $\mu/\mu_0 = 1$, so the expression for n is often written out as $n = \frac{c}{v} = \frac{\sqrt{\varepsilon_{air}}}{\sqrt{\varepsilon_0}}$). In Chapter 3, we saw how the dielectric constant ε in the most general terms is a dielectric function dependent on both frequency (ω) and wave vector $\mathbf{k}$: $\varepsilon(\omega,\mathbf{k})$. There is also a dispersion function for the magnetic permeability [$\mu(\omega, \mathbf{k})$]; thus, from Equation 5.23 we can expect the refractive index n to be frequency-dependent as well. When a material is subject to an applied electric field $\mathbf{E}$, the internal charge distribution is distorted, which generates an electric dipole moment per unit volume called the *electric polarization* $\mathbf{P}$. After introducing the Maxwell equations, we will link the dielectric function to this polarization $\mathbf{P}$ and to the dispersion of the refractive index n for metals, semiconductors, and dielectrics.

Dispersion of light is most familiar with respect to dispersion of white light on a prism. If a beam of white light enters a glass prism, what emerges from the other side is a spread-out beam of many-colored light. The various colors are refracted through different angles by the glass and are "dispersed," or spread out.

Reflectance and Total Internal Reflectance

To illustrate reflectance and total internal reflectance (TIR) we show a sequence of four images from ray optics in Figure 5.19. A light beam goes from medium 1 (say, water with refractive index n_1) into a medium 2 (say, air with refractive index n_2) at different angles θ_1. When the light coming from the water strikes the surface, part will be reflected and part will be refracted. Measured with respect to the normal line perpendicular to the surface, the reflected light comes off at an angle equal to that at which it entered, whereas that for the refracted light is larger than the incident angle. In fact, the greater the incident angle, the more the refracted light bends away from the normal.

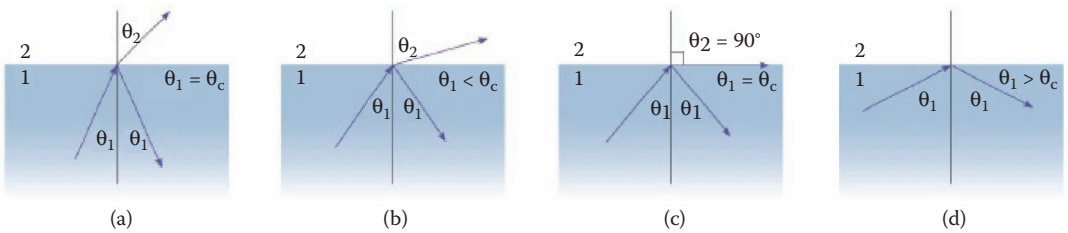


FIGURE 5.19 Illustration of reflection and total internal reflectance (TIR). Material 1 (bottom) has the higher index of refraction ($n_1 > n_2$). For explanation of (a)–(d), see text.

Thus, increasing the angle of incidence from the path shown in Figure 5.19a to the one shown in Figure 5.19b will eventually reach a point where the refracted angle is 90°, at which point the light appears to emerge along the surface between the water and air (Figure 5.19c). If the angle of incidence is increased further, the refracted light cannot leave the water (Figure 5.19d). It gets completely reflected. The interesting thing about TIR is that 100% of the light gets reflected back into the denser medium, as long as the angle at which it is incident to the surface is large enough. At a given critical angle when $\theta_1 = \theta_c$ TIR occurs (Figure 5.19c); the light cannot escape medium 1 with refractive index n_1 anymore, except for some really weak evanescent wave. Any light incident at an angle $\theta_1 > \theta_c$ is also totally reflected (Figure 5.19d). Thus, TIR is an effect that combines both refraction and reflection. The critical angle, also called the *Brewster angle*, named after the Scottish physicist Sir David Brewster (1781–1868), is given by:

$$\sin\theta_c = \frac{n_2}{n_1} \tag{5.24}$$

A common and useful application of TIR is in fiberoptics to keep light beams focused inside a fiber without significant loss (see Figure 5.20).

Further below, after launching the Maxwell equations, we calculate the reflectance R in terms of the refractive index n.

Light Polarization

Polarized light is light that has its electric vector (**E**) oriented in a predictable fashion with respect to the propagation direction. Ordinary white light is unpolarized and made up of numerous orthogonal electrical and magnetic field pairs (**E**, **B**) that fluctuate at all possible angles around the propagation direction. Even in short time intervals, it appears to be oriented in all directions with equal probability. Light is considered to be "linearly polarized" when it contains waves that only fluctuate in one specific plane (see Figure 5.21). Linear polarization "tries to" induce a linear charge motion, but the Lorentz force prevents this from being purely linear. For example, charge oscillation along the x-axis introduces a small E_y component (field rotation. Specifying the orientation of one of the field vectors (**E** or **B**) determines the polarization of the EM wave. The plane containing the **E**-field vector defines the polarization of the EM wave. Of course, the **B**-field vector would then always lie in a plane perpendicular to this plane containing the **E**-field vector. But according to theoretical and experimental evidence, it is the electric vector (**E**) rather than the magnetic vector (**B**) of a light wave that is responsible for all polarization effects. Therefore, the electric vector of a light wave, for all practical purposes, can be identified as the light vector in this respect.

The simplest way of producing linearly polarized light is by reflection from a dielectric surface. In the preceding section we learned that when light strikes a material such as glass, part of the light might get reflected and part of it refracted. Reflection also causes the light to become unpolarized, partially polarized, or completely polarized, depending on the angle at which the light strikes the surface, or the angle of incidence (see Figure 5.22). The light

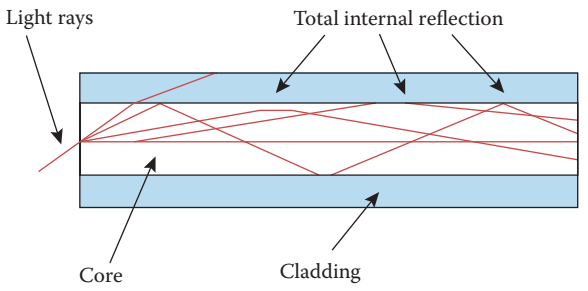


FIGURE 5.20 A fiberoptic is based on the principle of total internal reflectance.

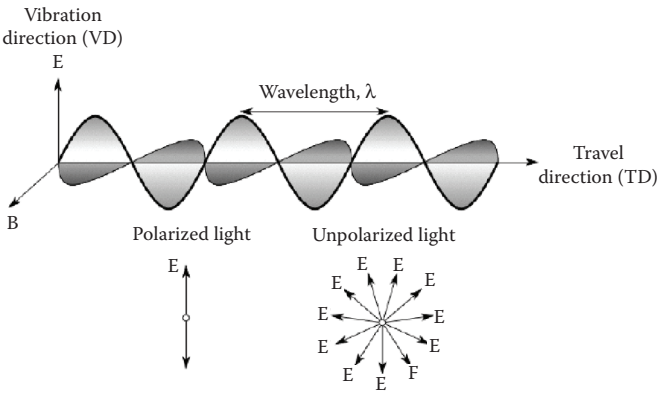


FIGURE 5.21 Unpolarized and polarized light.

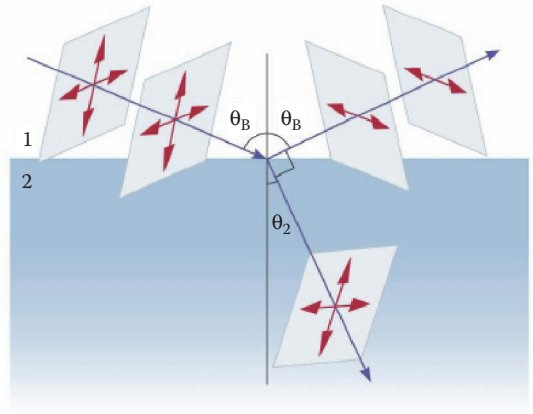


FIGURE 5.22 Reflection causes the light to become either unpolarized, partially polarized, or completely polarized, depending on the angle at which the light strikes the surface, or the angle of incidence. The angle θ_B is known as the *polarizing angle*, or the *Brewster angle*.

is unpolarized only if the incident light angle is 0°, which does not occur often. It is more common that polarization does occur, and the greater the angle of incidence, the greater the polarization. When the light strikes at an angle θ_B of 90° or perpendicular to the surface, the light is completely polarized. At this particular angle of incidence the reflectivity for light whose electric vector is in the plane of incidence becomes zero. Thus, the reflected light is linearly polarized at right angles to the plane of incidence. This angle θ_B is known as the *polarizing angle*, or the *Brewster polarizing angle*, because it was discovered by the same David Brewster of the critical angle θ_c (see previous section):

$$\tan\theta_B = \frac{n_2}{n_1} \tag{5.25}$$

Certain natural materials absorb linearly polarized light of one vibration direction much more strongly than light vibrating at right angles. Such materials are termed dichroic. Tourmaline is one of the best-known dichroic crystals, and tourmaline plates were used as polarizers for many years. Dichroic crystals were studied by Edwin Land in 1930; Land invented a polarizing material that would be called *Polaroid* and give a huge benefit to the camera industry.

Polarization of light is also possible by refraction, in which case the light is bounced off of numerous surfaces, instead of only one. Often a stack of thin films of a transparent material is used for this purpose. The beam of light that is refracted is continuously polarized because refraction causes polarization. Because of this polarization, the beam of light that is refracted is shown to increase in intensity. Polarization obviously occurs frequently: even the light from a rainbow is polarized because the angle of incidence of light off a water beam is similar to a Brewster angle.

Other natural materials exist in which the velocity of light depends on the vibration direction. These materials are called *birefringent*. We will discuss these in more detail after we have introduced the Maxwell equations.

Circular (and elliptical) polarization is possible because the propagating electric (and magnetic) fields can have two orthogonal components with independent amplitudes and phases (and the same frequency). The electric vector, at one point in time, describes a helix along the direction of wave propagation. Circular polarization may be referred to as *right* or *left*, depending on the direction in which the electric field vector rotates. The magnitude of the electric field vector is constant as it rotates. Circular polarization is a limiting case of the more general condition of elliptical polarization. A circularly polarized wave may be resolved into two linearly polarized waves, of equal amplitude, in phase quadrature (90° apart) and with their planes of polarization at right angles to each other.

Such polarization occurs in chiral* materials, for example, with helical polarizable molecules in a solution.

Maxwell's Equations

Introduction

As stated earlier, ray optics is a significant simplification of optics and fails to account for many important optical effects such as diffraction and polarization. The latter phenomena are only understood properly from Maxwell's equations. Light

* Definition of chirality by Kelvin: "I call any geometrical figure, or group of points, chiral, and say it has chirality, if its image in a plane mirror, ideally realized, cannot be brought to coincide with itself."

with its wave and particle character is not the easiest of natural phenomena to describe. Its ambivalent nature requires two complementary models for a satisfactory description of its properties. In Chapter 3, we mentioned how Dirac's quantum field theory removed the paradox of particle-wave duality: it showed that if a particle was probed in a way that was meant to demonstrate its particle-like properties, it would appear to be a particle, and if it was probed in a way that was meant to demonstrate its wave-like properties, it would appear to be a wave. Maxwell's electromagnetic (EM) wave model, the topic of this chapter, follows logically from Dirac's model and describes light as a continuous transfer of energy through space by a combination of waves of electrical and magnetic fields.

The Scottish physicist James Clerk Maxwell (1831–1879) unified in one magnificent theory—four elegant equations—all that was known of electricity and magnetism. For the first time it became apparent that EM fields could travel through space, and the theory suggested the revolutionary concept that light was nothing but an EM wave itself. Maxwell's equations are even more fundamental than Newton's three laws of motion because they agree with Einstein's relativity theory, whereas Newton's laws do not.

The refractive index of a material as introduced above comes about through the interaction of light with many, many atoms (10^{22} atoms/cm^3), and it is impossible to describe the individual response of so many electrons and atoms. So we have to make do with "effective" materials properties, such as the refractive index, an average over many atoms. This works very well for many materials because atom spacings in solids are much smaller than the wavelength of the light at hand. For visible light, e.g., green light, a volume of λ^3 corresponds to a volume of ~ (500 nm)^3 ~ (5 10^{-5} cm)^3 ~ 10^{-13} cm^3, and the number of atoms in such a volume ~$10^{-13} \times 10^{22} = 10^9$! So working with spatially averaged fields will give good results in many cases.

Before we launch into the Maxwell's equations for plane wave EM radiation, we remind the reader of some important vector relations. Because both the electric field and the magnetic field have a direction, we describe light as vector fields. Two important integral theorems we need for manipulating those equations are Gauss' divergence theorem and Stokes' theorem. We also introduce Ampère's law, Gauss' electric and magnetic laws, and Faraday's law, as they are the foundation for Maxwell's famous unifying equations.

Gauss and Stokes Theorems

The first theorem we introduce, the divergence theorem, comes from vector calculus and is also known as *Gauss'* (1777–1855) or *Green's* (1783–1841) *theorem*. The divergence of a vector field is a scalar (a number) and represents the net outward flux per unit volume (one can think of it as a "source density"). The divergence theorem relates volume and surface integrals and expresses the fact that the volume integral of the divergence of any conservative vector (or tensor) field equals the total outward flux of the vector through the surface that bounds the volume (http://mathworld.wolfram.com/ConservativeField.html). More simply stated: the sum of all sources minus the sum of all sinks gives the net flow out of a region. Mathematically:

$$\int_v \nabla \cdot \mathbf{A} dv = \oint_s \mathbf{A} \cdot d\mathbf{s} \qquad (5.26)$$

where **A** is a vector field or tensor field (for example, **A** could be a stress tensor $\underline{\sigma}$ or an electric field vector **E**); d**s** is a differential vector element of surface area *S*, with infinitesimally small magnitude and direction normal to surface *S*; and *dv* is a differential element of volume *V* enclosed by surface *S*. We remind the reader that in vector analysis $\nabla \cdot \mathbf{A}$ is the same as div **A**. This equation applies to any volume *V* bounded by surface *S*. Effectively this comes down to saying that "field lines are conserved." We see where field lines originate on the left side of the equation or count them as they leave through the surface on the right side of the equation (Figure 5.23).

The second theorem we need to refamiliarize ourselves with is Stokes' theorem, or the curl theorem. It is named after Sir George Gabriel Stokes (1819–1903), although the first known version of the theorem is by William Thomson (Lord Kelvin). Stokes' theorem deals with the problem of a 3D

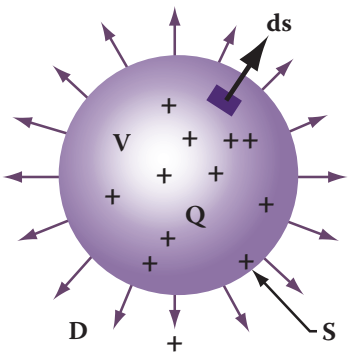


FIGURE 5.23 The divergence or Gauss' theorem illustrated for the case where **A** is equal to **D**, the electric displacement which is related to the electric field **E** via a materials-dependent constant called the *permittivity*, ε (as given by the constitutive relation, $\mathbf{D} = \varepsilon \mathbf{E}$; Equation 3.38 in Volume I). Consider a total amount of charge *Q* in a volume *V* bounded by the closed surface *S*. The area integral gives a measure of the net charge enclosed, and the divergence of the electric displacement gives the density of the sources (see Equations 5.34–5.36 below).

curve in space, and the line integral around such a curve; it relates a surface integral to a line integral:

$$\int_S (\nabla \times \mathbf{A}) \cdot d\mathbf{s} = \oint_C \mathbf{A} \cdot d\mathbf{l} \qquad (5.27)$$

where *S* is an open surface bounded by a closed curve *C*, and $d\mathbf{l}$ is a differential vector element of path length tangential to contour *C* enclosing surface area *S*. The vector $d\mathbf{s}$ is an outward surface normal vector of magnitude *ds*. Stokes' theorem says that if *S* is a surface in three dimensions having a closed curve *C* as its boundary, then the surface integral of the curl of a vector field over that open surface equals the closed line integral of the vector along the contour *C* bounding the open surface. Integrating the curl across a surface gives the same answer as integrating the tangential vector components along the line bounding that surface. Also, rotation in adjacent spots cancels out, and what is not canceled is present at the edge of the surface. This is a fairly remarkable theorem because *S* can be any surface bounded by the curve *C* and have just about any shape (Figure 5.24). We remind the reader that in vector analysis $\nabla \times \mathbf{A}$, rot **A**, and curl **A** are equivalent notations. Another way of saying this is that the curl of a field defines the "amount of rotation" in a field.

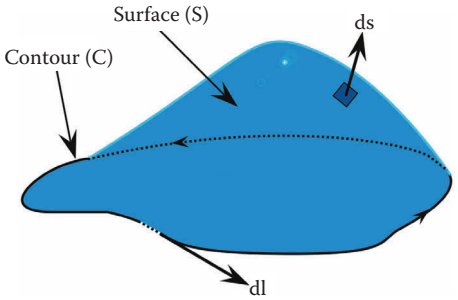


FIGURE 5.24 Illustrating Stokes' law. *S* is an open surface bounded by a closed curve *C*.

With Gauss' and Stokes' theorems Maxwell's equations can be transformed from differential format to integral, and vice versa.

Ampère's Law, Gauss' Electric and Magnetic Laws, and Faraday's Law

Ampère's Law

Ampère's law states:

> The magnetic circulation around a closed loop is proportional to the net electric current flowing through the loop, including the actual flow of electric charge as well as the displacement current, which is the time derivative of the electric flux. The positive directions of circulation, current and flux are related by the right-hand rule.

An electric current *J* running through a straight wire generates a magnetic field strength **H** (in A/m) around that wire (Figure 5.25). The integral of H along the circular path *C* is equal to the current *I*, which in turn can be obtained as the integral

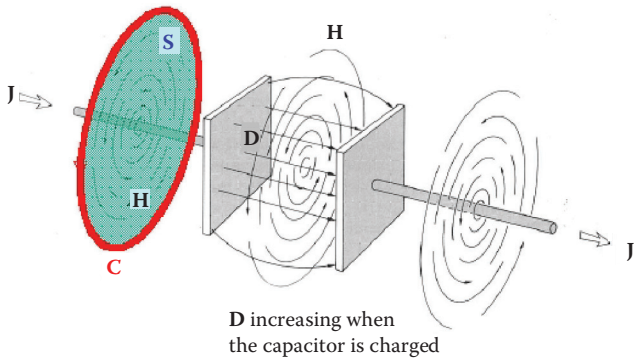


FIGURE 5.25 Ampère's law illustrated. The magnetic field is induced by electrical currents (**J**) and changes in electric field, $\frac{\partial \mathbf{D}}{\partial t}$ (the displacement current is also called *Maxwell's displacement* **D**).

of the current density **J** across the cross-section of the wire:

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = I = \int_S \mathbf{J} \cdot d\mathbf{s} \tag{5.28}$$

This is the original Ampère's Law (1775–1836).

Using Stokes' theorem (Equation 5.27) we transform Equation 5.28 to:

$$\int_S (\nabla \times \mathbf{H}) \cdot d\mathbf{s} = I = \int_S \mathbf{J} \cdot d\mathbf{s} \tag{5.29}$$

From the latter expression it follows that:

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{5.30}$$

James Clerk Maxwell made an important correction to Ampère's law; he recognized that a changing electric field, $\frac{\partial \mathbf{D}}{\partial t}$, is equivalent to a current **J**, and he called $\frac{\partial \mathbf{D}}{\partial t}$ the displacement current. The electric flux density, also called *Maxwell's displacement* **D**, is expressed in units of As/m^2. The more general Ampere's law, in which a changing electric field $\left(\frac{\partial \mathbf{D}}{\partial t}\right)$ and a current (**J**) both produce a magnetic field (**H**), is then given as:

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{5.31}$$

The electric displacement **D**, expressed in coulombs per square meter (C/m^2) is related to the electric field **E** (in units of V/m) via the permittivity, $\varepsilon(\omega,\mathbf{k})$, as:

$$\mathbf{D} = \varepsilon(\omega,\mathbf{k})\mathbf{E} \tag{3.37}$$

The permittivity ε (As/Vm) is a scalar if the medium is isotropic or a 3 × 3 matrix otherwise, and its frequency dependence defines the so-called *dielectric function*, i.e., $\varepsilon(\omega,\mathbf{k})$. To obtain $\varepsilon_r(\omega,\mathbf{k})$, the relative dielectric response of a material, one divides by ε_0 because $\varepsilon(\omega,\mathbf{k})/\varepsilon_0 = \varepsilon_r(\omega,\mathbf{k})$.

Similarly, the magnetic field or flux density **B** is linked to the magnetic field strength **H** (in A/m) within the material through a second constitutive relation:

$$\mathbf{B} = \mu\mathbf{H} \tag{5.32}$$

with $\mu = \mu_0\mu_r$ and μ_r the dimensionless relative permeability. The magnetic permeability μ is a tensor that reduces to a scalar in isotropic materials. The magnetic inductive capacity, μ, or permeability, is related to the ability of a medium, such as air, to store magnetic potential energy. The permeability of vacuum, μ_0, is 1.257×10^{-6} Hm^{-1}, and in the case of air is $\mu_{air} = 1.260 \times 10^{-6}$ Hm^{-1}. As in the case of the dielectric "constant" the response of the relative magnetic permeability μ_r of normal media to externally applied fields depends on the frequency of the field as well and on the wave vector, i.e., $\mu_r(\omega,\mathbf{k})$.

A third constitutive relation, the basis of Ohm's law (V = IR), introduced in Chapter 2, links current density **J** (A/m^2) to electrical conductivity σ (in A/Vm) of the material:

$$\mathbf{J} = \sigma\mathbf{E} \tag{2.2}$$

In the most general case, the electrical conductivity σ is a second-rank tensor and is also frequency-dependent, i.e., $\sigma(\omega,\mathbf{k})$. For an isotropic material, σ becomes a scalar.

Gauss' Law for Electric Fields

Gauss' law for electric fields states:

> The net electric flux exiting a closed surface is proportional to the net electric charge enclosed by the surface.

Gauss' law is an extended form of Coulomb's law. Both Coulomb's and Gauss' laws relate electric fields to charges: charges make electrical fields. Consider a total amount of charge *Q* in a volume *V* bounded by a closed surface *S*. The electric charger *Q* equals the surface integral of the electric flux density over the closed surface *S*:

$$\oint_S \mathbf{D} \cdot d\mathbf{s} = Q \tag{5.33}$$

With the divergence theorem this can be rewritten as:

$$\oint_S \mathbf{D} \cdot d\mathbf{s} = \int_v \nabla \cdot \mathbf{D} dv \tag{5.34}$$

or also:

$$\int_V \nabla \cdot \mathbf{D} dv = Q = \int_V \rho_V dv \tag{5.35}$$

and finally one obtains:

$$\nabla \cdot \mathbf{D} = \rho_v \tag{5.36}$$

where ρ_V is the electric charge per unit volume (C/m^3). For linear materials, $\mathbf{D} = \varepsilon \mathbf{E}$ (Equation 3.38) and Equation 5.36 can be rewritten as:

$$\nabla \cdot \mathbf{E} = \frac{\rho_V}{\varepsilon} \tag{5.37}$$

From Equation 5.37 we can derive the so-called *Poisson equation* for the potential:

$$\nabla^2 V = -\frac{\rho v}{\varepsilon} \tag{5.38}$$

because $\mathbf{E} = -\nabla V$. In the case that the total charge density is zero, Equation 5.38 reduces to Laplace's Equation:

$$\nabla^2 \mathbf{V} = 0 \tag{5.39}$$

marking a constant electric field.

Gauss' Law for Magnetic Fields

Gauss' law for magnetic fields states:

> The net magnetic flux exiting a closed surface is always zero.

We know that magnetic poles always occur in pairs (north and south). Therefore, a magnetic field will be produced in space if there is a changing electric field, but this magnetic field is changing as the electric field is changing. A changing magnetic field produces an electric field that is changing—in other words, we have a self-perpetuating system. Because a single magnetic charge does not exist, the equivalent expression to Equation 5.33 for magnetism reads:

$$\oint_S \mathbf{B} \cdot d\mathbf{s} = 0 \tag{5.40}$$

This means that the magnetic field lines must form closed loops, i.e., the mathematical formulation of the fact that the field lines cannot originate from somewhere or that there are no magnetic monopoles. Using Gauss' theorem we obtain:

$$\oint_S \mathbf{B} \cdot d\mathbf{s} = \int_V \nabla \cdot \mathbf{B} dv = 0 \tag{5.41}$$

In other words:

$$\nabla \cdot \mathbf{B} = 0 \tag{5.42}$$

Faraday's Law

Faraday's law states:

> The electric circulation (or emf) around a closed loop is proportional to the negative time derivative of the magnetic flux through the loop, where the positive circulation and flux directions are related by the right-hand rule.

Faraday (1791–1867) discovered that a voltage V (or emf) is induced in a coil when a magnetic flux Φ [Webers (Wb)] in the environment changes with time, and he formulated a law stating that a changing magnetic field $\left(\frac{\partial \mathbf{B}}{\partial t}\right)$ produces an electric field ($\mathbf{E}$):

$$\oint_C \mathbf{E} \cdot dl = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s} \tag{5.43}$$

With Equation 5.27 (Stokes) we rewrite this expression as:

$$\int_S (\nabla \times \mathbf{E}) \cdot d\mathbf{s} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s} \tag{5.44}$$

or:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{5.45}$$

Finally, we introduce the continuity equation for the conservation of charge. Say that in a time dt, a charge dQ is lost from volume V; because of this flux of charge, a current density $\mathbf{J}$ leaves the surface element ds. We use the divergence theorem to convert the integral over the surface S into an integral over the volume V. We assume that the volume does not change, so only the charge density ρ in the volume changes. This leads to the continuity equation:

$$\frac{\partial \rho_V}{\partial t} + \nabla \cdot \mathbf{J} = 0 \tag{5.46}$$

From Equation 5.46 the current density flux flowing out of a closed surface equals the rate of decrease of the positive charge density.

Example 5.1: Place some charges on a conductor with a conductivity σ; how long does it take for the charges to spread? From the continuity relation (Equation 5.46) we have $\nabla \cdot \mathbf{J} = -\frac{\partial \rho_V}{\partial t}$. In the conductor we have Ohm's law stating $\mathbf{J} = \sigma \mathbf{E}$

(Equation 2.2), and because $\sigma\nabla \cdot \mathbf{E} = -\frac{\partial \rho_V}{\partial t}$ and using Poisson's equation $\left(\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon}\right)$, the solution is:

$$\rho = \rho_0 e^{-\frac{t}{\tau}} \tag{5.47}$$

with the time constant $\tau = \varepsilon/\sigma$ With typical values for ε and σ one obtains a τ of 1.5 10^{-19} s. Thus, charge density decays exponentially with time. For a very good conductor, charges flow instantly to the surface to form a surface charge density and (for time-varying fields) a surface current is established. Inside a perfect conductor ($\sigma \to \infty$) and $\mathbf{E} = \mathbf{H} = 0$.

Maxwell's Elegant Summary

It was in 1864 that Maxwell summarized the above set of equations for all frequency ranges and size scales larger than atoms. He was the first to show that the four equations describing electric and magnetic fields could be rearranged to form a wave equation, with a speed of propagation equal to the speed of light. Maxwell also corrected Ampère's law by introducing an additional term for the displacement current (changing electric fields act like currents, likewise producing magnetic fields; see above). The elegant modern mathematical formulation introduced here is by Oliver Heaviside (1850–1925), who in 1884 reformulated Maxwell's equations using vector calculus. His more symmetric mathematical representation reinforces the physical symmetries between the magnetic and electrical fields.

Maxwell's equations, as remarked earlier, can be transformed from differential format to integral representation, and vice versa, by using the theorems from Gauss and Stokes. Remarkably, as we will see below, Maxwell's equations are perfectly consistent with the transformations of special relativity.

Maxwell's equations in the absence of electric charge density ρ (in units of $C/m^3 = 0$) and current $\mathbf{J}$ (in units of A/m^2) in a vacuum in differential and integral form are:

$$\nabla \times \mathbf{B} = \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \tag{5.48a}$$

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 \varepsilon_0 \int_S \frac{\partial \mathbf{E}}{\partial t} \cdot d\mathbf{s} \quad \text{Ampere} \tag{5.48b}$$

$$\nabla \cdot \mathbf{E} = 0 \tag{5.49a}$$

$$\int_S \mathbf{E} \cdot d\mathbf{s} = 0 \quad \text{Electrical Gauss} \tag{5.49b}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{5.50a}$$

$$\oint_S \mathbf{B} \cdot d\mathbf{s} = 0 \quad \text{Magnetic Gauss} \tag{5.50b}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{5.51a}$$

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s} \quad \text{Faraday} \tag{5.51b}$$

where $\mathbf{E}$ (in V/m) is the electrical field strength and $\mathbf{B}$ (in Tesla, T, or $kgA^{-1}s^{-2}$) the magnetic flux density, and μ_0 and ε_0 the permeability and electric permittivity of free space, respectively; $d\mathbf{l}$ is a line element, and $d\mathbf{s}$ a surface element. The permittivity of free space, ε_0, is 8.854 pF/m. The permeability of vacuum, μ_0, is 1.257 10^{-6} H m^{-1}. Vacuum is a linear, homogeneous, isotropic, and dispersionless medium. Except for a multiplicative scalar, electric and magnetic fields appear in these equations with remarkable symmetry, and this mathematical symmetry implies physical symmetry. We note here that the four Maxwell equations are not independent. In fact, the two divergence equations can be derived from the two curl equations.

The time-varying Maxwell's equations, in a medium and in the presence of charges and currents, in their differential and integrated forms are given as:

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{5.48c}$$

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \left(\mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}\right) \cdot d\mathbf{s} \quad \text{Ampere} \tag{5.48d}$$

$$\nabla \cdot \mathbf{D} = \rho_V \tag{5.49c}$$

$$\int_S \mathbf{D} \cdot d\mathbf{s} = Q \quad \text{Electrical Gauss} \tag{5.49d}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{5.50c}$$

$$\oint_S \mathbf{B} \cdot d\mathbf{s} = 0 \qquad \text{Magnetic Gauss} \qquad (5.50d)$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad (5.51c)$$

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s} \qquad \text{Faraday} \qquad (5.51d)$$

where **D** [in coulombs per square meter (C/m^2)] is the electric displacement, **H** (in A/m) is the magnetic field strength, **J** is current density (A/m^2), Q (C) is the total charge, and ρ_V is the volume density of free electric charges (C/m^3). Note that no constants such as μ_0, ε_0, μ, ε, c, χ … appear when the equations are written in this way. The divergence equations basically say that flux lines start and end on charges or poles:

1. Diverging D fields relate to charges: $\nabla \cdot \mathbf{D} = \rho_V$
2. Diverging magnetic flux does not exist: $\nabla \cdot \mathbf{B} = 0$

The curl equations state that changes in magnetic fluxes give rise to electrical fields and changes in currents give rise to H fields or changing E field results in changing H field, which results in changing E field.… (Figure 5.26):

3. Rotating E fields relate to changing currents: $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
4. Rotating H flux lines relate to changing D, and several currents could be added (e.g., magnetization): $\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}$

If we now consider solutions for time-harmonic electric and magnetic fields (e.g., sinusoidal) and replace the time derivative $\partial/\partial t$ by $-i\omega$ [because $\frac{\partial}{\partial t}(e^{-i\omega t}) = -i\omega e^{-i\omega t}$], we arrive at the following Maxwell expressions:

$$\nabla \times \mathbf{H} = \mathbf{J} - i\omega\mathbf{D} \qquad (5.48e)$$

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \mathbf{J} - i\omega\mathbf{D} \qquad \text{Ampere} \qquad (5.48f)$$

$$\nabla \cdot \mathbf{D} = \rho_V \qquad (5.49e)$$

$$\int_S \mathbf{D} \cdot d\mathbf{s} = Q \qquad \text{Electrical Gauss} \qquad (5.49f)$$

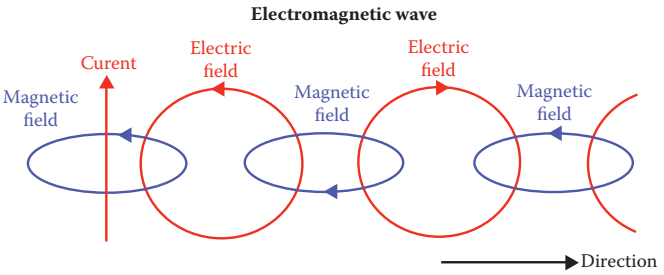


FIGURE 5.26 Changing *E* field results in changing *H* field, which results in changing *E* field.

$$\nabla \cdot \mathbf{B} = 0 \qquad (5.50e)$$

$$\oint_S \mathbf{B} \cdot d\mathbf{s} = 0 \qquad \text{Magnetic Gauss} \qquad (5.50f)$$

$$\nabla \times \mathbf{E} = +i\omega\mathbf{B} \qquad (5.51e)$$

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = +i\omega \int_S \mathbf{B} \cdot d\mathbf{s} \qquad \text{Faraday} \qquad (5.51f)$$

Boltzmann was so impressed with Maxwell's equation that he used a quotation from Goethe to describe them: "*War es ein Gott der diese Zeichen schrieb*?" (Is it a God who has written these signs?). These equations, together with the continuity equation (Equation 5.46), the Lorentz force* (Equation 3.64), and the constitutive equations for ε, μ, and σ, form the foundation of all macroscopic electromagnetic phenomena.

In the case of no time variation ($\partial/\partial t = 0$), the Maxwell equations describe electrostatics and magnetostatics as summarized in Table 5.1.

In the next section we will find that Maxwell's equations allow for solutions that have the form of propagating plane waves, and that the wavelength is affected by the amount of polarization **P** of the material, leading to a refractive index n.

Plane Wave Solutions of Maxwell's Equations

Plane Wave Propagation

Earlier we saw how electromagnetic (EM) plane waves are observed when moving far enough away from the source, just like dropping a rock in a pond and looking at waves a few hundred feet away from the

* An electric charge moving in the presence of an electric and a magnetic field experiences a vector force per unit charge that is equal to the sum of the electric field and the cross product of the velocity with the magnetic field.

Table 5.1 Electrostatics (Capacitor Equations) and Magnetostatics (Inductor Equations)

Magnetostatics	
$\nabla \cdot \mathbf{D} = \rho_V$	$\nabla \cdot \mathbf{B} = 0$
$\nabla \times \mathbf{E} = 0$	$\nabla \times \mathbf{H} = \mathbf{J}$
$\oint_S \mathbf{D} \cdot d\mathbf{s} = Q$	$\oint_S \mathbf{B} \cdot d\mathbf{s} = 0$
$\oint_C \mathbf{E} \cdot d\mathbf{l} = 0$	$\oint_C \mathbf{H} \cdot d\mathbf{l} = I$
Electrical capacitance (in Farad) $C = Q/V$	Electric inductance (in Henry) $L = \Phi/I$*

*Φ is magnetic flux, and I is current.

impact point so that the wavefronts are essentially flat. When close enough to the EM source the radiated field must be considered mathematically spherical rather than planar. The first situation is called *far field*, and the second is *near field* (Figures 5.4 and 5.5). In Chapter 2, we defined a plane wave,* $e^{i(\mathbf{kr}-\omega t)}$, as a wave that is constant in a plane perpendicular to $\mathbf{k}$ (in rad/m) and is periodic parallel to it. The wave number or wave vector $\mathbf{k}$ measures the number of wavelengths in a complete cycle; thus, it has a value of $2\pi/\lambda$, and $\omega = 2\pi\nu = c\mathbf{k}$ (free space) is the period of the wave. In the plane wave, equation $\mathbf{r}$ is the 3D position vector. A bit simpler: a plane wave is a continuous wave (CW) whose amplitude and phase are constant in the directions transverse to the propagation direction.

We will now embark on a discussion of the interaction of light with materials. In optical experiments one measures reflectivity, transmission coefficient, luminescence, or light scattering to learn a great deal about energy band structure, impurities, phonons, defects, etc. In particular, optical measurements are used to determine the complex dielectric or complex conductivity, which is directly related to band structure.

Wave Equations for a Medium without Charges or Currents

We now derive the wave equations for plane wave propagation in a simple linear isotropic, homogeneous, nonconducting medium characterized by ε and μ without any charges or currents present (source-free region, i.e., a region where ρ_V and $\mathbf{J}$ both are zero) from the Maxwell equations. Taking the curl of Equation 5.51c leads to the homogenous vector wave equations (Equations 5.53 and 5.54) for electromagnetic (EM) radiation in such a medium we obtain:

$$\nabla \times (\nabla \times \mathbf{E}) = -\frac{\partial(\nabla \times \mathbf{B})}{\partial t} \tag{5.52}$$

L.H.S: $\nabla(\nabla \cdot \mathbf{E}) - \nabla^2\mathbf{E} = -\nabla^2\mathbf{E}$
($\nabla \cdot \mathbf{E} = 0$; Equation 5.49a)

R.H.S: $-\mu\varepsilon\dfrac{\partial^2\mathbf{E}}{\partial^2 t}$ (Equation 5.48c with $\mathbf{J} = 0$ and $\mathbf{D} = \varepsilon\mathbf{E}$ and $\mathbf{B} = \mu\mathbf{H}$)

$$\text{or } \nabla^2\mathbf{E} = \mu\varepsilon\frac{\partial^2\mathbf{E}}{\partial^2 t} \tag{5.53}$$

and similarly

$$\nabla^2\mathbf{B} = \mu\varepsilon\frac{\partial^2\mathbf{B}}{\partial^2 t} \tag{5.54}$$

For the left-hand side (L.H.S.) of Equation 5.52 we used the "BAC-CAB" rule.† Equations 5.53 and 5.54 represent wave equations for electromagnetic (EM) waves in bulk materials. In free space, μ and ε in these equations are replaced by μ_0 and ε_0. The Laplacian ∇^2 operates on each component of $\mathbf{E}$ and $\mathbf{B}$, so that the two vector equations (Equations 5.53 and 5.54) represent six 1D, homogeneous scalar wave equations. One of these expressions, the E_x component, in Cartesian coordinates, is written out as:

$$\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} = \varepsilon\mu\frac{\partial^2 E_x}{\partial t^2} \tag{5.55}$$

Therefore, each component of the EM field (E_x, E_y, E_z, B_x, B_y, B_z) obeys a generic scalar differential wave equation of the type:

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2\Psi}{\partial t^2} \tag{5.56}$$

provided that the following important identity holds:

$$v = \frac{1}{\sqrt{\varepsilon\mu}} \tag{5.20}$$

* The complex exponential representation for periodic functions is convenient for adding waves, taking derivatives of wave functions, and so on. It is equivalent to a linear combination of a sine and cosine function because $e^{i\theta} = \cos\theta + i\sin\theta$ (Euler's formula).

† We are taking advantage in this derivation of the "BAC-CAB" rule: $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}(\mathbf{A} \cdot \mathbf{B})$, or in the current case: $\nabla \times (\nabla \times \mathbf{E}) = \nabla(\nabla \cdot \mathbf{E}) - (\nabla \cdot \nabla)\mathbf{E} = -\nabla^2\mathbf{E}$.

The velocity v, the propagation speed of the wave, is solely determined by the medium. For a wave traveling in free space we substitute the values for μ_0 and ε_0 in Equation 5.20 and calculate the speed for all EM waves as $v = c = 2.998\ 10^8$ m/s. At the time of Maxwell, the values for μ_0, ε_0, and c were uncertain enough that the correspondence between c and Equation 5.20 was not all that definitive. However, it was close enough that Maxwell wrote in 1864: "This velocity is so nearly that of light that it seems we have strong reason to conclude that light itself (including radiant heat and other radiations) is an electromagnetic disturbance in the form of waves propagated through the electromagnetic field according to electromagnetic laws." It is largely for this reason that this whole set of equations is named after Maxwell. When Heinrich Hertz (1857–1894) produced low-frequency radio waves (low-frequency light) and demonstrated that these waves were identical in all respects to light, save frequency, he confirmed Maxwell's hypothesis. It was also Hertz who, in 1887, discovered the photoelectric effect (see Chapter 3). Maxwell's equations show that electric and magnetic forces travel at a definite predicted speed—this is not instantaneous "action at a distance." James Clerk Maxwell's *Treatise on Electricity & Magnetism* (1873) is the last word on classical electricity and magnetism and ranks with Newton's work as one of the great accomplishments of physics. Newton's and Maxwell's equations describe all "classical physics." In Chapter 3 we introduced Bohr's correspondence principle, according to which the predictions of the quantum theory must also cover the predictions of the classical theories of physics. In other words, the above equations derived here from Maxwell's equations should also be a solution of Schrödinger's equation. We demonstrated that this is so in Chapter 3, "Solution of the Wave Equation for Free Particles."

Wave Equations for a Medium with Charges and Currents

To obtain the most general wave equations—in the presence of sources (charges, ρ) and a current ($\mathbf{J}$) in a medium—we first combine the constitutive relations with the general time-varying Maxwell's equations to obtain the following two equivalent expressions (using the BAC-CAB rule again):

$$\nabla \times (\nabla \times \mathbf{E}) = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2\mathbf{E}$$

$$= -\mu\frac{\partial \mathbf{J}}{\partial t} - \mu\frac{\partial^2 \mathbf{D}}{\partial t^2} = -\mu\sigma\frac{\partial \mathbf{E}}{\partial t} - \mu\varepsilon\frac{\partial^2 \mathbf{E}}{\partial t^2}$$

$$\nabla^2\mathbf{E} - \mu\sigma\frac{\partial \mathbf{E}}{\partial t} - \mu\varepsilon\frac{\partial^2 \mathbf{E}}{\partial t^2} = \nabla\left(\frac{\rho_v}{\varepsilon}\right) \qquad (5.57)$$

and similarly:

$$\nabla^2\mathbf{H} - \mu\sigma\frac{\partial \mathbf{H}}{\partial t} - \mu\varepsilon\frac{\partial^2 \mathbf{H}}{\partial t^2} = 0 \qquad (5.58)$$

In solving these equations, one often invokes the electric scalar and the magnetic vector potential. The magnetic vector potential $\mathbf{A}$ (in Weber per meter) is related to $\mathbf{B}$ as:

$$\mathbf{B} = \nabla \times \mathbf{A} \qquad (5.59)$$

If one can find $\mathbf{A}$ for a given current distribution, then $\mathbf{B}$ can be obtained by taking the differential (curl) from $\mathbf{A}$. This is similar to obtaining the electrical field $\mathbf{E}$ from the scalar electrical potential V ($\mathbf{E} = -\nabla V$). Substituting Equation 5.59 into the differential Faraday expression (Equation 5.51a) gives:

$$\nabla \times \mathbf{E} = -\frac{\partial}{\partial t}(\nabla \times \mathbf{A}) \text{ or}$$

$$\nabla \times \left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}\right) = 0 \qquad (5.60)$$

The electric field intensity stems from both charge accumulations (through the $-\nabla V$ term) and from the time-varying magnetic field (through the $-\frac{\partial \mathbf{A}}{\partial t}$ term), or:

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} - \nabla V \qquad (5.61)$$

For the static case (with $\frac{\partial \mathbf{A}}{\partial t} = 0$), this reduces back to the familiar $\mathbf{E} = -\nabla V$, so that we can obtain $\mathbf{E}$ from V alone and $\mathbf{B}$ from Equation 5.59. With varying fields, $\mathbf{E}$ depends on both V and $\mathbf{A}$. The inhomogeneous

vector potential wave equation to be solved then becomes:

$$-\nabla^2\mathbf{A} + \mu\sigma\frac{\partial \mathbf{A}}{\partial t} + \mu\varepsilon\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\sigma\nabla V \qquad (5.62)$$

Using the constitutive equations and substituting Equations 5.59 and 5.61 into 5.48e, we obtain:

$$\nabla \times \nabla \times \mathbf{A} = \mu\mathbf{J} + \mu\varepsilon\frac{\partial}{\partial t}\left(-\frac{\partial \mathbf{A}}{\partial t} - \nabla V\right) \qquad (5.63)$$

or also:

$$\nabla^2\mathbf{A} - \mu\varepsilon\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\mathbf{J} + \nabla\left(\nabla\cdot\mathbf{A} + \mu\varepsilon\frac{\partial V}{\partial t}\right) \qquad (5.64)$$

By choosing the divergence of **A** such that:

$$\nabla\cdot\mathbf{A} + \mu\varepsilon\frac{\partial V}{\partial t} = 0 \qquad (5.65)$$

which is the so-called *Lorentz condition for potentials* (for static fields this condition becomes $\nabla\cdot\mathbf{A} = 0$). The Lorentz condition enables us to simplify Equation 5.64 to:

$$\nabla^2\mathbf{A} - \mu\varepsilon\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\mathbf{J} \qquad (5.66)$$

This is the nonhomogeneous wave equation for vector potential **A**; solutions are waves traveling with a velocity v equal to $\frac{1}{\sqrt{\mu\varepsilon}}$. The corresponding wave equation for the scalar potential V is obtained by substituting Equation 5.61 in Equation 5.49c, or:

$$-\nabla\cdot\varepsilon\left(\nabla V + \frac{\partial \mathbf{A}}{\partial t}\right) = \rho_v \qquad (5.67)$$

With ε constant this can be written out as:

$$\nabla^2 V + \frac{\partial}{\partial t}(\nabla\cdot\mathbf{A}) = -\frac{\rho_v}{\varepsilon} \qquad (5.68)$$

With the Lorentz condition (Equation 5.65) applied, this leads to:

$$\nabla^2 V - \mu\varepsilon\frac{\partial^2 V}{\partial t^2} = -\frac{\rho_v}{\varepsilon} \qquad (5.69)$$

This result represents nonhomogeneous wave equation for the scalar potential V. Both wave equations (Equations 5.66 and 5.69) reduce to the Poisson's equations in static cases.

The above derived wave equations are the basis for modeling electromagnetic fields in microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS).

For more details on deriving and working with Maxwell's equations, consult David K. Cheng's *Field and Wave Electromagnetics*[1] and Minoru Taya's *Electronic Composites*.[2]

Plane Wave Solutions for Media without Charges or Currents

The Overall Solutions Are Plane Waves: $\mathbf{E(r,t)} = \mathbf{E}_m e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$ *and* $\mathbf{B(r,t)} = \mathbf{B}_m e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$ The solutions of the wave equations for plane wave propagation in a simple linear isotropic, homogeneous, nonconducting medium (ρ_V and **J** both are zero) characterized by ε and μ (Equations 5.53 and 5.54) in one dimension (say with the E field propagating along the z-axis, the B field along y, and the propagation vector in the x-direction) are:

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_m e^{i(\mathbf{k}_x\cdot\mathbf{x}-\omega t)} \text{ and } \mathbf{B} = \mathbf{B}_m e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \qquad (5.70)$$

Rewriting these expressions in sines and cosines instead of the complex variable form ($e^x = \cos x + i \sin x$), one obtains as the simplest form of a light wave:

$$\mathbf{E} = \mathbf{E}_m\sin(\mathbf{k}_x x - \omega t) \text{ and } \mathbf{B} = \mathbf{B}_m\sin(\mathbf{k}_x x - \omega t) \qquad (5.71)$$

Generalizing for 3D, Equation 5.70 converts to:

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_m e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \text{ and } \mathbf{B}(\mathbf{r},t) = \mathbf{B}_m e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \qquad (5.72)$$

with **r** the 3D position vector and **k** a complex propagation constant, with the real part of it corresponding to the wave vector and the imaginary part corresponding to the absorption (analogous to our treatment of x-rays in Chapter 2). The complex propagation constant **k** is obtained by substitution of this solution in the wave equations (see Equations 5.53 and 5.54).

It should be noted that plane waves are not the only solutions to Maxwell's equations; the differential

wave equation allows many solutions, including spherical waves.

Constraints to the Plane Wave Solutions: $\mathbf{E} \perp \mathbf{k}$ *and* $\mathbf{B} \perp \mathbf{k}$ *(I)* Let us assess what we have so far: the source-free Maxwell equations (Equations 5.48a–5.51a) were manipulated in such a way as to obtain the vector wave Equations 5.53 and 5.54, and solutions for these equations are given by $\mathbf{E}(\mathbf{r},t) = \mathbf{E}_m e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$ and $\mathbf{B}(\mathbf{r},t) = \mathbf{B}_m e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$ (Equation 5.72). But we are not done yet: in addition to satisfying the vector wave equation, we must also verify these solutions satisfy the other Maxwell equations. We can write down lots of functions that satisfy the wave equations but that are not valid electromagnetic fields for the given boundary conditions (ρ_V and $\mathbf{J}$ both are zero). We first plug these proposed solutions into the divergence Equations 5.49a and 5.50a to find the following additional constraints:

$$\nabla \cdot \mathbf{E} = 0 \,(\text{Equation 5.49a}) \rightarrow \mathbf{k} \cdot \mathbf{E} = 0 \qquad (5.73)$$

$$\nabla \cdot \mathbf{B} = 0 \,(\text{Equation 5.50a}) \rightarrow \mathbf{k} \cdot \mathbf{B} = 0 \qquad (5.74)$$

In other words, the divergence equations require that the electric and magnetic fields are orthogonal to the propagation direction: $\mathbf{E} \perp \mathbf{k}$ and $\mathbf{B} \perp \mathbf{k}$. The $\mathbf{E}$ and $\mathbf{B}$ fields have no component of vibration in the direction of propagation: there are no longitudinal components, or $\mathbf{E}$ and $\mathbf{B}$ are exclusively transverse in nature.

Constraints to the Plane Wave Solutions: $\mathbf{E} \perp \mathbf{B} \perp \mathbf{k}$ *(II)* Let us now substitute these same solutions in the curl equations (Equations 5.48a and 5.51a) and establish how this adds the following constraints:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \,(\text{Equation 5.51a}) \longrightarrow \mathbf{k} \times \mathbf{E} = \omega \mathbf{B} \qquad (5.75)$$

$$\nabla \times \mathbf{B} = \mu\varepsilon \frac{\partial \mathbf{E}}{\partial t} \,(\text{Equation 5.48a}) \longrightarrow \mathbf{k} \times \mathbf{B} = \frac{\omega}{c^2}\mathbf{E} \qquad (5.76)$$

Consider $\nabla \times \mathbf{B} = \mu\varepsilon \frac{\partial \mathbf{E}}{\partial t}$ one of the curl equations: on the right-hand side, we have the time derivative of $\mathbf{E}$, pointing in the same direction as $\mathbf{E}$ because the time derivative does not change the vector's direction. On the left, we have $\nabla \times \mathbf{B}$; this is perpendicular to $\mathbf{B}$ because the curl of any vector is orthogonal to that vector. Therefore, the first thing the curl equations tell us is that radiation's electric and magnetic fields are orthogonal to each other. Second, the curl equations also require that $\mathbf{E} \times \mathbf{B}$ is parallel to the propagation vector $\mathbf{k}$. In other words, the symmetry of Maxwell's equations results in $\mathbf{E} \perp \mathbf{B} \perp \mathbf{k}$ as illustrated in Figure 5.1.

In electromagnetic waves, the $\mathbf{E}$ and $\mathbf{B}$ fields are obviously intimately linked. To test this further, let us try out the solutions, $\mathbf{E}_x = \mathbf{E}_m e^{i(\mathbf{k}\cdot x-\omega t)}$ and $\mathbf{B}_y = \mathbf{B}_m e^{i(\mathbf{k}\cdot x-\omega t)}$ [$\mathbf{E}$ along the z-axis, $\mathbf{B}$ along the y-axis, and propagation direction $\mathbf{k}$ along the x-axis $\left(\frac{\partial E_x}{\partial x} = 0\right)$] in Equation 5.51a, which leads to:

$$\begin{aligned} \nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \text{ or} \\ \frac{\partial E_z}{\partial x} &= -\frac{\partial B_y}{\partial t} \text{ and thus } \mathbf{E}_m(ik)e^{i(\mathbf{k}\cdot x-\omega t)} \\ &= -\mathbf{B}_m(-i\omega)e^{i(\mathbf{k}\cdot x-\omega t)} \\ \text{or } \frac{E_m}{B_m} &= \frac{\omega}{k} \end{aligned} \qquad (5.77)$$

For planar waves, $\mathbf{k} = 2\pi/\lambda$ and $\omega = 2\pi\nu = \nu\mathbf{k}$ with v being the speed of the wave, and because E and B are in phase we conclude that E and B are related at any point in space by the velocity of the propagating wave:

$$\frac{E}{B} = \frac{\omega}{k} = \nu \qquad (5.78)$$

This relates the wave vector $\mathbf{k}$, relating to phase variations in space, to ω, the angular frequency, relating to phase variations in time. In the case of a wave in free space, $\nu = c = 2.998 \times 10^8$ m/s. Thus, boundary conditions for planar waves propagating in a medium without charges lead to Equation 5.78 expressing an important constraint for the proposed solution in a dispersion relation, i.e., a relation between the energy of the system and its corresponding momentum $\mathbf{k}$ (ω). In the case of free space, this is a linear function ($\omega = \mathbf{k}c$) with a slope c, the so-called *light line* (see Figure 5.27).

Summarizing, we conclude that in a source-free region for time-harmonic signals Maxwell's equations are reduced to two coupled curl equations, and these two curl equations combine to produce the wave equation for **E** or **B**. The eigenfunctions for these wave equations are plane waves described by propagation direction vectors called **k**-*vectors* that have a length of $2\pi/\lambda$ and result in wave velocity v (the **k**-vector is reciprocal to the space variation of the wavelength). The two curl equations have zero divergence, and this requires the vectors **E** and **B** to be perpendicular to the direction of propagation, whereas the **E** and **B** vectors must also be perpendicular to each other.

In the section below on electromagnetic waves at boundaries, we will learn that to solve electromagnetic problems involving adjacent regions of different constitutive parameters, it is necessary to establish the boundary conditions that the field vectors **E**, **D**, **B**, and **H** must satisfy at the interfaces between the different optical materials. These boundary conditions, we will see, are established by applying the integral form of Maxwell's equations to a small region at the interface of the contacting media.

Plane Wave Solutions for Media with Charges or Currents

We now substitute the time-harmonic electric field wave $\mathbf{E}(\mathbf{r},t) = \mathbf{E}_m e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} = \mathbf{E}_m e^{i\mathbf{k}\cdot\mathbf{r}} e^{-i\omega\cdot t} = \mathbf{E}_r(\mathbf{r}) e^{-i\omega\cdot t}$ in the most general wave equation for media with charges and currents $\nabla^2\mathbf{E} - \mu\sigma\dfrac{\partial \mathbf{E}}{\partial t} - \mu\varepsilon\dfrac{\partial^2 \mathbf{E}}{\partial t^2}$ (Equation 5.57). We will assume here nonmagnetic materials with $\mu = \mu_0$ (remember that the magnetic permeability $\mu = \mu_0 \cdot \mu_r$, where μ_r is the dimensionless relative permeability, which is 1 in this case) to obtain:

$$\nabla^2\mathbf{E}(\mathbf{r}) = -i\omega\sigma\mathbf{E}(\mathbf{r}) - \varepsilon\omega^2\mathbf{E}(\mathbf{r})$$

$$\nabla^2\mathbf{E}(\mathbf{r}) = -\frac{\omega^2}{v^2}\varepsilon(\omega)\mathbf{E}(\mathbf{r}) \tag{5.79}$$

with an effective dielectric contstant:

$$\varepsilon(\omega) = 1 + \frac{i\sigma}{\omega}$$

Substituting $\mathbf{E}_r(\mathbf{r}) = \mathbf{E}_m e^{i\mathbf{k}\cdot\mathbf{r}}$ in Equation 5.79 leads to the following constraint expressed in a new dispersion equation:

$$\mathbf{k}^2 = \frac{\omega^2}{v^2}\varepsilon(\omega) \tag{5.80}$$

This is a relation between the wave vector **k**, related to phase variations in space, and ω, the angular frequency, related to phase variations in time. Note that a higher temporal frequency yields a higher spatial frequency, and a large permittivity will result in a large **k** (short wavelength). The "slow-down" factor of an electromagnetic wave in a material relative to free space is by definition the refractive index $n = \dfrac{c}{v}$ (Equation 5.21). With $n = \dfrac{c}{v}$ we can rewrite Equation 5.80 in terms of the refractive index n and the velocity of light in free space c as:

$$\mathbf{k}^2 = \frac{n^2}{c^2}\omega^2 \text{ or } |\mathbf{k}| = \frac{n}{c}\omega \tag{5.81}$$

which in the case of free space with n = 1 simplifies back to $\omega = \mathbf{k}c$. Thus, the wave equation again imposes a constraint on the length of the **k**-vector. Below we will analyze this dispersion relation in terms of the phase velocity of a wave ($v_p = \omega/\mathbf{k} = \mathbf{E}/\mathbf{p}$) and the group velocity ($v_g = d\omega/d\mathbf{k} = d\mathbf{E}/d\mathbf{p}$) of that wave.

Characteristic Vacuum and Medium Impedance—η_0 and η

The vacuum impedance η_0 is a universal constant relating the magnitudes of the electric and magnetic fields of electromagnetic radiation traveling through free space. From Equation 5.75:

$$\eta_0 = \frac{|\mathbf{E}|}{|\mathbf{H}|} = \frac{\omega\mu_0}{|\mathbf{k}|} = \mu_0 c = \frac{|\mathbf{k}|}{\omega\varepsilon} = \frac{1}{\varepsilon_0 c} = \sqrt{\frac{\mu_0}{\varepsilon_0}} \tag{5.82}$$

which equals 376.73 ohms. More general, for a uniform plane wave the ratio of the magnitudes of **E** and **H** is the intrinsic impedance of the medium and is designated η. Again from Equation 5.75 it follows:

$$\eta = \frac{|\mathbf{E}|}{|\mathbf{H}|} = \frac{\omega\mu}{|\mathbf{k}|} = \frac{|\mathbf{k}|}{\omega\varepsilon} = \sqrt{\frac{\mu}{\varepsilon}} \tag{5.83}$$

Like the velocity of propagation, the intrinsic impedance is independent of the source and only determined by the properties of the medium.

Dispersion Relations in Vacuum and in Media

In Vacuum

In physics, the dispersion relation is the relation between the energy of a system and its corresponding momentum. Here we analyze the dispersion relation $\omega(\mathbf{k})$ from Equation 5.81 in terms of the phase velocity of a wave ($v_p = \omega/\mathbf{k} = \mathbf{E}/\mathbf{p}$) and the group velocity ($v_g = d\omega/d\mathbf{k} = dE/d\mathbf{p}$).

Remember that the phase velocity of a wave ($v_p = \omega/\mathbf{k} = \mathbf{E}/\mathbf{p}$) is not necessarily the same as the group velocity ($v_g = d\omega/d\mathbf{k} = dE/d\mathbf{p}$) of that wave. Two velocities must be defined when dealing with multiple waves, with the phase velocity corresponding to the speed of the high-frequency components and the group velocity to the speed of a localized envelope. Also, the phase velocity of the wave is the rate at which the phase of that wave propagates in space, or it is the velocity at which the phase of any one frequency component of the wave propagates and the group velocity we saw earlier also corresponds to the light "particle" speed (Chapter 3).

Because the group velocity v_g is given as $d\omega/d\mathbf{k}$, we calculate:

$$v_g = \frac{d\omega}{d\mathbf{k}} = \frac{d}{d\mathbf{k}}(\mathbf{k}v_p) = v_p + k\frac{dv_p}{d\mathbf{k}} \qquad (5.84)$$

In a nondispersive medium such as a vacuum $\frac{dv_p}{d\mathbf{k}} = 0$, the group and phase velocity is the same ($v_p = v_g$), or:

$$v_p = \frac{\omega}{\mathbf{k}} = v_g = \frac{d\omega}{d\mathbf{k}} = c \qquad (5.85)$$

This is of course the same result we obtained earlier as a constraint when substituting a planar wave $[\mathbf{E}(\mathbf{r},t) = \mathbf{E}_m e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}]$ in the wave equation for free space $\nabla^2\mathbf{E} = \mu\varepsilon\frac{\partial^2\mathbf{E}}{\partial^2 t}$ (Equation 5.53). Thus, the dispersion relation—the relation between the energy of the system and its corresponding momentum $\mathbf{k}(\omega)$—is a linear function of ω with a slope $c = v_g = v_p$ in the case of vacuum as shown in Figure 5.27.

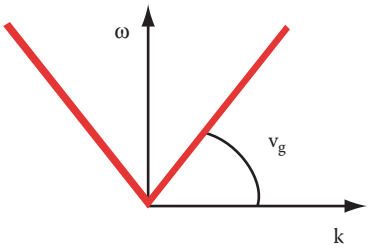


FIGURE 5.27 The dispersion relation—the relation between the energy of the system and its corresponding momentum $\mathbf{k}(\omega)$—is a linear function of ω in the case of vacuum with a slope $c = v_g = v_p$.

In a Medium

Dispersion occurs when the phase velocity v_p depends on $\mathbf{k}$ (or λ), i.e., the group velocity does not equal the phase velocity (see Equation 5.85):

$$\frac{dv_p}{d\mathbf{k}} \neq 0 \text{ or } v_g \neq v_p \qquad (5.86)$$

A wave packet or group velocity with $v_g = d\omega/d\mathbf{k}$ less than the phase velocity $v_p = \omega/\mathbf{k}$ is illustrated in Figure 5.28 for a case where $v_g = v_p/2$.

We can rewrite the group velocity ($v_g = d\omega/d\mathbf{k}$) of light in a medium also in terms of refractive index n and the velocity of light c. For the group velocity we then derive for the most general case:

$$v_g = \frac{d\omega}{d\mathbf{k}} = \frac{c}{n} = \frac{c}{\sqrt{\varepsilon\mu}} \qquad (5.87)$$

where we used the fact that $n(\omega) = \frac{c}{v_g}$ and $n(\omega)^2 = \varepsilon\mu$ (Equations 5.21 and 5.22), and the dispersion relation then becomes:

$$\mathbf{k}(\omega) = \frac{\omega}{v_g} = \frac{n(\omega)\omega}{c} = \frac{\omega}{c}\sqrt{\varepsilon\mu} \qquad (5.88)$$

We will use this important equation to construct a μ-ε *quadrant* to classify all materials according to their magnetic permeability μ and their dielectric constant (or permittivity) ε.

From Equation 5.88 we can also calculate the group velocity v_g as a function of the refractive index:

$$v_g = \frac{c}{n + \frac{\omega dn}{d\omega}} = \frac{c/n}{1 + \frac{\omega dn}{n d\omega}} \qquad (5.89)$$

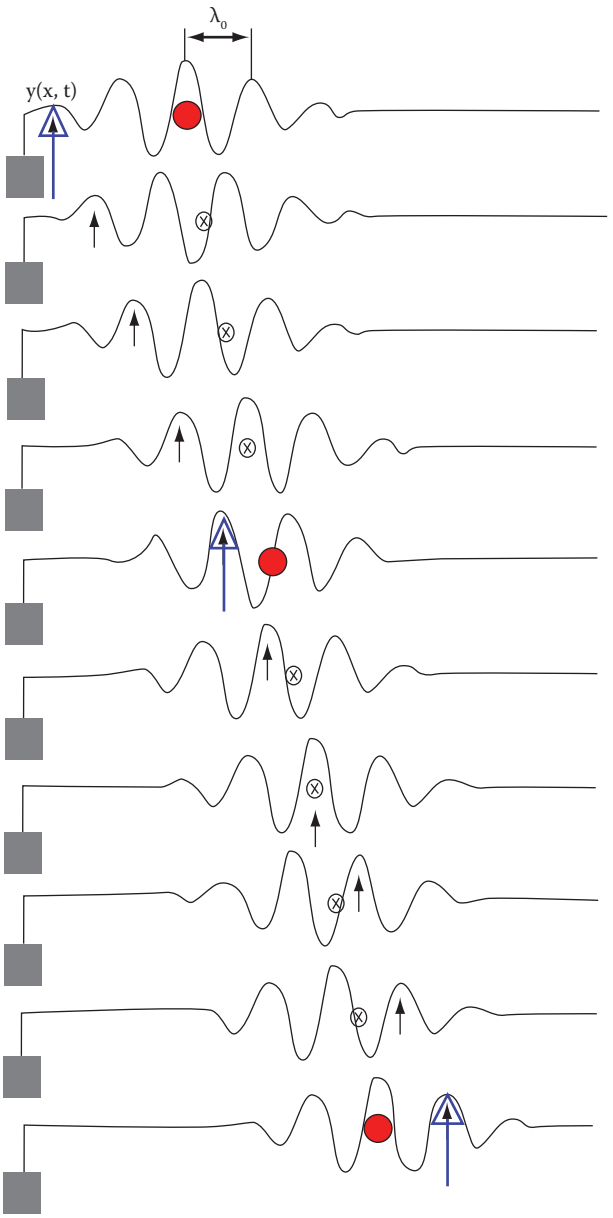


FIGURE 5.28 A wave packet or group velocity with $v_g = d\omega/d\mathbf{k}$ less than the phase velocity $v_p = \omega/\mathbf{k}$ with $v_g = v_p/2$. Phase velocity v_p is marked by blue arrows, and the group velocity v_g is marked with red dots.

and finally:

$$v_g = \frac{v_p}{1 + \frac{\omega dn}{n d\omega}} \tag{5.90}$$

The group velocity v_g equals the phase velocity when $dn/d\omega = 0$, such as in vacuum. Otherwise, because n usually increases with ω, $dn/d\omega > 0$ and v_g is less than the phase velocity v_p. This is illustrated in Figure 5.29, where we plot the refractive index as a function of ω. We see that the refractive index is positive for most frequencies ($v_g < v_p$). The exception is in regions of anomalous dispersion, regions that are not only dispersive but also absorbing, that is, near a resonance. To explain these anomalous regions we will introduce the Lorentz model further below.

We can summarize the frequency dependence of the refractive index n shown here as follows:

- Low frequencies: All transitions contribute their low-frequency response ⇒ high n.
- Between resonances: n slowly increases = "normal dispersion."
- Close to resonances: n rapidly decreases = "anomalous dispersion."
- Above resonance, contribution vanishes ⇒ $n(\omega)$ mostly because of higher resonances.
- At high frequencies (~x-ray wavelengths) n approaches 1.

We more often think of the refractive index in terms of wavelength, so let us write the group velocity also in terms of the vacuum wavelength λ_0. Using the chain rule we can write:

$$\frac{dn}{d\omega} = \frac{dn}{d\lambda_0}\frac{d\lambda_0}{d\omega} \tag{5.91}$$

Now $\lambda_0 = 2\pi c/\omega$ and:

$$\frac{d\lambda_0}{d\omega} = \frac{-2\pi c}{\omega^2} = \frac{-2\pi c}{\left(\frac{2\pi c}{\lambda_0}\right)^2} = \frac{-\lambda_0^2}{2\pi c} \tag{5.92}$$

Substituting this in Equation 5.89, we obtain:

$$v_g = \frac{c}{n + \frac{\omega dn}{d\omega}} = \frac{c/n}{1 - \frac{\lambda_0 dn}{n d\lambda_0}} \tag{5.93}$$

Right- and Left-Handed Materials (Metamaterials)

From Maxwell's equations we derived that for planar waves with a positive μ and ε, the vectors **E** and **H** and **k** do constitute a right set of vectors and thus one defines right-handed materials (RHMs) as illustrated in Figure 5.30a. The direction of $\mathbf{E} \times \mathbf{H}$ is called the *beam* or *ray direction*.

In vacuum and for most materials the "right-hand rule" relates **E**, **H**, and **k** because normally

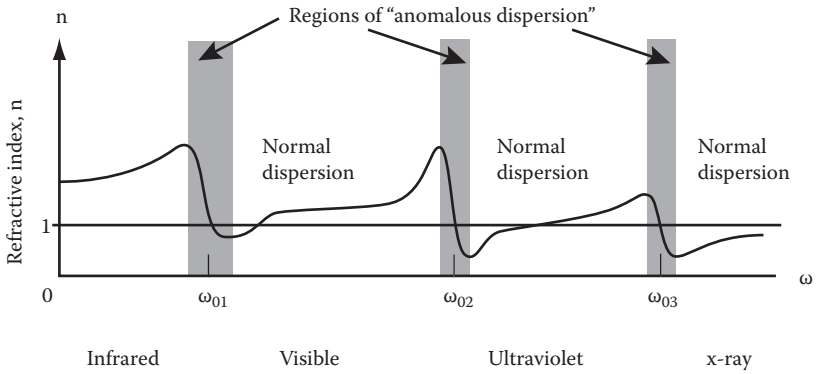


FIGURE 5.29 Refractive index n as a function of frequency ω: $n(\omega)$. The refractive index is positive, so $v_g < v_p$ for most frequencies! The exception is in regions of anomalous dispersion, absorbing regions near a resonance.

$\mu > 0$ and $\varepsilon > 0$. But in 1968, Victor G. Veselago postulated that if the permittivity ε and permeability μ of materials are negative simultaneously, there is no violation of any fundamental physical rules, and the materials will have special optical properties.[3] For a material with μ and ε negative, the vectors **E** and **H** and **k** form a left set of vectors, and one refers to them as left-handed materials (LHMs) (Figure 5.30b). For left-handed materials $n(\omega)^2 = \varepsilon\mu$ (Equation 5.22) is modified to:

$$\begin{aligned} n(\omega) &= \pm\sqrt{\varepsilon\mu} \text{ which for } \varepsilon < 0 \text{ and } \mu < 0 \\ n(\omega) &= -\sqrt{\varepsilon\mu} \end{aligned} \qquad (5.94)$$

So we are forced by Maxwell's equations to choose a negative square root for the refractive index. Because no naturally occurring material or compound has ever been demonstrated with both ε and μ negative, Veselago wondered whether this apparent asymmetry in material properties was just happenstance or perhaps had a more fundamental origin. From 1996 onward, strong theoretical and experimental evidence started emerging that left-handed materials, now called *metamaterials*, could be fabricated (at least in the RF range, not yet in the optical range). In Figure 5.17 we show a picture of a drinking straw in a glass with a negative refractive index water n = –1.3 and observe an apparent bending of the straw in the opposite direction of what we are used to in the case of "normal" water. We will detail metamaterials further below. Whereas in conventional materials properties derive from their constituent atoms, in metamaterials properties derive from their constituent units. These units can be engineered as we please.

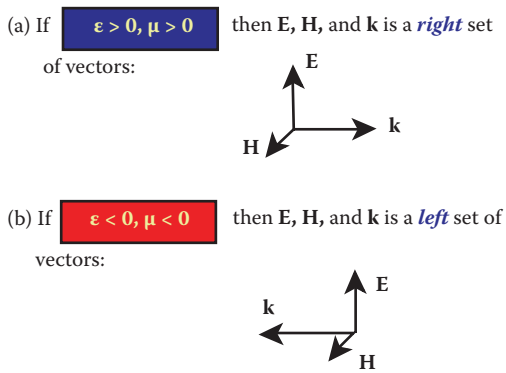


FIGURE 5.30 (a) Right-handed (RHM) and (b) left-handed materials (LHM).

Birefringence—Double Refraction

In the daily world we are most familiar with EM waves or light that propagates according to the right-hand rule (**k** is in the **E** × **H** direction) as illustrated in Figure 5.30a. Materials in which this is the case are isotropic media, and the dielectric constant ε in the constitutive relation (**D** = ε **E**; Equation 3.38) in this case is a scalar. Glass is an example of an isotropic material because it does not matter how light passes through it. The characteristics of the glass are such so that it has no different effects on the light even though the direction of the beam varies.

In nature one also has materials in which the direction of **k** is not exactly along the **E** × **H** direction; the angle between the two is then smaller than 90°. In mathematical terms, this means $\mathbf{k} \cdot (\mathbf{E} \times \mathbf{H}) > 0$. These are optically anisotropic media, such as birefringent calcite crystals ($CaCO_3$), quartz, ice, sugar,

and so on. Birefringence, also called *double refraction*, was first observed and explained in the 17th century after sailors visiting Iceland brought back to Europe transparent calcite crystals (Iceland spar) that showed double images of objects that were viewed through them (Figure 5.31). The effect was first explained by the Dutch physicist Christiaan Huygens (1629–1695), who called the effect *double refraction* that involved an ordinary and an extraordinary wave. The extraordinary ray is so named because it does not exhibit ordinary Snellius law (Equation 5.19) behavior on refraction at the crystal surface. As light travels through an anisotropic material, such as calcite, the electromagnetic waves become split into two principal vibrations, which are oriented mutually perpendicular to each other and perpendicular to the direction that the waves propagate. The wave whose electric vector vibrates along the major axis is termed the ordinary or slow wave because the refractive index for this wave is greater than the refractive index for the other wave. The wave vibrating perpendicular to the slow wave is termed the *fast* or *extraordinary wave*.

The dielectric tensor in the constitutive Equation 3.38 ($\mathbf{D} = \varepsilon\mathbf{E}$) in the case of an optically anisotropic material with a single optical axis (calcite, quartz, or any tetragonal, hexagonal, or rhombohedral crystal) is a tensor of rank two given by:

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} & 0 & 0 \\ 0 & \varepsilon_{xx} & 0 \\ 0 & 0 & \varepsilon_{zz} \end{bmatrix} \tag{5.95}$$

In a crystal like this there exists a single optical axis, the ordinary optical axis, where light propagation is isotropic. In all other directions, two orthogonally polarized waves with different velocities do propagate; these are called the *extraordinary rays*.

Birefringence is widely used in optical devices, such as liquid crystal displays, light modulators, color filters, wave plates, optical axis gratings, and so on. In Volume III, Chapter 6 on metrology we encounter the phenomenon in both ellipsometry (measures the polarization of light), and in Volume II, Chapter 10 (Figure 10.8), we compare birefringence of polymer MEMS structures made with different replication techniques to evaluate their stress levels.

Birefringence should not be confused with dichroism, which itself has two related but distinctly different meanings in optics. A dichroic material is either one that causes visible light to be split up into distinct beams of different colors, or one in which light rays having different polarizations are absorbed by different amounts. Light that passes through an isotropic crystal is absorbed in the same way in all directions, but in an anisotropic crystal, the light is absorbed differently, depending on its direction of travel through the crystal. Isotropic colored crystals show no color change as they are rotated in plane polarized light, whereas anisotropic crystals show a distinct change in color when they are viewed in different directions. An anisotropic crystal may produce a number of different colors, known as *pleochroism*. When only two colors are observed, it is called *dichroism*. Such devices include mirrors and filters, usually treated with optical coatings, which are designed to reflect light over a certain range of wavelengths and transmit light that is outside that range. An example is the dichroic prism, used in some camcorders, which uses several

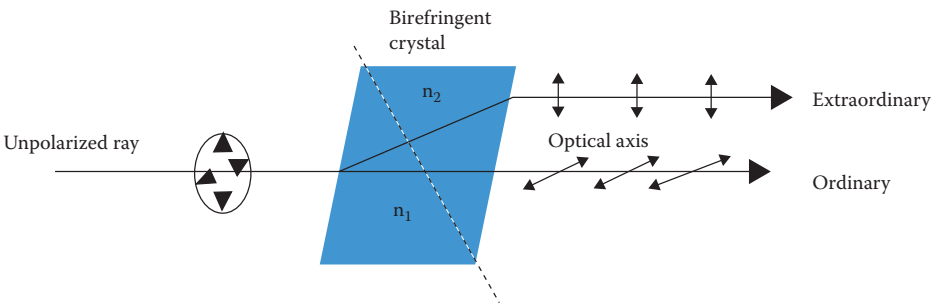


FIGURE 5.31 As light travels through an anisotropic material, such as calcite, the electromagnetic waves become split into two principal vibrations. In this double diffraction mode, an ordinary and an extraordinary wave are involved that are oriented mutually perpendicular to each other and perpendicular to the direction that the waves propagate.

coatings to split light into red, green, and blue components for recording on separate charge-coupled device (CCD) arrays. This kind of dichroic device does not usually depend on the polarization of the light. The second meaning of dichroic refers to a material in which light in different polarization states traveling through it experiences a varying absorption. The term came about because of early observations of the effect in crystals such as tourmaline. In these crystals, the strength of the dichroic effect varies strongly with the wavelength of the light, making them appear to have different colors when viewed with light having differing polarizations. This is more generally referred to as *pleochroism*, and the technique can be used in mineralogy to identify minerals. Dichroic crystals were researched by Edwin Land in 1930, who invented a polarizing material that would be called *Polaroid* and result in a huge benefit to the camera industry. The material he invented was herapathite (iodoquinine sulfate), in which the effect is not strongly dependent on wavelength, and so the term dichroic is something of a misnomer but still used.

Energy and Momentum and Electromagnetic Waves

Energy in EM—The Poynting Vector

Electromagnetic waves transfer energy from one place to another. The Poynting vector **S** describes that energy flux ($Jm^{-2}s^{-1}$). It is named after its "inventor," the English physicist John Henry Poynting (1852–1914). The vector **S** points in the direction of energy flow, and its magnitude is the power per unit area crossing a surface normal to it. The total energy, U, in an EM wave is the sum of the electrical and magnetic contributions:

$$U = \frac{1}{2}\varepsilon_0 E^2 + \frac{1}{2}\frac{B^2}{\mu_0} \tag{5.96}$$

From Equation 5.78, $E/B = c$, and because $(e_0\mu_0)^{-1/2} = 1/c$, we can rewrite Equation 5.96 as:

$$\text{Total energy} = \varepsilon_0 E^2 \tag{5.97}$$

The **E** field and the **B** field each contribute half of the total energy in an electromagnetic wave. As we just wrote out the total energy in terms of the **E** field, we may also write it out in terms of magnetic field **B** to obtain:

$$U = \frac{B^2}{\mu_0} \tag{5.98}$$

Writing the total energy out in terms of both **E** and **B**, we derive:

$$U = \varepsilon_0 E^2 = \varepsilon_0 EcB = \frac{\varepsilon_0 EB}{\sqrt{\varepsilon_0\mu_0}} = \sqrt{\frac{\varepsilon_0}{\mu_0}}EB \tag{5.99}$$

Now let us get back to the Poynting vector **S**: the energy the wave transports per unit time per unit area is given by:

$$S = \frac{1}{A}\frac{dU}{dt} = \frac{EB}{\mu_0} \tag{5.100}$$

The direction of **S** is along **k** and perpendicular to **E** and **B** for right-handed materials; for left-handed materials, **k** is still perpendicular to **E** and **H** (we can use **B** or **H** because $B = \mu H$) but opposite the Poynting vector as shown in Figure 5.32. Mathematically more correct, the Poynting vector is usually written out as:

$$S = \frac{1}{\mu_0}(E \times B) \tag{5.101}$$

Because **S**, **E**, and **H** always form a right-handed set, two types of substances can be identified: right-handed substances with **S** parallel to **k**, and left-handed substances with **S** antiparallel to **k** (see also Figure 5.30). Because **k** denotes the direction of phase velocity ($v_p = \omega/k = E/p$) and **S** denotes the direction of group velocity ($v_g = d\omega/dk = dE/dp$), in right-handed materials phase velocity (along **k**)

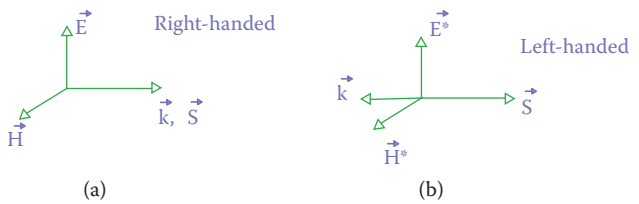


FIGURE 5.32 The phase velocity (along **k**) and group velocity (along the Poynting **E** × **H** vector) are in the same direction (a) but in left-handed materials. The group and phase velocities are in opposite directions (b) (see also Figure 5.30).

and group velocity (along the Poynting $\mathbf{E} \times \mathbf{H}$ vector) are in the same direction. In left-handed materials, group and phase velocity are in opposite directions. The left-handed substances have negative group velocity that occurs in anisotropic systems or when there is spectral dispersion. We delve into these matters deeper when we discuss metamaterials in more detail further below.

Momentum of EM: Radiation Pressure

Because magnetic waves carry energy, see above, they also deliver punch or momentum to a surface they are reflected off or absorbed in (think about a sun sail; Figure 5.33).

The force they exert on the material as a result is given by $\mathbf{F} = d\mathbf{P}/dt$. The force per unit area exerted by electromagnetic waves is called *radiation pressure*. Maxwell predicted the existence of radiation pressure and showed that if a beam of electromagnetic radiation is fully absorbed by a material, then the momentum transfer is given by:

$$\Delta \mathbf{P} = \frac{\Delta U}{c} \tag{5.102}$$

with ΔU the energy absorbed by the material over a time Δt and c the speed of light. If in contrast the radiation is completely reflected, the momentum transfer becomes twice as large, or:

$$\Delta \mathbf{P} = \frac{2\Delta U}{c} \tag{5.103}$$

For intermediate cases between total absorption and total reflection, one generalizes Equations 5.102 and 5.103 to read:

$$\Delta \mathbf{P} = \frac{\kappa \Delta U}{c} \tag{5.104}$$

Because a real surface will always absorb and reflect some of the energy, κ is a constant between 1 and 2. From Newton's second law we can calculate the force and the pressure. The average rate of energy in the beam is related to the Poynting vector by:

$$\frac{dU}{dt} = \mathbf{S}A = \frac{A}{\mu_0}(\mathbf{E} \times \mathbf{B}) \tag{5.105}$$

with A the cross-section of the object that the beam is hitting. The radiation pressure is given by:

$$\mathbf{P} = \frac{\mathbf{F}}{A} = \frac{1}{A}\frac{d\mathbf{P}}{dt} = \frac{1}{Ac}\frac{dU}{dt} = \frac{\mathbf{S}}{c}$$

if light is fully absorbed and (5.106)

$$\mathbf{P} = \frac{2\mathbf{S}}{c}$$

if light is fully reflected

Laser tweezers are another example dramatically illustrating the utility of radiation pressure (Figure 5.34). With optical tweezers one can manipulate nanometer and micrometer-sized dielectric particles by relying on the small forces contained in a highly focused laser beam (even a single atom can be manipulated!. Arthur Ashkin and colleagues first demonstrated laser tweezers in 1970 at Bell Labs.[4] With forces in the pN range optical tweezers can be used to manipulate and move small objects, measure small forces, and build up structures of small particles. The "Lilliputian" optical force pulls the object toward the focal volume of the beam, and

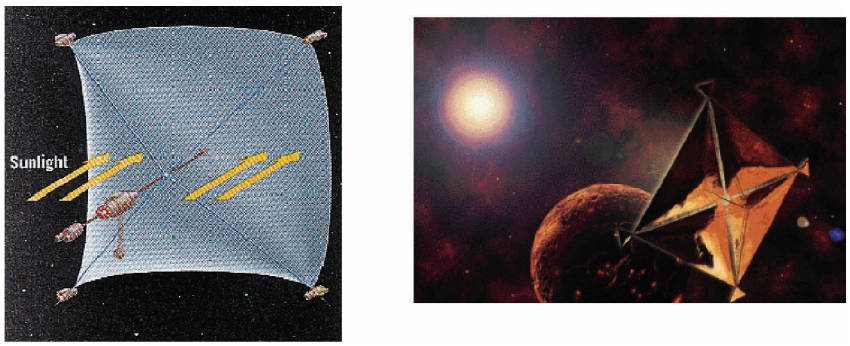


FIGURE 5.33 Sun sail. Sunlight striking the sail creates a force that pushes the spaceship away from the sun, much as the wind propels a sailboat.

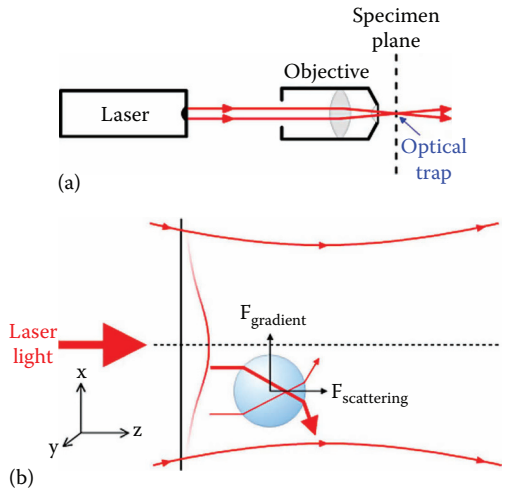


FIGURE 5.34 Laser tweezers—an example of the use of radiation pressure. (a) The beam is focused through a microscope objective. The beam waist, contains a very strong electric field gradient. Dielectric particles are attracted along the gradient to the region of strongest electric field, which is the center of the beam. (b) This spot creates an "optical trap" which is able to hold a small particle at its center. The forces felt by this particle consist of the light scattering and gradient forces due to the interaction of the particle with the light.

by moving the focal volume of the beam one can cause the object to move. In Figure 5.34a we show a laser beam focused by a microscope objective into a small spot on a specimen (e.g., a nanoparticle). This focussed spot constitutes an "optical trap" capable of holding a small particle at its center. In Figure 5.34b we illustrate how exactly such an optical trap works. Two forces are shown to be at play : 1) $F_{\text{scattering}}$ which is the scattering force and points in the direction of the incident light (z, see axis), and 2) F_{gradient}, which is the gradient force, due to the gradient nature of the Gaussian light intensity profile (x-y plane). As we just learned, light carries momentum and momentum is proportional to its energy in the direction of its propagation. A change in direction of the incoming light by reflection or refraction, results in a change of the momentum of the incoming photons. That momentum change, conservation of momentum dictates, causes the dielectric particles in the path of the light to undergo an equal and opposite momentum change and this is what causes a force to act on particles. When the light of a Gaussian laser beam interacts with a small particle, the light rays are bent (two example rays are illustrated in Figure 5.34b). Summing all the forces results in a resultant force for $F_{\text{scattering}}$ (z-direction) and one for F_{gradient} (x-y plane). With the contribution to $F_{\text{scattering}}$ of the refracted rays larger than that of the reflected rays a restoring force is created along the z-axis, and a stable trap is established. In modern equipment the image of the trapped particle can be projected onto a quadrant photodiode to measure nm-scale displacements. A continuous wave (CW), low-power laser was used to "optically trap" individual bacteria and protozoa. Steven Chu, who was at Stanford at the time, showed that lasers could be used to manipulate molecules. He and his team famously attached polystyrene beads at the end of DNA and pulled on the beads to stretch the DNA molecule. Others have used laser tweezers to study kinesin motors (see Volume III, Figure 2.33). Laser scissors were in use earlier than laser tweezers.[5] Laser scissors are short-pulsed and high-irradiance (energy reaching the surface in a given amount of time) beams, whereas tweezers are continuous and of low irradiance. The CW lasers can cause a 1.15–1.45°C increase per 100 mW of laser power. If the heat dissipation is inefficient, the laser can produce 10 times that power. Therefore, in the case of tweezers, the target must be transparent to not store significant energy.

The Complex Refractive Index n_c and the Complex Wave Vector **k**

As we saw above, at optical frequencies, electric fields in homogeneous media have solutions of the form $\mathbf{E} = \mathbf{E}_m e^{i(k_x \cdot x - \omega t)}$ and $\mathbf{B} = \mathbf{B}_m e^{i(k_x \cdot x - \omega t)}$ ($\mathbf{E}$ along the z-axis, $\mathbf{B}$ along the y-axis, and $\mathbf{k}$ along the x-axis; see Equation 5.70). For nonmagnetic materials with $\mu = \mu_0$, this results in a dispersion equation $\mathbf{k}^2 = \frac{\omega^2}{v^2}\varepsilon(\omega)$ (Equation 5.80), and with $n = \frac{c}{v}$ this can also be expressed as $\mathbf{k}^2 = \frac{n^2}{c^2}\omega^2$ (Equation 5.81). This dispersion relation for the case where the refractive index n is a real number correctly describes transparent materials. To describe a material that is also absorbing and features the anomalous regions we typically observe in plots of the refractive index as a function of frequency (see, for example, Figure 5.29), we need to introduce a complex wave vector $\mathbf{k}$ and a complex refractive index n_c.

Thus, the real wave vector $\mathbf{k} = \frac{n}{c}\omega = \frac{2\pi}{\lambda / n}$ with a real refractive index n must be replaced by a complex wave vector $\mathbf{k} = \frac{n_c}{c}\omega = \frac{2\pi}{\lambda / n_c}$ with a complex refractive index n_c.

The complex refractive index n_c is given as:

$$n_c = n + i\kappa \tag{5.107}$$

with n the refractive index and κ the extinction coefficient of the material. It follows that the dispersion Equation 5.81 then is transformed as:

$$\mathbf{k}(\omega) = \pm\frac{(n + i\kappa)\omega}{c} \tag{5.108}$$

The Kramers-Krönig relation links the real and complex parts of the complex refractive index (i.e., the refractive index n and κ the extinction coefficient). If you know one, you can calculate the other one.

The macroscopic quantities n and κ in Equation 5.108 will now be derived from the microscopic functions χ and ε_r, i.e., from the complex electric susceptibility and the complex permittivity, respectively.

Substituting the complex wave vector in the proposed wave solution (Equation 5.70), we obtain:

$$E_x(x,t) = E_m(x,\omega)e^{\frac{\omega\kappa}{c}x}\, e^{i\left(\frac{\omega n}{c}x - \omega\kappa\right)} \tag{5.109}$$

damping/ absorption — propagation

which represents a traveling damped wave (Figure 5.35).

The optical intensity of this electromagnetic wave is calculated as:

$$I \propto |E|^2 = I_0 e^{\frac{2\omega\kappa}{c}x} = I_0\, e^{-\alpha x} \tag{5.110}$$

with the absorption coefficient $\alpha = 2\frac{\kappa\omega}{c} = \frac{4\pi\kappa}{\lambda} = 2\mathbf{k}\kappa$. Note that $\alpha > 0$ means absorption and $\alpha < 0$ means gain (see lasers at the end of this chapter), and a propagation coefficient is defined as $\beta = \left(\frac{\omega}{c}\right)n = \mathbf{k}n$. The complex wave vector in Equation 5.108 can be written out in terms of this absorption coefficient α and the propagation coefficient β as:

$$\mathbf{k} = \mathbf{k}' + i\mathbf{k}'' = \beta - i\frac{\alpha}{2} \tag{5.111}$$

The real part of the complex wave vector ($\mathbf{k}$) is $\mathbf{k}' = \beta$ and the imaginary part $\mathbf{k}'' = \alpha/2$; in other words, the real part of the complex wave vector corresponds

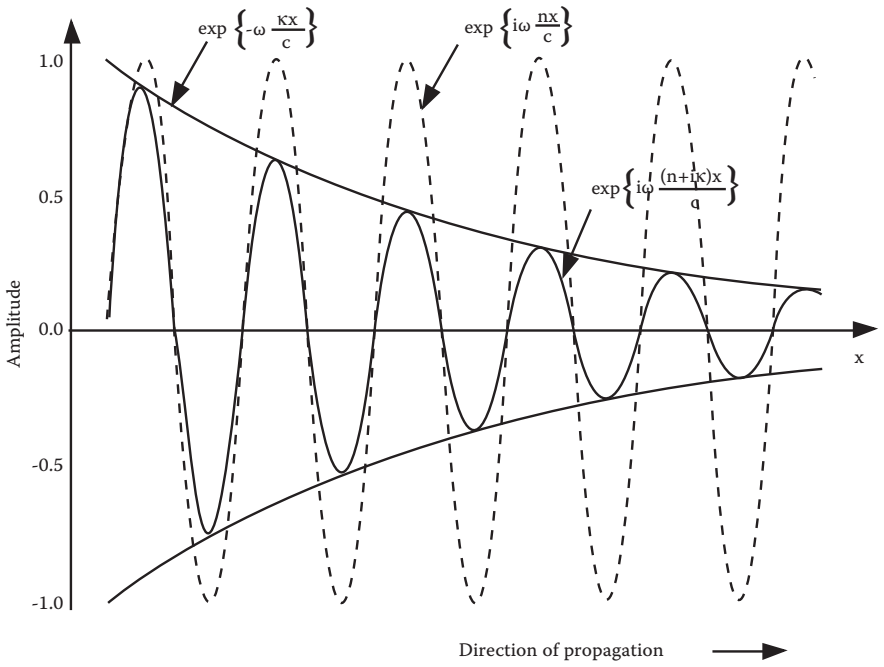


FIGURE 5.35 A traveling damped wave.

to the propagation coefficient and the imaginary part corresponds to the absorption coefficient of the material. Substituting the complex wave vector, written out in this format, into the proposed wave solution [$\mathbf{E} = \mathbf{E}_m e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$] (for simplicity we have dropped the sub *x* for the wave vector **k**), we obtain:

$$\begin{aligned}\mathbf{E}_x(x,t) &= \mathbf{E}_m(x,\omega)e^{-\mathbf{k}''x}e^{i(\mathbf{k}'x-\omega t)} \\ &= \underbrace{\mathbf{E}_m(x,\omega)e^{\frac{\alpha}{2}x}}_{\text{absorption}}\underbrace{e^{i(\beta x-\omega t)}}_{\text{propagation}}\end{aligned} \tag{5.112}$$

The complex refractive index *n* results from a complex electric susceptibility tensor $\chi_e(\omega)$ as:

$$\begin{aligned}n_c(\omega) &= n(\omega) + i\kappa(\omega) = \sqrt{1+\chi_e(\omega)} \\ &= \sqrt{1+\chi'(\omega)+i\chi''(\omega)}\end{aligned} \tag{5.113}$$

or also:

$$n_c(\omega) = n - i\frac{\alpha}{2\mathbf{k}} = \sqrt{1+\chi'(\omega)+i\chi''(\omega)} \tag{5.114}$$

We still need to introduce the relationship between the complex refractive index n_c and the complex dielectric constant. To do this we take the square of Equation 5.114 and group the real parts and imaginary parts together and relate n' and n'' to ε_r' and ε_r''. From $n_c(\omega)^2 = \varepsilon_r = \varepsilon_r' + i\varepsilon_r'$, it follows that:

$$\begin{aligned}&n_c(\omega) = n(\omega) + i\kappa(\omega) = \sqrt{\varepsilon_r'(\omega) + i\varepsilon_r''(\omega)} \\ &\varepsilon_r'(\omega) = [n(\omega)]^2 - [\kappa(\omega)]^2 = 1 + \chi'(\omega) = \frac{\varepsilon'}{\varepsilon^0} \\ &\varepsilon_r''(\omega) = 2n(\omega)\kappa(\omega) = \frac{\varepsilon''(\omega)}{\varepsilon_0(\omega)} = \chi''(\omega)\end{aligned} \tag{5.115}$$

and:

$$\kappa(\omega) = \frac{\varepsilon_r''(\omega)}{2n} = \frac{\chi''(\omega)}{2n} \tag{5.116}$$

to finally obtain:

$$\begin{aligned}n(\omega) &= \sqrt{\frac{(\varepsilon_r')^2 + (\varepsilon_r'')^2 + \varepsilon_r'}{2}} \\ \kappa(\omega) &= \sqrt{\frac{(\varepsilon_r')^2 + (\varepsilon_r'')^2 - \varepsilon_r'}{2}}\end{aligned} \tag{5.117}$$

for transparent materials $\kappa \lll 1$, and Equation 5.115 reduces to:

$$n_c(\omega) = n(\omega) = \sqrt{\varepsilon_r'(\omega)} \tag{5.118}$$

For weakly absorbing materials (dilute media) with both the real and imaginary parts of the electric susceptibility tensor $|\chi'(\omega)|$ and $||\chi''(\omega)| \lll 1$, $\sqrt{1+\chi'(\omega)+i\chi''(\omega)} \approx 1 + \frac{1}{2}[\chi'(\omega) + i\chi''(\omega)]$ one obtains the following relation between *n* and $\chi'(\omega)$ and κ''' and $\chi'(\omega)$:

$$n(\omega) \approx 1 + \frac{1}{2}\chi'(\omega) \tag{5.119}$$

and:

$$\kappa(\omega) \approx \frac{\chi''(\omega)}{2} \tag{5.120}$$

So the absorption coefficient for small $\chi'(\omega)$ and $\chi''(\omega)$ is given by:

$$\alpha(\omega) \approx \frac{\omega}{c}\chi''(\omega) \tag{5.121}$$

The different real and imaginary partner pairs we introduced so far: $\mathbf{k}'$ and $\mathbf{k}''$, *n* and κ, χ' and χ'', ε' and ε'' all describe the same physics, but for some problems one set is preferable over the others. The complex wave vector pair $\mathbf{k}'$ and $\mathbf{k}''$ and the refractive index *n* and the extinction κ pair is mostly used when discussing wave propagation. The other pairs, χ' and χ'' and ε' and ε'', are used most often when discussing the microscopic origin of optical effects. Each imaginary component can always be calculated from the real component, and vice versa (see Kramers-Krönig relationship).

Summarizing, the polarization of matter (**P**) induced by an applied electromagnetic field (**E**) is not instantaneous, making **P** and therefore also χ, *n*, and **k** frequency-dependent:

$$\mathbf{P}(\mathbf{k},\omega) = \chi_e(\mathbf{k},\omega)\varepsilon_o\mathbf{E}(\mathbf{k},\omega) \Rightarrow$$

$$\begin{array}{llll} \chi'(\omega) & n(\omega) & \mathbf{k}'(\omega) = \beta = \mathbf{k}n & \text{Phase propagation} \\ \Rightarrow & \Rightarrow & & \\ \chi''(\omega) & \kappa(\omega) & \mathbf{k}''(\omega) = \frac{\alpha}{2}\mathbf{k}\kappa & \text{Absorption} \end{array}$$

Example 5.2: Magnitude of κ
With $\alpha = 1\ cm^{-1}$, then $\kappa = \alpha/2\mathbf{k}$. And $\mathbf{k}$ is about $10^7\ m^{-1}$ for visible light, so $\kappa \sim 10^{-5}$, $\ll n$. Hence, $n^2 = 1 + \chi'(\omega)$ is a good approximation, provided $\alpha < 10^4\ cm^{-1}$ (see Equation 5.115). Then we can say that refraction is solely related to χ' away from resonances.

Electromagnetic Waves at Boundaries

Reflectance R and Transmittance T of Electromagnetic Radiation

Different from mechanical waves, transferred through a medium, light does not need a medium. Electrons in materials are vibrated and emit energy in the form of photons, which propagate across the universe (photons have no mass but are pure energy). Electromagnetic (EM) waves are waves that are made up of these "photons." When these photons come in contact with physical boundaries, EM waves interact like other waves would.

Incoming radiation that interacts with an object may follow three paths: reflection, absorption, and transmission. Reflectance, *R*, is the amount of reflected radiation over the amount of incoming radiation power. Absorption, *A*, is the amount of absorbed radiation over the amount of incoming radiation power, and transmittance, *T*, is the amount of transmitted radiation over the incident power. Power continuity dictates that:

$$R + A + T = 1 \tag{5.122}$$

This expression is known as *Kirchoff's law*. For an opaque object (T = 0), what is not absorbed is reflected, or R + A = 1. For real surfaces, *A*, *T*, and *T* all vary with the incoming wavelength λ. Incoming light striking a solid object is schematically represented in Figure 5.36.

The decay of the intensity of an EM wave propagating in the *x*-direction in an absorbing material, Beer's law [$I_0 = I(x = 0)$], is calculated in Chapter 7 as:

$$I(x) = I_0 e^{-\alpha x} \tag{5.110}$$

so that we calculate for the transmittance:

$$T = (1 - R)^2 e^{-\alpha l} \tag{5.123}$$

where *l* is the thickness of the solid piece and where we have assumed that the reflectivity of the front face and back face of that piece of material is identical.

Link between Reflectance R and Complex Refractive Index n_c

If we could relate the relatively easy-to-measure reflectance *R* of a material to the more difficult to access complex refractive index n_c, we then might be able to obtain both the real and imaginary parts of the complex refractive index n_c, i.e., *n* and κ, by using the Kramers-Krönig relation, which links the real and complex parts of the complex refractive index.

It was Augustin-Jean Fresnel (1788–1827) who first derived the equations that describe how light is reflected and transmitted at an interface between two media with different refractive indices. His equations give the reflection and transmission for waves parallel and perpendicular to the plane of incidence. For a dielectric medium where Snell's law (Equation 5.19) can be used to relate the incident and transmitted

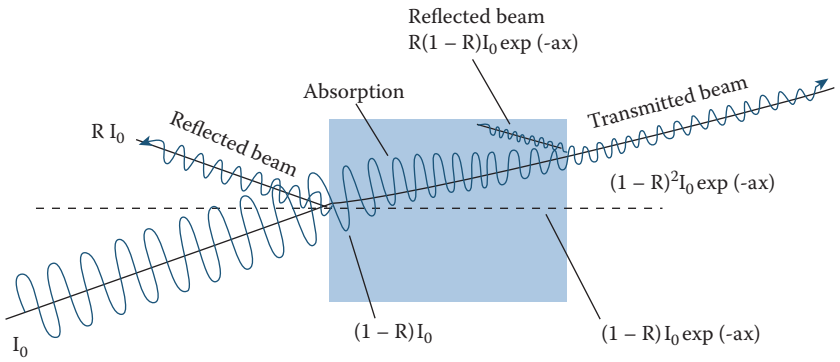


FIGURE 5.36 Radiation incident on a solid. With incoming beam of intensity I_0, externally reflected beam (RI_0), the intensity of the beam entering the solid $(1 - R_0)I_0$, the absorption of the beam inside the solid $(1 - R)I_0e^{-ax}$, that attenuated beam internally reflected $R(1 - R)I_0e^{-ax}$, and finally the transmitted beam: $(1 - R)^2I_0e^{-ax}$, where we have assumed that the reflectivity *R* is the same for the light coming in as it is for the light going out.

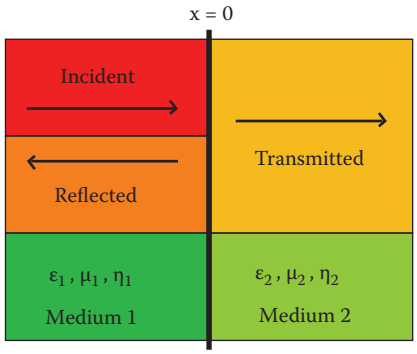


FIGURE 5.37 Wave reflection and transmission at an interface (x = 0).

angles, Fresnel's equations can be stated in terms of angles of incidence and transmission. The reflection of light that the equations predict is known as *Fresnel reflection*. In Volume III, Chapter 6 on metrology, when describing ellipsometry, we will take full advantage of these equations to relate reflectance R of a material with its complex dielectric constant n_c.

To establish the relation between R and n_c, we consider a plane wave moving through medium 1 (ε_1, μ_1, and η_1) and incident normal on a medium 2 (ε_2, μ_2, and η_2) (see Figure 5.37). We now need to express the incident (i), reflected (r), and transmitted (t) waves in each medium and then apply the boundary conditions. Because the wave is incident normal onto the surface, all **E** and **B** fields are parallel to the surface. We are using the solutions of the wave equations (Equation 5.70) [with **E** along the z-axis, **B** along the y-axis, and the propagation vector **k** along the x-axis, and where the complex propagation vector (Equation 5.108)].

Boundary conditions for EM waves at an interface are established by applying the integral form of Maxwell's equations to a small region at the interface of the contacting media. The integral equations are assumed to remain valid for regions containing discontinuous media as in the current case. We are assuming an interface between two lossless media (no charges or currents) and use the integral in Equation 5.51b to find that the following constraint is imposed on the tangential **E** field ($\mathbf{E}_t$):

$$\mathbf{E}_{1t} = \mathbf{E}_{2t} \tag{5.124}$$

This simply states that the tangential component of a field is continuous across an interface. When media 1 and 2 are dielectrics with permittivities ε_1 and ε_2, respectively, we can add that:

$$\frac{\mathbf{D}_{1t}}{\varepsilon_1} = \frac{\mathbf{D}_{2t}}{\varepsilon_2} \tag{5.125}$$

From the integral in Equation 5.48b, the tangential component of the **H** field ($\mathbf{H}_t$) also is continuous across the boundary of the two lossless media:

$$\mathbf{H}_{1t} = \mathbf{H}_{2t} \tag{5.126}$$

and for linear media we can deduce:

$$\frac{\mathbf{B}_{1t}}{\mathbf{B}_{2t}} = \frac{\mu_1}{\mu_2} \tag{5.127}$$

Similarly, the boundary conditions for the normal components of the **D** and **B** fields ($\mathbf{D}_n$ and $\mathbf{B}_n$) are obtained from the integrals in Equations 5.49b and 5.50b:

$$\mathbf{B}_{1n} = \mathbf{B}_{2n} \tag{5.128}$$

or, the normal component of a **B** field is continuous across an interface, and for linear media it holds that:

$$\mu_1\mathbf{H}_{1n} = \mu_2\mathbf{H}_{2n} \tag{5.129}$$

Finally, one also finds that:

$$\mathbf{D}_{1n} = \mathbf{D}_{2n} \tag{5.130}$$

and for linear media:

$$\varepsilon_1\mathbf{E}_{1n} = \varepsilon_2\mathbf{E}_{2n} \tag{5.131}$$

These results are summarized in Table 5.2. In the section on surface plasmonics, below, we will analyze these boundary conditions for the case of an interface between a dielectric and a conductor.

Let us denote the incident wave with $\mathbf{E}^i(x) = \mathbf{E}^i_m e^{i(k\cdot x-\omega t)}$ and $\mathbf{H}^i(x) = \frac{\mathbf{E}^i_m}{\eta_1} e^{i(k\cdot x-\omega t)}$, the reflected wave

TABLE 5.2 Boundary Conditions between Two Lossless Media

$\mathbf{E}_{1t} = \mathbf{E}_{2t} \Rightarrow \frac{\mathbf{D}_{1t}}{\varepsilon_1} = \frac{\mathbf{D}_{2t}}{\varepsilon_2}$
$\mathbf{H}_{1t} = \mathbf{H}_{2t} \Rightarrow \frac{\mathbf{B}_{1t}}{\mathbf{B}_{2t}} = \frac{\mu_1}{\mu_2}$
$\mathbf{D}_{1n} = \mathbf{D}_{2n} \Rightarrow \varepsilon_1\mathbf{E}_{1n} = \varepsilon_2\mathbf{E}_{2n}$
$\mathbf{B}_{1n} = \mathbf{B}_{2n} \Rightarrow \mu_1\mathbf{H}_{1n} = \mu_2\mathbf{H}_{2n}$

with $\mathbf{E}^r(x) = \mathbf{E}^r_m e^{i(\mathbf{k}\cdot x-\omega t)}$ and $\mathbf{H}^r(x) = \frac{\mathbf{E}^r_m}{\eta_1} e^{i(\mathbf{k}\cdot x-\omega t)}$, and the transmitted wave with $\mathbf{E}^t(x) = \mathbf{E}^t_m e^{i(\mathbf{k}\cdot x-\omega t)}$ and $\mathbf{H}^t(x) = \frac{\mathbf{E}^r_m}{\eta_2} e^{i(\mathbf{k}\cdot x-\omega t)}$. From the expression for the intrinsic impedance $\eta = \frac{|\mathbf{E}|}{|\mathbf{H}|} = \frac{\omega\mu}{|\mathbf{k}|} = \frac{|\mathbf{k}|}{\omega\varepsilon} = \sqrt{\frac{\mu}{\varepsilon}}$ (Equation 5.83), we also introduced here the fact that $\frac{\mathbf{E}^i_m}{\eta_1} = \mathbf{H}^i_m$, $\frac{\mathbf{E}^r_m}{\eta_1} = \mathbf{H}^r_m$, and $\frac{\mathbf{E}^t_m}{\eta_2} = \mathbf{H}^t_m$. Like the velocity of propagation c, the intrinsic impedance is independent of the source and is determined only by the properties of the medium. The two unknowns in these equations are $\mathbf{E}^r_m$ and $\mathbf{E}^t_m$, which we will now retrieve from the above-derived boundary conditions, i.e., the tangential **E** and **H** fields must be continuous at x = 0 (Equations 5.124 and 5.126). Fields parallel and perpendicular to the plane of incidence are considered *p*- and *s*-polarized, respectively. These two components are independent and can be calculated separately. Fresnel calculated the amount of light reflected and transmitted between materials of different refractive index, and for the tangential fields (||) in medium 1 he deduced:

$$\begin{aligned} E_{t1}(x=0) &= E^i_m(x=0) + E^r_m(x=0) \\ &\text{Total tangential E} \quad \text{Incident E} \quad \text{Reflected E} \\ H_{t1}(x=0) &= H^i_m(x=0) \\ &\text{Total tangential H} \quad \text{Incident H} \\ + H^r_m(x=0) &= \frac{1}{\eta_1}(E^i_m - E^r_m) \\ &\text{Reflected H} \end{aligned} \tag{5.132}$$

And for medium 2 he derived:

$$E_{t2}(x=0) = E^t(x=0) = E^t_m$$

Therefore, at x = 0:

$$E^i_m + E^r_m = E^t_m \tag{5.133}$$

$$H_{t2} = H^t(x=0) = \frac{1}{\eta_2} E^t_m$$

Therefore, from Equations 5.132 and 5.133, we obtain:

$$\frac{1}{\eta_1}(E^i_m - E^r_m) = \frac{1}{\eta_2} E^t_m \tag{5.134}$$

and we can solve now for the unknowns $\mathbf{E}^r_m$ and $\mathbf{E}^t_m$ in terms of the known $\mathbf{E}^i_m$:

$$E^r_m = \frac{\eta_2 - \eta_1}{\eta_2 + \eta_1} E^i_m = r(\omega) E^i_m$$

and

$$E^t_m = \frac{2\eta_2}{\eta_2 + \eta_1} E^i_m = t(\omega) E^i_m \tag{5.135}$$

where $r(\omega)$, the dimensionless reflectivity coefficient, is a complex function defined at the solid surface as the ratio of the reflected electric field to the incident electric field, or for the tangential field (||) we obtain:

$$r(\omega)_{\parallel} = \left(\frac{E^r_m}{E^i_m}\right)_{\parallel} = \left(\frac{\eta_2 - \eta_1}{\eta_2 + \eta_1}\right)_{\parallel} \tag{5.136}$$

And for $t(\omega)$, the dimensionless transmission coefficient of the tangential field (||), we get:

$$t(\omega)_{\parallel} = \left(\frac{E^t_m}{E^i_m}\right)_{\parallel} = \left(\frac{2\eta_2}{\eta_2 + \eta_1}\right)_{\parallel} \tag{5.137}$$

Equations 5.136 and 5.137 are the Fresnel expressions for the tangential fields. For the normal field (⊥) one obtains these additional Fresnel expressions:

$$r(\omega)_{\perp} = \left(\frac{E^r_m}{E^i_m}\right)_{\perp} = \left(\frac{\eta_4 - \eta_3}{\eta_4 + \eta_3}\right)_{\perp} \tag{5.138}$$

and:

$$t(\omega)_{\perp} = \left(\frac{E^r_m}{E^i_m}\right)_{\perp} = \left(\frac{2\eta_4}{\eta_4 + \eta_3}\right)_{\perp} \tag{5.139}$$

In Figure 5.38 we have sketched a planar interface, and the "incident" medium occupies the half-space $y > 0$ and has an index of refraction n_i. The

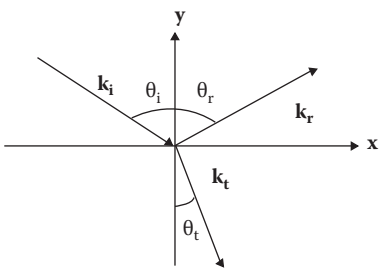


FIGURE 5.38 Light reflects and refracts according to Snell's law.

"transmission medium" occupies the half-space $y < 0$ and has an index of refraction n_t. For simplicity, we only consider the electric fields, and we obtain the four amplitude coefficients (the Fresnel equations) as:

$$r(\omega)_{\parallel} = \left(\frac{E_m^r}{E_m^i}\right)_{\parallel} = \left[\frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}\right]_{\parallel} \quad (5.140)$$

$$r(\omega)_{\perp} = \left(\frac{E_m^r}{E_m^i}\right)_{\perp} = \left[-\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}\right]_{\perp} \quad (5.141)$$

$$t(\omega)_{\parallel} = \left(\frac{E_m^t}{E_m^i}\right)_{\parallel} = \left[\frac{2\sin\theta_t \cos\theta_i}{\sin(\theta_i + \theta_t)\cos(\theta_i - \theta_t)}\right]_{\parallel} \quad (5.142)$$

$$t(\omega)_{\perp} = \left(\frac{E_m^r}{E_m^i}\right)_{\perp} = \left[\frac{2\sin\theta_t \cos\theta_i}{\sin(\theta_i + \theta_t)}\right]_{\perp} \quad (5.143)$$

Note that the reflection coefficient r in Equations 5.136 and 5.138 can be positive or negative, depending on the relative sizes of the intrinsic impedances of the two media. The transmission coefficient t, on the other hand, is always positive (see Equations 5.137 and 5.139). Importantly, Equations 5.136–5.139 also hold when the media are dissipative, i.e., when η_1 and/or η_2 are complex. A complex r (or t) means that a phase shift is introduced at the interface on reflection (or transmission). Reflection and transmission are related by:

$$t(\omega) = 1 + r(\omega) \quad (5.144)$$

The same boundary conditions derived above require the components of the complex wave vector $\mathbf{k}$ (see Equation 5.111) parallel to the surface to be the same on both sides of the boundary. When a sinusoidal plane wave, with angular frequency ω and propagation wave vector $\mathbf{k}$, travels in the x-direction in a linear dielectric with a scalar permittivity and no free charge or currents present, any dissipative process leads to phase lag and a complex wave vector. Thus, a transmitted wave in a medium is attenuated because, by the dispersive relation for electromagnetic waves, the wave vector in the medium is related to the incident $\mathbf{k}$ in vacuum by $(n + i\kappa)\mathbf{k}$, or the incident wave, $\mathbf{E}^i(\mathbf{x}) = \mathbf{E}_m^i e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$, is modified to:

$$\begin{aligned} E_m^t &= E_m \exp[i(n + i\kappa)\mathbf{k}x - \omega t)] \\ &= \exp(-\kappa\mathbf{k}x)\exp[i(n\mathbf{k}x - \omega t)] \quad (5.145) \\ &\qquad \text{attenuation} \qquad \text{propagation} \end{aligned}$$

Continuity of $\mathbf{E}$ and $\mathbf{B}$ at the interface (surface of the medium) dictates that $\mathbf{E}_m^i + \mathbf{E}_m^r = \mathbf{E}_m^t$ (Equation 5.133), or:

$$\begin{aligned} kE_m^i - kE_m^r &= (n + i\kappa)kE_m^t \quad \text{or} \\ E_m^i - E_m^r &= (n + i\kappa)(E_m^i + E_m^r) \quad (5.146) \\ (n + i\kappa - 1)E_m^i &= (1 + n + i\kappa)E_m^r \end{aligned}$$

Thus, we obtain for the reflectivity at normal incidence:

$$r(\omega) = \frac{E_m^r}{E_m^i} = \left(\frac{n + i\kappa - 1}{n + i\kappa + 1}\right) = \frac{n_c - 1}{n_c + 1} \quad (5.147)$$

But the quantity measured in experiments is reflectance R, defined as the ratio of the reflected intensity to the incident intensity (see above), or:

$$R(\omega) = \frac{\left|E_m^r\right|^2}{\left|E_m^i\right|^2} = \left|\frac{n_c - 1}{n_c + 1}\right|^2 = \frac{(n-1)^2 + \kappa^2}{(n+1)^2 + \kappa^2} \quad (5.148)$$

When $R = 0$, there is no reflection, and this occurs with $n = 1$, $\kappa = 0$, and $R = 1$ when either n or κ becomes large. It is difficult to measure the phase $\theta(\omega)$ of the reflected wave, but it can be shown that it can be calculated from the measured reflectance $R(\omega)$ if this is known over all frequencies. With both $R(\omega)$ and $\theta(\omega)$ known we can use Equation 5.148 to calculate $n(\omega)$ and $\kappa(\omega)$. Above we saw that $n(\omega)$ and $\kappa(\omega)$ are related to the complex dielectric function ε_r as:

$$n_c(\omega) = n(\omega) + i\kappa(\omega) = \sqrt{\varepsilon_r'(\omega) + i\varepsilon_r''(\omega)} \quad (5.113 \text{ and } 5.115)$$

Thus, plugging the values for $n(\omega)$ and $\kappa(\omega)$ in Equation 5.113 we obtain the complex dielectric function.

The values for the refractive index n and extinction coefficient κ are not independent; they are linked by the Kramers-Krönig relations. The Kramers-Krönig relations enable us to find the real part of the response of a linear passive system if we know

Table 5.3 Optical Coefficients

Relative Dielectric Constant	$\varepsilon_r = 1 + \chi_e(\omega)$
Real part of refractive index	$n' = \sqrt{\dfrac{(\varepsilon_r')^2 + (\varepsilon_r'')^2 + \varepsilon_r'}{2}}$
Imaginary part of refractive index	$n'' = \sqrt{\dfrac{(\varepsilon_r')^2 + (\varepsilon_r'')^2 - \varepsilon_r'}{2}}$
Absorption coefficient	$\alpha = -2\dfrac{\omega}{c}n''$
Reflectivity	$r(\omega) = \dfrac{E_m^r}{E_m^i} = \left(\dfrac{n + i\kappa - 1}{n + i\kappa + 1}\right) = \dfrac{n_c - 1}{n_c + 1}$
Reflectance	$R(\omega) = \dfrac{\lvert E_m^r \rvert^2}{\lvert E_m^i \rvert^2} = \dfrac{(n-1)^2 + \kappa^2}{(n+1)^2 + \kappa^2}$

the imaginary part of the response at all frequencies, and vice versa:

$$n(\omega) - 1 = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{\kappa(\omega')}{\omega' - \omega} d\omega'$$
$$\kappa(\omega) = \frac{-1}{\pi} P \int_{-\infty}^{\infty} \frac{n(\omega') - 1}{\omega' - \omega} d\omega' \qquad (5.149)$$

where P denotes the Cauchy principal value. Either of these equations, plus a knowledge of $R(\omega)$ at all frequencies, permits one to disentangle the separate values of $n(\omega)$ and $\kappa(\omega)$.

The previous discussion makes it clear that the refractive index of a semiconductor can also be determined from a sometimes more practical transmission measurement.

In Table 5.3 we summarize the important optical coefficient properties deduced so far. In Table 5.4 we list a number of semiconductors with their transparency range, the refractive index, and the wavelength corresponding to their bandgap.

Example 5.3*: Surface reflectance R of Ge. At 400 nm the complex refractive index n_c for Ge = 4.141 + i2.215. Calculate the 1) wave propagation speed v, 2) absorption coefficient α, and 3) reflectance R.

1. $v = c/n = 2.998 \times 10^8/4.141 = 7.24 \times 10^7$ m/s
2. $\alpha = 4\pi \times 2.215/400 \times 10^{-9} = 6.96 \times 10^7$ m^{-1}
3. $R = \dfrac{(4.141 - 1)^2 + 2.215^2}{(4.141 + 1)^2 + 2.215^2} = 47.1\%$

* All three examples are from Mikael Mulot's class notes.

Table 5.4 Semiconductors: Transparency Range (T), Bandgap Wavelength, and Refractive Index

Crystal	Transparency Range (μm)	λ_g (μm)	n
Ge	1.8–23	1.8	4.00
Si	1.2–15	1.1	3.42
GaAs	1.0–20	0.87	3.16
CdTe	0.9–14	0.83	2.67
CdSe	0.75–24	0.71	2.50
ZnSe	0.45–20	0.44	2.41
ZnS	0.4–14	0.33	2.20

Example 5.4: Transmittance T through Si. The reflectance of a 10-μm-thick Si slice at 633 nm = 35% and $\alpha = 3.8 \times 10^{-5}$ m^{-1}. Calculate the transmittance T.

For $\alpha l = (3.8 \times 10^{-5})(10 \times 10^{-6}) = 3.8$. As a result we obtain for $T = (1 - R)^2 e^{-\alpha l} = (1–35\%)^2 e^{-3.8} = 0.0095$.

Example 5.5: Reflectance R and absorption coefficient α of NaCl. The complex relative dielectric constant at 60 μm for NaCl is –16.8 + i91.4. Calculate 1) α and 2) **R**:

$$n = \frac{1}{\sqrt{2}}\sqrt{-16.8 + \sqrt{(-16.8)^2 + 91.4^2}} = 6.17$$

1. With $\kappa = \dfrac{1}{\sqrt{2}}\sqrt{+16.8 + \sqrt{(-16.8)^2 + 91.4^2}} = 7.41$

we derive for $\alpha = \dfrac{4\pi\kappa}{\lambda} = \dfrac{4\pi \times 7.41}{60 \times 10^{-6}} = 1.55 \times 10^6\ \text{m}^{-1}$

2. $R = \dfrac{(6.17 - 1)^2 + 7.41^2}{(6.17 + 1)^2 + 7.41^2} = 76.8\%$

Bulk, Surface, and Localized Plasmons: Plasmonics

Introduction

Plasmons are electron density waves that can carry huge amounts of data and that can be squeezed into miniscule metal structures, holding the promise of a new generation of superfast computer chips, ultrasensitive molecular detectors, microscopes with improved subwavelength resolution (near-field scanning optical microscopes; NSOMs), better exploitation of nonlinear optics, sensitive optical studies of surfaces and interfaces, fabrication of an infinite variety of brightly colored metal nanoparticles,

better light-emitting diodes, higher-resolution lithography, and perhaps even the possibility of altering the electromagnetic field around an object as to render it invisible (cloaking). In 2000, Harry A. Atwater from Caltech gave the name of "plasmonics" to this promising emerging field.[6]

In Chapter 3 we saw how a metal may exhibit collective longitudinal oscillations of the free electron gas density, often at optical frequencies. A plasmon is a quantum of a plasma oscillation that is excited by passing an electron through a material or by reflecting an electron or a photon off a material. It is a result of the quantization of the plasma oscillations just as photons and phonons are quantizations of light and sound waves, respectively. In other words, reflected or transmitted electrons/photons lose an amount of energy that equals an integral number of times the plasmon energy. A plasmon can also couple with a photon to create a plasma polariton. The most discussed types of polaritons are phonon-polaritons, resulting from the coupling of an infrared photon with an optical phonon; exciton-polaritons, resulting from the coupling of visible light with an exciton; and surface plasmon-polaritons, resulting from the coupling of surface plasmons with light.

Plasmons can be observed experimentally using electron energy loss spectroscopy (EELS). In EELS, electrons are accelerated to a defined energy and are then shot onto a thin foil of the material under investigation. The number and energy of the transmitted electrons are measured. These electrons are able to excite plasmons, which then appear as a more or less sharp dip in the number of transmitted electrons at an energy loss equal to the plasma frequency.

Bulk Plasmons

In Chapter 3 we explained electron oscillations in a metal using the Drude model of metals. In the Drude model the metal is treated as a 3D crystal of positively charged ions with a delocalized electron gas moving in a periodic potential of this ion grid, i.e., as a plasma. Based on that model, we calculated the complex dielectric constant of the bulk plasma as $\varepsilon(\omega,k)=\varepsilon_0\left[1-\frac{\omega_p^2}{(\omega^2+i\gamma\omega)}\right]$ (Equation 3.40). The plasma exhibits resonance absorption at the bulk plasmon frequency ω_p, defined as $\omega_p=\sqrt{\frac{ne^2}{m_e\varepsilon_0}}$ (Equation 3.41), and at frequencies above ω_p, the metal becomes transparent and dielectric-like, whereas below it is reflective and conductive. The bulk plasmon resonance of silver, for example, occurs at $\hbar\omega_p = 3.76$ eV. In Figure 5.27, we saw that the dispersion relation, the relation between the energy of the system and its corresponding momentum $\mathbf{k}(\omega)$, in a vacuum, is a linear function of ω with a slope $c = v_g = v_p$ (velocity of light = group velocity = phase velocity). The straight line in Figure 5.27 is referred to as the *light line*. We also established that the dispersion relation in a medium is no longer a straight line but is given by: $\mathbf{k}(\omega)=\frac{\omega}{v_g}=\frac{n(\omega)\omega}{c}=\frac{\omega}{c}\sqrt{\varepsilon\mu}$ (Equation 5.88). If we consider a highly conductive system, where the damping factor γ is very small, the real component of the dielectric constant $\varepsilon_r'(\omega,\mathbf{k})\approx 1-\frac{\omega_p^2}{\omega^2}$ (Equation 3.44) dominates with the imaginary component $\varepsilon_r''(\omega,\mathbf{k})\approx\frac{\omega_p^2\gamma}{\omega^3}$ (Equation 3.45) tending to zero. In other words, Equation 3.44 is the expression for the dielectric constant when only the conduction electrons contribute, and damping is neglected. The Drude model, in this case, predicts a monotonous decrease of $\varepsilon_r(\omega,\mathbf{k})$ for decreasing frequency, and experiments confirm that $\varepsilon_r'(\omega,\mathbf{k})$ becomes zero when $\omega=\omega_p$ in such case (see Figure 3.9. When $\varepsilon_r'(\omega)$ becomes zero at $\omega=\omega_p$, the material supports free longitudinal collective modes for which all electrons oscillate in phase. The dispersion relation for bulk plasmons from Equation 5.88 is then given as:

$$\omega^2-\omega_p^2=c^2\mathbf{k}^2 \qquad (5.150)$$

as shown in Figure 5.39, where we also mark the light line ($\omega = c\mathbf{k}$). It is clearly seen that when ω is less than ω_p there is no propagating solutions and the wave vector is imaginary.

Surface Plasmons

Surface plasmons are those plasmons that are confined to surfaces and that interact strongly with light, resulting in a polariton. They occur at the interface of a vacuum or a material with a positive

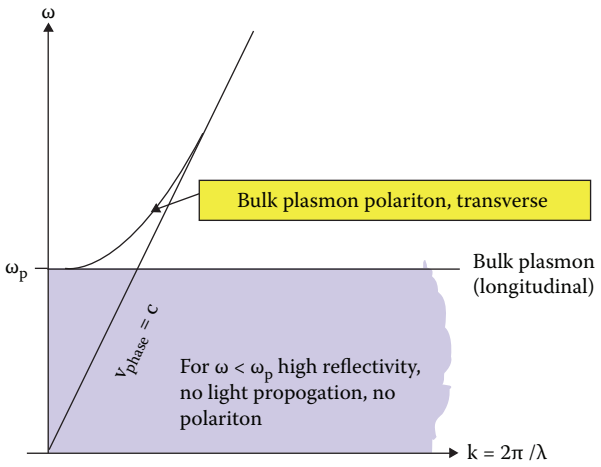


FIGURE 5.39 Dispersion of bulk plasmon. The light line is also marked ($v_{phase} = c$).

dielectric constant (a dielectric) and that of a negative dielectric constant (usually a metal or doped dielectric) (Figure 5.40).

Surface plasmons are a subset of the "eigenmodes" of the free electrons in a metal or doped dielectric, with the electronic oscillations being excited parallel to the surface of the metal and a surrounding dielectric. Because all electrons may participate in the oscillations, plasmons can interact strongly with light of the right frequency. With wave vectors much smaller than the Fermi wave vectors of metals, these wave modes can be described by Maxwell's equations. Quantum mechanical considerations only are necessary for the shortest plasmon wavelengths. At low-surface plasmon wave vectors, the behavior of these surface modes can be understood intuitively. In this regime, they are essentially transverse in character, and these transverse fields generate a polarization in the dielectric, which is aligned with the stimulating field. In the metal, however, the polarization is in the opposite direction of the applied field owing to its negative dielectric constant. Now the stimulating field is creating equal and opposite electric displacements (D), in phase with

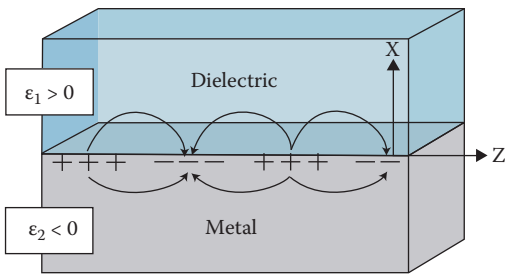


FIGURE 5.40 Surface plasmons at metal/dielectric interface.

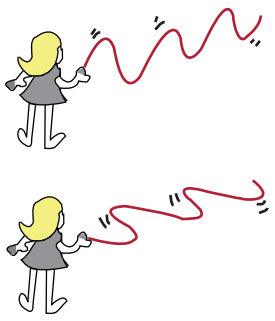
each other across an interface. These opposing electric displacements serve to attract and confine the current to the interface, thus generating the collective electron oscillations of the surface plasmon. Figure 5.41 illustrates vertical and horizontal polarized waves. In vertical polarization (*s*-polarization from the German *senkrecht*), the **E** field is perpendicular to the plane of incidence. In horizontal polarization (*p*-polarization), the **E** field is parallel to the plane of incidence.

Starting from the Maxwell's equations, we derive the characteristics for the plasmonic system shown in Figure 5.42a. Light that is *p*-polarized (see Figure 5.41) will create polarization charges at the interface as illustrated in Figure 5.42b. These charges give rise to surface plasmon modes. With an *s*-polarized incident radiation no such polarization charges are created at the interface; thus, these EM waves cannot excite surface plasmon modes.

We assume that the free current **J** is zero, that the relative permeability (μ_r) of both media to be unity, and that the wave is propagating in the x-direction. The EM field of surface plasmons propagates on the surface in the x-direction. The $\mathbf{H}_y$ is the magnetic field in the y-direction of the incident *p*-polarized wave. Thus, the components of the **E** and **H** fields are $\mathbf{E} = (\mathbf{E}_x, 0, \mathbf{E}_y)$ and $\mathbf{H} = (0, \mathbf{H}_y, 0)$, and the interface between the metal and dielectric is at $z = 0$.

As we saw above, the constraints imposed on the transverse **E** and **H** field components at a boundary are that:

$$\begin{aligned} \mathbf{E}_{1x} &= \mathbf{E}_{2x} \\ \mathbf{H}_{1y} &= \mathbf{H}_{2y} \end{aligned} \tag{5.151}$$



Illustrating vertical and horizontal polarized waves.

FIGURE 5.41 The *p*-polarization: **E** field is parallel to the plane of incidence, and *s*-polarization: **E** field is perpendicular to the plane of incidence (German *senkrecht* = perpendicular).

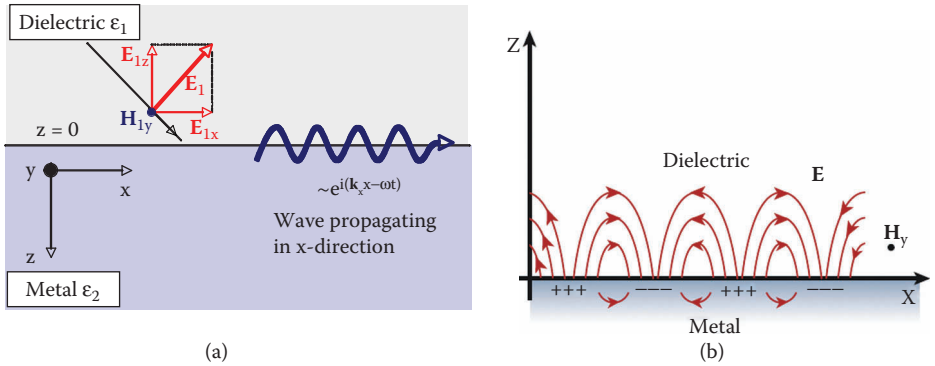


FIGURE 5.42 (a) Surface plasmons a subset of the "eigenmodes" of the free electrons in a metal. Electronic oscillations being excited parallel to the surface of the metal and a surrounding dielectric. The local field intensity depends on the wavelength. (b) Polarization charges are created at the interface between two materials.

i.e., the components of the **E** and **H** fields are continuous across the interface from metal to insulator.

It is also informative to look at the boundary condition for the normal component of the **D** field:

$$\mathbf{D}_{1z} = \mathbf{D}_{2z} \tag{5.152}$$

which can be rewritten as:

$$\begin{aligned} \varepsilon_0\mathbf{E}_{0z} + \mathbf{P}_{1z} &= \varepsilon_0\mathbf{E}_{0z} + \mathbf{P}_{2z} \\ \downarrow \qquad & \qquad \downarrow \end{aligned} \tag{5.153}$$

Creation of the polarization charges

If one of the contacting materials is a metal, the electrons respond to the polarization by the *p*-polarized light (Figure 5.42, right), and this gives rise to surface plasmon modes. To model these waves we are looking for a localized surface mode, decaying into both materials as shown in Figure 5.43a, where the exponential dependence of the field E_z is shown. The decay in the two media is characterized by the penetration depth, i.e., the depth at which value of E_z falls to $1/e \approx 1/|\kappa_z|$ (see Equations 5.111 and 5.112: $\mathbf{k} \sim \mathbf{k}'' \sim \alpha \sim \kappa$ (because the absorption coefficient $\alpha = 2\kappa\omega/c = 2\mathbf{k}\kappa$, with κ the extinction coefficient).

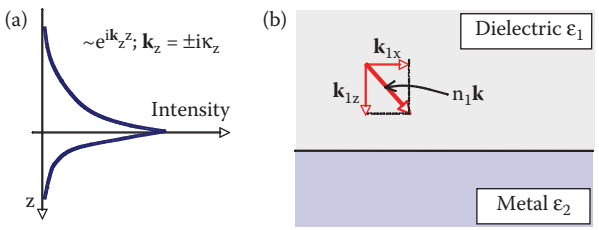


FIGURE 5.43 (a) We are looking for a localized surface mode, decaying into both materials so that $\mathbf{k}_z$ has to be imaginary; (b) n_1 is the dielectric constant of the dielectric.

Thus, a suggested solution for the surface plasmon mode of wave propagation in the dielectric may be written as:

$$\begin{aligned} \mathbf{H}_1 &= (0, \mathbf{H}_{1y}, 0)e^{i(\mathbf{k}_{1x}x + \mathbf{k}_{1z}z - \omega t)} \\ \mathbf{E}_1 &= (\mathbf{E}_{1x}, 0, \mathbf{E}_{1y})e^{i(\mathbf{k}_{1x}x + \mathbf{k}_{1z}z - \omega t)} \end{aligned} \tag{5.154}$$

and for the metal:

$$\begin{aligned} \mathbf{H}_2 &= (0, \mathbf{H}_{2y}, 0)e^{i(\mathbf{k}_{2x}x + \mathbf{k}_{2z}z - \omega t)} \\ \mathbf{E}_2 &= (\mathbf{E}_{2x}, 0, \mathbf{E}_{2y})e^{i(\mathbf{k}_{2x}x + \mathbf{k}_{2z}z - \omega t)} \end{aligned} \tag{5.155}$$

Substituting these suggested solutions in Maxwell's equations and applying the continuity boundary conditions, i.e., $\mathbf{k}_{1z} = \mathbf{k}_{2z}$ (because continuity holds for both **E** and **H**: Equation 5.151), we find the following additional constraint for the wave vector:

$$\frac{\varepsilon_1}{\mathbf{k}_{1z}} = \frac{\varepsilon_2}{\mathbf{k}_{2z}} \tag{5.156}$$

From the vector diagram in Figure 5.43b, we see that $n_1\mathbf{k}$ is given as:

$$(n_1\mathbf{k})^2 = \mathbf{k}_{1x}^2 + \mathbf{k}_{1z}^2 \text{ or also } \varepsilon_1(\omega, \mathbf{k})\frac{\omega^2}{c^2} = \mathbf{k}_{1x}^2 + \mathbf{k}_{1z}^2 \tag{5.157}$$

and this enables us to establish the sign of $\mathbf{k}_z$:

$$\begin{aligned} \mathbf{k}_{1z} &= -\sqrt{(n_1\mathbf{k})^2 - \mathbf{k}_{1x}^2} \\ \mathbf{k}_{2z} &= +\sqrt{(n_1\mathbf{k})^2 - \mathbf{k}_{2x}^2} \end{aligned} \tag{5.158}$$

because $\mathbf{k}_{1z}$ and $\mathbf{k}_{2z}$ are of opposite signs, the condition in Equation 5.156 will be satisfied only if ε_1 and

ε_2 are of opposite signs as well. This is the case when one material is dielectric, $\varepsilon_1 > 0$, and the second material is metal, $\varepsilon_2 < 0$. It also follows from this equation, because $\mathbf{k}_{1z}$ and $\mathbf{k}_{2z}$ must both be imaginary, that:

$$\left(n_1\mathbf{k}\right)^2 - \mathbf{k}_{1k}^2 < 0 \quad \text{so that } \mathbf{k}_{1x} > n_1\mathbf{k} \tag{5.159}$$

and:

$$\left(n_2\mathbf{k}\right)^2 - \mathbf{k}_{2x}^2 < 0 \quad \text{so that } \mathbf{k}_{2x} > n_2\mathbf{k} \tag{5.160}$$

The condition in Equation 5.160 is always satisfied for metals because:

$$\mathbf{k}_{2x}^2 > \varepsilon_2\mathbf{k}^2 \tag{5.161}$$

with ε_2 for a metal <0. Of course, this result is to be expected from our initial assumptions. By defining the mode to have an exponential decay normal to the surface, we ensured an imaginary wave vector in this dimension. The absolute square of this positive quantity then adds to the light line wave vector to determine $\mathbf{k}_{1x}$ and $\mathbf{k}_{2x}$ (see Equations 5.159 and 5.160); hence $\mathbf{k}$ must always be greater than that of the free space field. This means that a surface plasmon has a greater momentum than a free photon at the same frequency. This makes a direct excitation (by shining light on a dielectric/metal interface) of a surface plasmon mode impossible (see inset in Figure 5.44)!

To calculate the dispersion relation for surface plasmon polaritons, we calculate $\mathbf{k}$ by substituting the expression for $\mathbf{k}_{1z}$ and $\mathbf{k}_{2z}$ from Equation 5.158 in Equation 5.156 to obtain:

$$\mathbf{k}_x = \mathbf{k}\sqrt{\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1+\varepsilon_2}} = \frac{\omega}{c}\sqrt{\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1+\varepsilon_2}} \tag{5.162}$$

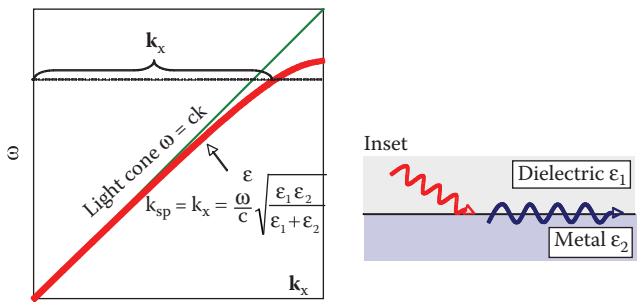


FIGURE 5.44 Dispersion of a surface plasmon. The surface plasmon mode always lies beyond the light line, i.e., it has greater momentum than a free photon of the same frequency. Inset: It is impossible to excite a plasmon mode by shining light on a dielectric/metal interface.

The surface plasmon condition is $\varepsilon_1 = -\varepsilon_2$, which occurs at high $\mathbf{k}_x$ when $\omega = \omega_{sp}$, which we calculate from:

$$\varepsilon_1 = 1 - \frac{\omega_p^2}{\omega^2} = -\varepsilon_2 \text{ so that } \omega_{sp} = \frac{\omega_p}{\sqrt{1+\varepsilon_2}} \tag{5.163}$$

In the case we deal with air as the dielectric ($\varepsilon_1 = 1$), the resonance frequency $\omega_{sp} = \omega_p/\sqrt{2}$ (see Table 5.6).

The power of plasmonics can be understood from an inspection of the dispersion relationship in Equation 5.162. From this dispersion relationship it is important to notice that the wave vector is no longer a linear function of the permittivity of a standard dielectric. Because we have the sum of dielectrics of opposite sign in the denominator, very large wave vectors are possible. Therefore, surface plasmons have the unique ability to achieve very short wavelengths ($\mathbf{k}_s = 2\pi/\lambda_{sp}$) at modest optical frequencies. The reason surface plasmon modes can achieve these anomalously high wave vectors at visible frequencies is because they are mediated by electrons rather than free space optical fields.

Because these modes are mediated by collective electron oscillations with subangstrom wavelengths, they may achieve very large optical wave vectors. From the same equation it follows that, using geometries of conductors (such as metals or doped semiconductors) with dielectrics (such as air or glass), surface plasmon modes at optical frequencies can be created with effective indices of refraction that are orders of magnitude higher than those of the constituent materials. In fact, these indices can be so high as to create x-ray wavelengths (less than 10 nm) with visible frequencies. This is of major consequence for microscopy: it is evident that deeply subwavelength focal spots cannot be formed through conventional focusing using a lens system or microscope objective. This is due primarily to the lack of high-index media at visible frequencies. But here we show that we are able to achieve a high effective refractive index with conventional optical materials. Although the permittivities of these material combinations may be modest, their geometry and interactions create an effective refractive index much larger than that available from conventional transparent media.

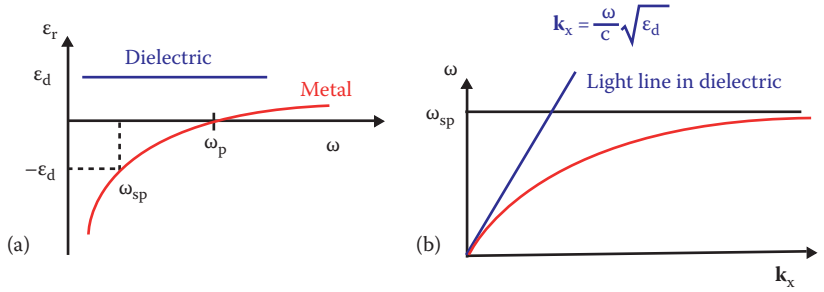


FIGURE 5.45 (a) Plot of dielectric constant for metal and dielectric. (b) Dispersion plot of surface plasmon (plot of ω, **k**).

Stimulation of surface plasmons with a purely optical field is not easy. The dispersion relation always lies at higher wave vectors than the light line ($\omega = c\mathbf{k}$) (see Figure 5.44). Because of this difference in wave vector, the plasmon field cannot efficiently couple to radiating modes. Conversely, free space optical fields cannot directly stimulate surface plasmons unless a mechanism introduces additional momentum. For example, this can be done when the light travels through a media of refractive index *n* greater than the index of refraction of the vacuum/glass. Thus, by tuning the angle of incidence of the light, the **k** vector (parallel to the surface) will match the resonant condition for SP excitation. In Figure 5.45a we show the dielectric constant of the metal (ε_m, red line) and the dielectric (blue line). At $\omega = \omega_{sp}$ we have $\varepsilon_m = -\varepsilon_d$. Figure 5.45b, is one of ω versus $\mathbf{k}_x$. At $\omega = \omega_{sp}$ (with $\varepsilon_m = -\varepsilon_d$), $\mathbf{k}_x \to \infty$, and at low frequency (where $\varepsilon_m \to \infty$) the slope of the dispersion curve approaches the light line in the dielectric, or:

$$\mathbf{k}_x \approx \frac{\omega}{c}\sqrt{\varepsilon_d} \tag{5.164}$$

In Figure 5.45b, we see that $\mathbf{k}_x$ based on Equation 5.162 must automatically be lying below the light line based on Equation 5.164.

Special geometries are required to match the surface wave vector $\mathbf{k}_{sp}$ with that of the incoming light producing it. Surface plasmons can be excited by:

Light (photons)
- Excitation from high-index medium in the Kretschmann geometry (Figure 5.46a and b)
- Coupling using grating (Figure 5.46c)
- Coupling using subwavelength scatter points (Figure 5.46d)

Electrons

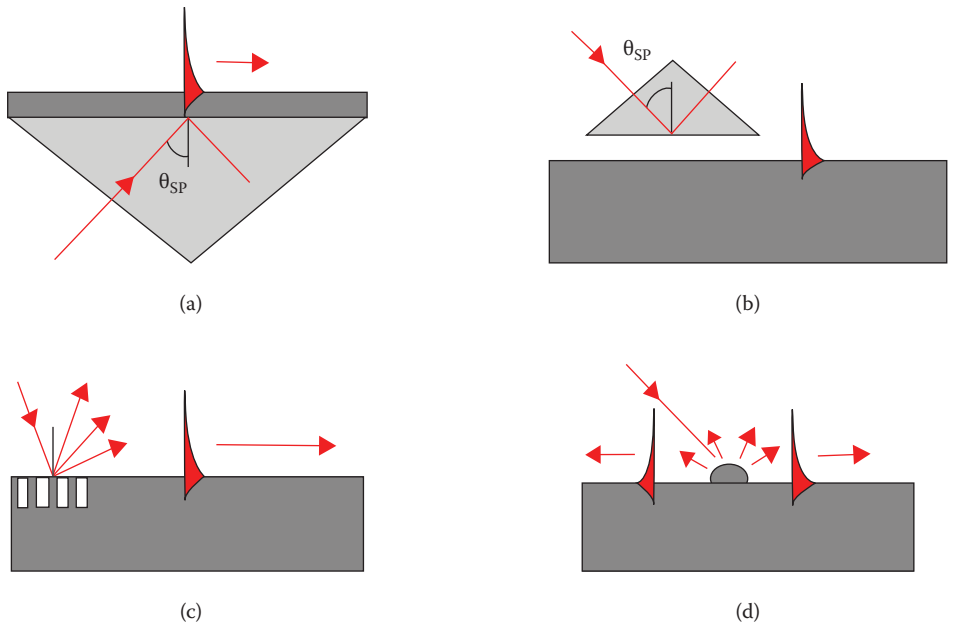


FIGURE 5.46 Excitation methods of surface plasmons. Excitation from high-index medium in Kretschmann geometry (a, b). Coupling using a grating (c). Coupling using subwavelength scatter points (d).

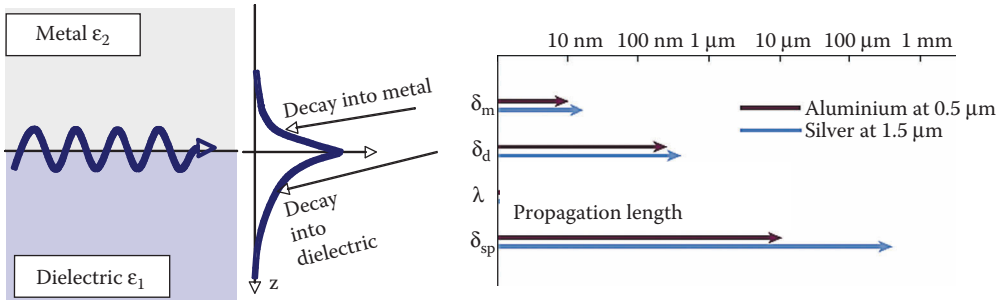


FIGURE 5.47 Decay lengths in a dielectric material (δ_d) and in a metal (δ_m), and the propagation length of the surface plasmon at resonance (δ_{sp}) for aluminum and silver.

From Equation 5.158 the penetration depth in the metal and the dielectric can be calculated:

$$\begin{aligned} k_{1z} &= -i\kappa_{1z} \\ k_{2z} &= +i\kappa_{2z} \end{aligned} \tag{5.165}$$

($\sim e^{ik_z z}$; $\kappa_z = \pm i\kappa_z$). As expected, and illustrated in Figure 5.47, the decay length in the dielectric material (δ_d) is considerably longer than in the metal (δ_m). The propagation length of the surface plasmon at resonance is the longest (δ_{sp}).

In Table 5.5 we list $\hbar\omega_{sp}$ for Au, Al, and Ag on GaN. The penetration depths into GaN are also listed.

Localized Plasmons

Whereas in semiconductor structures with reduced dimensionalities (quantum wells, quantum dots, and quantum wires) electron confinement leads to quantization of energetic states (see Chapter 3), metallic nanostructures exhibit major changes in their optical spectra because of light absorption and the creation of plasmonic waves at the surfaces of these structures. Whereas resonant coupling of light into a surface plasmon at a metal dielectric interface is difficult, coupling into surface plasmons on a metal nanoparticle is relatively easy. In the case of metal nanoparticles, surface plasmon waves are localized and are not characterized by a surface plasmon wave vector, $\mathbf{k}_{sp}$, and one refers to *localized surface plasmons*. There is no need for a special geometry to induce these localized oscillations. If the sphere is small compared to a wavelength of light, and the light has a frequency close to that of the SP, then the SP will absorb energy. The frequency of the SP depends on the dielectric function of the nanomaterial and the shape of the nanoparticle. For a gold spherical particle, the frequency is about 0.58 of the bulk plasma frequency. Thus, although the bulk plasma frequency is in the UV, the SP frequency is in the visible (close to 520 nm). Because it is possible to control the nanoparticle's surface shape, one can control the types of surface plasmons that can couple to it and propagate across it. This in turn controls the interaction of light with the surface. These effects are illustrated by the historic stained glass that adorns medieval cathedrals (Figure 5.48). In this case, the color is given by metal nanoparticles of a fixed size that interact with the optical field to give the glass its vibrant color. In modern science, these effects have been engineered for both visible light and microwave radiation. To produce optical range surface plasmon effects involves producing surfaces that have features <400 nm.

TABLE 5.5 $\hbar\omega_{sp}$ for Au, Al, and Ag on GaN and the Penetration Depths into GaN

Metal on GaN	h*ħ*sp (eV)	Penetration Depth (nm)
Au	2.2 (~560 nm)	33
Al	5 (~250 nm)	77
Ag	2.92 (~410 nm)	40

Plasmonic Applications I

As mentioned in the introduction of this section, plasmons have been considered a means of transmitting information on computer chips because plasmons can support much higher frequencies (into the 100-THz range, whereas conventional wires become very lossy in the tens of GHz). At first glance, guiding light signals with metal structures appear odd as metals are known for high optical losses, and it is obvious that fiberoptic networks are able to carry more information and with greater fidelity than electrons in copper or coaxial cable. Optical fibers are

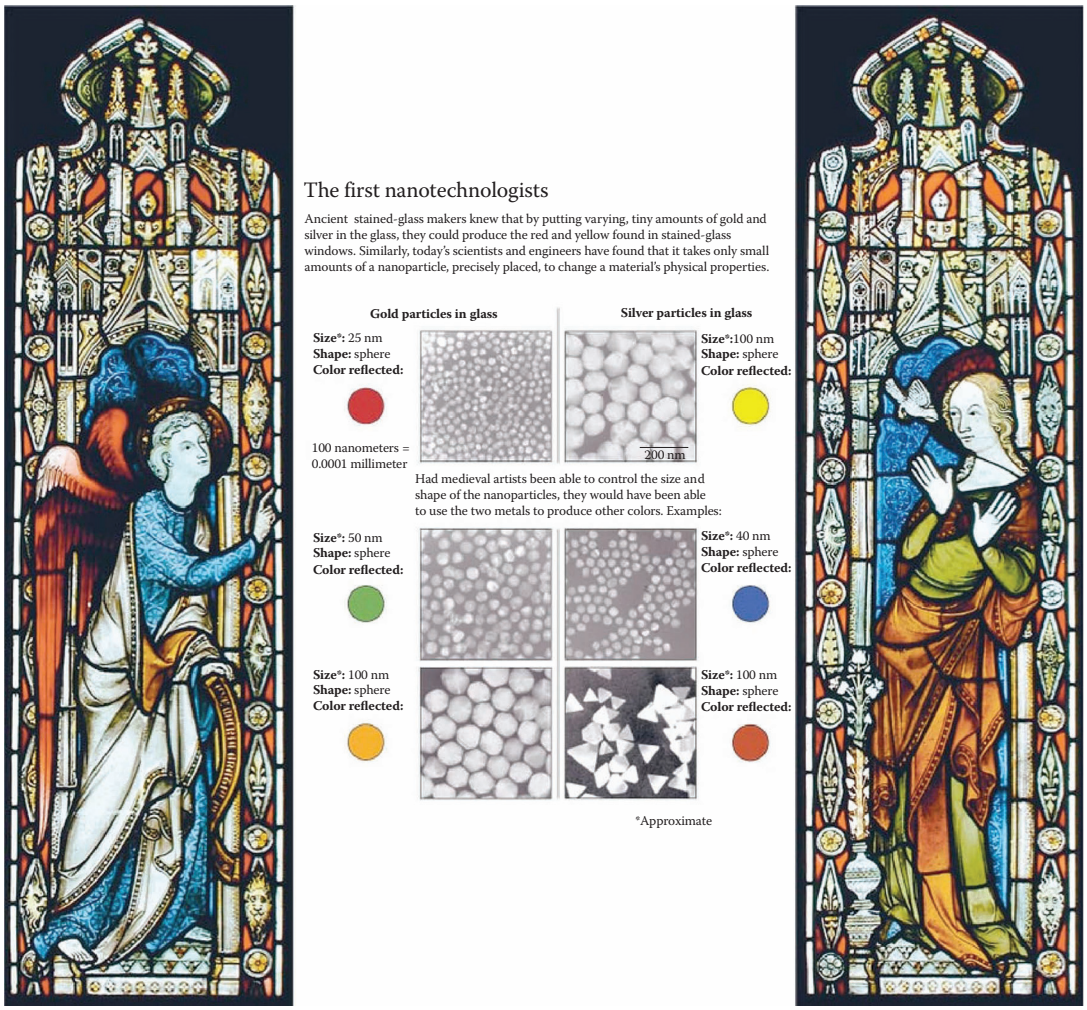


FIGURE 5.48 Medieval artisans unknowingly became nanotechnologists when they made red stained glass by mixing gold chloride into molten glass. That created tiny gold spheres, which absorbed and reflected sunlight in a way that produces a rich ruby color.

also immune to electrical interference, and they are chemically inert (glass), safe (they carry light instead of electron currents), and lightweight. They are very difficult to tap into (safe for sensitive data transfer), and finally photons travel much faster than electrons: $c = 3.0 \times 10^8$ m/s (vacuum) versus v_F (Fermi speed) = 1.57×10^6 m/s (copper wire). No wonder many scientists predicted that photonic devices would someday replace electronic ones. The reason why this has not happened is that the performance and size of photonic devices are diffraction limited so that the width of an optical fiber needs to be at least twice that of the light's wavelength. In case we use near-infrared wavelengths of about 1500 nm, we end up with photonic structures more than 15 times larger than Si-based electronic devices with features smaller than 65 nm. In other words, photonics, although high speed and high bandwidth, are harder to miniaturize than electronics because of the diffraction of light in the former (see Figure 5.49b). Moreover, in the case of a fiberoptic, one also has to deal with bending losses (see Figure 5.49a).

Plasmonics offers a way out of this dilemma as using metal to guide light avoids the optical fiber bending problem and the diffraction problem of light. Envision alternating charge densities on metal surfaces—surface plasmons—as the equivalent of an AC current traveling along an ordinary wire with the important difference that the frequency of the optical signal is more than 400,000 GHz, whereas the AC current is 60 Hz. This makes for plasmon circuits that can carry much more data than electronic ones.

Many photonic applications of surface plasmons and localized surface plasmons derive from the local

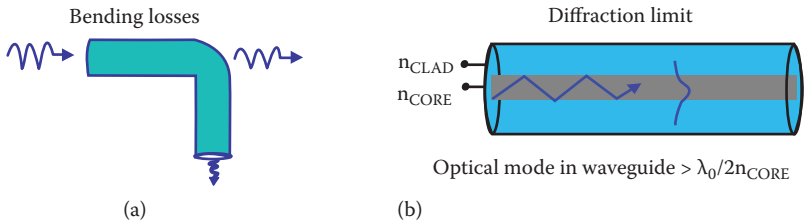


FIGURE 5.49 Problems with photonic devices. Bending (a) and diffraction losses (b).

electrical field enhancement under plasmon resonance conditions. Plasmon resonances, associated with large time-varying polarizations, can lead to extremely strong local electric fields. The local electric fields resulting from these plasmon effects are evanescent, and so they have very short ranges. The field enhancement in the small volume of the medium surrounding the metal nanoparticles or metal/dielectric interface has been used, for example, for guiding light through dimensions smaller than the wavelength of the incident light in plasmonic guiding. Plasmonic guiding in lithography is described further below and in Volume II, Chapter 2 on next-generation lithographies (NGLs) where plasmonics is proposed as a new high-resolution lithography. In Volume III, Chapter 6 on metrology and further below we learn that plasmonics also enables aperture-less near-field scanning optical microscopes (NSOMs). One other extremely interesting and important phenomenon related to plasmons is surface-enhanced Raman spectroscopy (SERS) that can be used to detect single molecules. In Raman spectroscopy the light scattered inelastically from a molecule provides information about the molecule's quantum states (see Figures 5.104 and 5.105). Discovered around 1974, it was found that a Raman signal from pyridine on a silver film is ~10^6 times larger than what one would expect based on the Raman signal for the same molecule in solution. The record enhancement inferred for SERS is now ~10^{14} (!!) if the molecules are attached to nanometer-sized metal surfaces.

Beyond Maxwell

The Lorentz Transformation

Hendrik Antoon Lorentz (1853–1928) (Figure 5.50), born in Arnhem, the Netherlands, refined Maxwell's electromagnetic theory and was the joint winner (with Pieter Zeeman) of the Nobel Prize for Physics in 1902 for his theory of electromagnetic radiation, which, confirmed by findings of Zeeman, gave rise to Albert Einstein's special theory of relativity. Lorentz proposed that light waves were caused by oscillations of an electric charge in the atom. What is remarkable is that he conceived this idea even before the existence of electrons was proved. Lorentz is also the man of the Lorentz force we introduced in Chapter 3 (Equation 3.64). He did get it a bit wrong with his proposal that the motion of a body through the ether resulted in a tiny contraction of all physical bodies in the direction of the motion—called the *Lorentz-FitzGerald contraction*, or *aether squeeze*. But the so-called *Lorentz transformation* got Einstein's attention, and he chose it over the Galilean transformation, doing away with the notion of an ether once and for all.

The speed of light c appears as a fundamental constant in Maxwell's equations, and one might ask against what relative frame of reference is this measured? Maxwell viewed light as waves moving in the medium aether. Because it was assumed that a substance filled all space and provided a medium (aether) through which light waves traveled, it was postulated that the velocity of light c should be measured with respect to the rest frame of the aether.

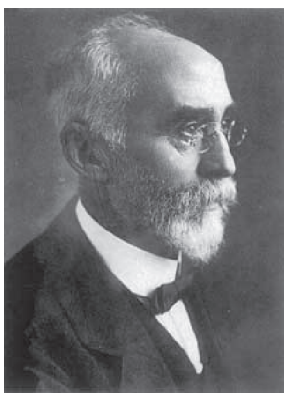


FIGURE 5.50 Hendrik Antoon Lorentz (1853–1928).

Relativity tells us how to relate measurements in different frames, and its principle was first stated by Galileo: no experiment inside a steadily moving ship will show that it is moving; only by looking outside can one detect motion—i.e., relative motion.

This principle is stated mathematically in the Galilean transformation for going from one reference frame to another. In the case of a reference frame moving with a speed v in the x-direction with respect to another reference frame, this yields:

$$\begin{aligned} x' &= x - vt \\ y' &= y \\ z' &= z \\ t' &= t \end{aligned} \qquad (5.166)$$

Thus, in Galilean relativity one has a simple velocity addition law (Figure 5.51).

$$v = v' + v_b \qquad (5.167)$$

Newton's laws were found to remain the same under the Galilean transformation. As a consequence, it is impossible to tell, by making mechanical experiments, whether a system is moving, i.e., they are invariant to a Galilean transformation. Newton articulated the idea that there is no such thing as "absolute motion" or reference frame. All "inertial" frames are equivalent, and the laws of physics are the same in all reference frames.

But there was trouble ahead: the Galilean transformation did not apply to Maxwell's equations. In other words, Maxwell's equations do not remain in the same form under the Galilean transformation (they are NOT invariant to a Galilean transformation). As a consequence, in a moving spaceship electrical and optical phenomena should be different from those in a stationary spaceship, and one should be able to use these optical phenomena to determine the absolute speed of the spaceship. Soon it was discovered that when the speed of light was involved the Galilean transformation was not applicable and that another transformation between reference frames, the Lorentz transformation, had to be invoked.

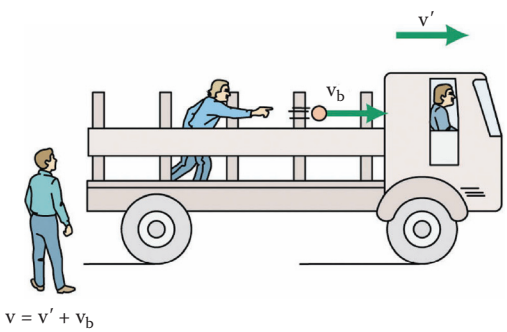


FIGURE 5.51 Illustration of Galilean addition of velocities.

The realization that there was something special about the speed of light came from the Michelson-Morley experiment. The Michelson-Morley experiment (Figure 5.52) was yet another attempt to detect the aether. If aether existed, then the speed of light should be dependent on it. If the speed of light were constant with respect to the aether, then the measured speed of light in perpendicular directions should be different because of the movement of the earth through the aether. In 1887, A.A. Michelson and E.W. Morley refined an experiment that Michelson had performed a few years earlier. They split a beam of light in two, sending the two halves in different directions. Each ray of light was reflected back and the two beams combined again. Thus, if the earth is traveling at v with the aether, then light

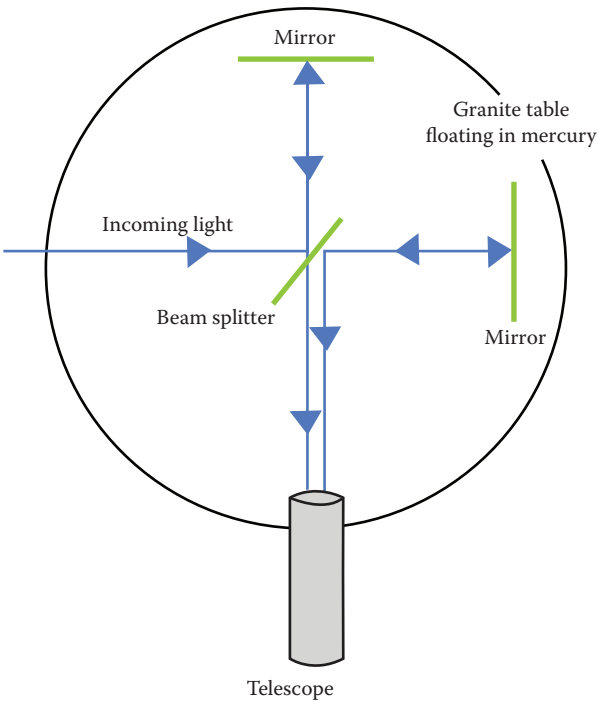


FIGURE 5.52 Michelson-Morley experiment. In 1887 Michelson and Morley set up this apparatus in which a light beam is split into two perpendicular beams and then recombined. By observing interference between the two beams, they showed that the speed of light was constant.

should have a velocity of $c + v$ (with respect to the aether). In the perpendicular direction, the velocity would just be c. By observing the interference patterns, they were able to determine the difference in the speed of light in the two perpendicular directions. Their result was astonishing: zero, or light speed, did not combine with Earth's velocity. The Michelson-Morley experiment implied that the velocity of light in vacuum, c, is independent of the frame of reference. Their result violated the Galilean law of addition of velocities (Equation 5.167).

However, the scientific community was not ready to dispense with the idea of an aether; to explain the Michelson-Morley experiment, Lorentz and the Irish physicist George Francis FitzGerald independently introduced the notion that motion through the aether resulted in a tiny contraction of all physical bodies in the direction of motion (the aether squeeze):

$$L_{\text{parallel}} = L_0\sqrt{1 - v^2/c^2} \tag{5.168}$$

In this way, the arm of the Michelson interferometer that pointed in the direction of Earth's motion (L_{parallel}) was supposed to be shorter than the arm at right angles. This explained the apparent absence of the relative motion of the Earth with respect to the aether, as indicated by the experiments of Michelson and Morley, and also paved the way for Einstein's special theory of relativity (see next section). In 1904, still believing in the concept of the aether, Lorentz thus replaced the Galilean transformation of reference frame with the following transformations:

$$\begin{aligned} x' &= \frac{x - vt}{\sqrt{1 - v^2/c^2}} \\ y' &= y \\ z' &= z \\ t' &= \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}} \end{aligned} \tag{5.169}$$

This transformation rendered Maxwell's equations reference frame-independent. The Lorentz transformation will result in a stationary observer recording an effect equivalent to the Lorentz contraction when observing an object in uniform motion relative to his system of coordinates. However, Einstein showed that this effect is not caused by the actual deformation of the body in question, as Lorentz had originally supposed, but by a change in the way space and time are measured. The consequences of the Lorentz transformation are mind-boggling: they imply that time is not universal, the relativity of simultaneous events, Fitzgerald-Lorentz contraction, and time dilation.

Einstein's Special and General Relativity

Special Relativity

Thus, science faced the following mutually exclusive choices: either continue to accept Newton's law and Galilean invariance or adapt Maxwell's equations and Lorentz invariance. Einstein, in 1905, introduced the theory of special relativity in which, following a suggestion made by Poincaré, he picked the Lorentz transformation instead of the Galilean transformation as the correct one. He proposed that "all physical laws should remain unchanged under Lorentz transformation."

Einstein's two postulates of special relativity are:

Postulate 1—The principle of relativity. The laws of nature are the same in all inertial frames of reference. Implications are that there is no way to detect absolute motion and that there is no preferred frame of reference (no aether).

Postulate 2—The universal speed of light. The speed of light, c, in a vacuum is the same in all inertial frames of reference. This implies that c does not depend on the source or the observer's velocity.

As a consequence of the second postulate, time appears dilated. Time-measuring instruments (clocks) must apparently slow down:

$$t' = \frac{t_0}{\sqrt{1 - v^2/c^2}} \tag{5.170}$$

and lengths seem to suffer (Lorentz) contraction for moving objects:

$$x' = \frac{x - vt}{\sqrt{1 - v^2/c^2}} \tag{5.171}$$

And finally, two events that are simultaneous in one inertial system are not, in general, simultaneous in another:

$$t_2' - t_1' = \frac{u(x_1 - x_2)/c^2}{\sqrt{1 - u^2/c^2}} \tag{5.172}$$

Einstein convinced the world that the speed of light is the same in any direction, which explains the null result of Michelson and Morley, and he also showed that there is no such thing as aether (nor any need for it): light is perfectly happy traveling in a vacuum.

Instead of the Galilean velocity addition (Equation 5.167) with $E = m_0c^2 + \frac{mv^2}{2}$, where m_0c^2 is for the stationary situation, Einstein suggested the world's most famous equation:

$$E = mc^2 \tag{5.173}$$

When applying a force to a stationary particle, we supply kinetic energy to the particle and it starts to move. From Equation 5.173, the total energy E and mass m must go up because c is constant, and according to the conservation of energy:

$$\frac{dE}{dt} = \mathbf{F} \cdot d\mathbf{v} \tag{5.174}$$

From Newton's second law:

$$\mathbf{F} = \frac{d(m\mathbf{v})}{dt} \tag{5.175}$$

and:

$$\frac{d(mc^2)}{dt} = \mathbf{v}\frac{d(m\mathbf{v})}{dt} \tag{5.176}$$

Multiplying both sides of Equation 5.176 with $2m$ gives us:

$$c^2 2m\frac{dm}{dt} = 2m\mathbf{v}\frac{d(m\mathbf{v})}{dt} \text{ or } c^2\frac{d(m^2)}{dt} = \frac{d(m^2\mathbf{v}^2)}{dt} \tag{5.177}$$

Integration of both sides results in:

$$m^2c^2 = m^2v^2 + C \tag{5.178}$$

The constant C can be obtained by substitution $v = 0$ at which $m = m_0$, or:

$$m^2c^2 = m^2v^2 + m_0^2c^2 \tag{5.179}$$

or:

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{5.180}$$

Thus, Newton's laws are the low-velocity limits of Einstein's special relativity theory. The Einstein modification of $p = mv$ is then given as:

$$p = \frac{m_0 v}{\sqrt{1 - v^2/c^2}} \tag{5.181}$$

What happens now if a constant force acts on a body? In Newtonian mechanics the body keeps picking up speed, whereas in relativity the body keeps picking up momentum, not speed. This implies that m must be a function of speed v as evident from Equation 5.180.

General Relativity

Einstein's magnum opus is the general relativity theory. Einstein recognized that his special relativity theory was violated by Newton's theory of gravity because the latter implied that gravity effects were felt instantaneously, i.e., faster than the speed of light! The general relativity theory is the geometric theory of gravitation, incorporating and extending the theory of special relativity to accelerated frames of reference and introducing the principle that gravitational and inertial forces are equivalent. From this theory, concentrations of matter or energy cause space-time to curve, and this curvature deflects the trajectories of particles, just as observed for particles in a gravitational field. Einstein was able to calculate from this new theory that gravity disturbances move at precisely the speed of light. According to Einstein, the agent of gravity then is the fabric of the cosmos, and in the presence of a massive object like the sun that fabric becomes distorted. An artistic rendition of this concept is shown in Figure 5.53.

Summarizing, all macroscopic electromagnetic phenomena can be explained in terms of Maxwell's equations, the continuity equation, and the Lorentz force equation. In Chapter 3 we saw that a description of light as purely consisting of waves fails to explain a large body of experimental facts such as the photoelectric effect and blackbody radiation. Here is where quantum mechanics comes in, with as one of its basic tenets that light and material

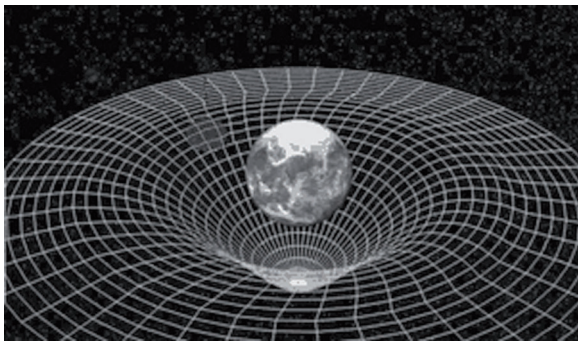


FIGURE 5.53 The fabric of the cosmos is distorted by a massive object. The Earth creates a deep hollow in the fabric of space-time.

objects both display wave and particle behavior. Special relativity becomes manifest when things are moving very fast, and general relativity comes into play when things are very massive and the warps in space and time are correspondingly large. The general relativity theory is inherently classical, and this bothers many physicists as it makes for a difficult marriage with quantum physics describing the behavior of particles and forces in the atomic and subatomic realm. A quantum theory of gravity is a holy grail for physicists. A promising approach to a quantum theory of gravity is string theory, which has been explored since the 1970s.

Two excellent books to read on the quest for the ultimate theory are Brian Greene's *The Elegant Universe*[7] and Barry Parker's *Search for a Super-Theory.*[8] A list of recommended books to read on quantum mechanics was presented in the Chapter 3 Further Reading list.

Optical Properties of Materials

Introduction

Now that we know how reflectance R, transmittance T, and absorbance A relate to refractive index n and extinction coefficient κ, as well as how these relate to electrical susceptibility $\chi(\omega)$, we must yet establish a physical model for $\chi(\omega)$. Optical properties of materials are determined by two main factors: first, how easy is it to move charges using an EM wave [$\mathbf{P} = \varepsilon_0 \chi(\omega)\mathbf{E}(\omega)$], and, second, how many charges are available (n)? Based on these criteria, optical behavior of materials can crudely be divided in:

- Atoms/ions, where optical transitions take place between electron orbits and lead to absorption lines, ionization energies.
- Molecules, where optical transitions again take place between electron orbits, but where one also has vibrations (in the meV range) and the optical excitation of those molecular vibrations (IR and Raman).
- Metals, with large concentration of free electrons, exhibit high reflectivity and strong absorption.
- Insulators, where electrons are strongly bound to atoms ($E_{gap} > \sim 4$ eV), resulting in a low absorption and refraction.
- Semiconductors, where electrons are "somewhat bound" ($E_{gap} < \sim 4$ eV) and are transparent in the IR and absorb in the visible and UV, and light can generate excitons (bound electron-hole pairs).

In this section we establish the optical dispersion of the dielectric constant [$\varepsilon(\omega)$], the refractive index [$(n)\omega$)], the absorption [$\alpha(\omega)$], and the reflectance [$R(\omega)$] of these types of materials based on a microscopic physical model, either the Drude model or the more sophisticated Drude-Lorentz model. First we define the polarization **P** of a material and briefly reiterate the salient points of the Drude model, encountered first in Chapter 3, and then, to accommodate both free and bound electrons (electrons bound to atom core) in a material, we introduce the Drude-Lorentz physical model for $\chi(\omega)$; this allows us to cover optical properties not only of all metals but also of insulators and semiconductors.

The Microscopic Origin of the Frequency Dependency of χ, ε, and n in Solids-Polarization **P**

The polarization **P** of a medium is the induced dipole moment per unit volume (C/m^3) (Figure 5.54). It is not the first time we touch on polarization **P** of matter in this book. We broached the topic first when introducing the Drude model, where we discussed the polarization of the free electron gas (FEG) with respect to the immobile ions in a metal. Polarization of the medium comes about as a result of small displacements of charges in an electrical field. For

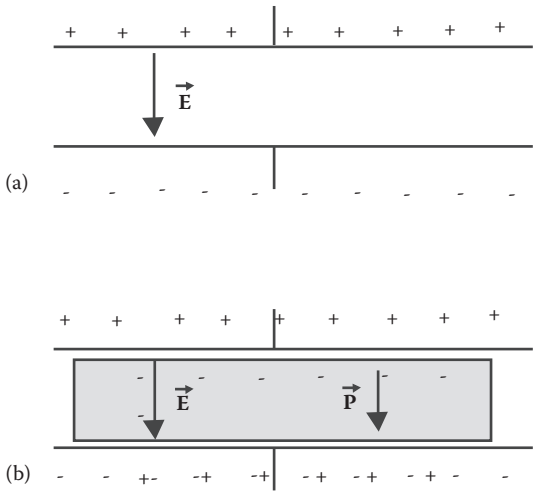


FIGURE 5.54 Polarization **P** of a medium. (a) P = 0 (no dielectric). (b) **P** ≠ 0 (dielectric). **P** points from – to +.

example, polar molecules align in a field, and nonpolar molecules have their electron cloud distorted. It is easy to see how at optical frequencies electrons might still swing along with a fast-changing electromagnetic field but that nuclei might be too heavy to follow, causing an electron cloud polarization. Different polarization mechanisms exhibit different dynamic behavior in time, and as a consequence, electromagnetic radiation absorption occurs at different frequencies. Polarization mechanisms include space-charge polarization, orientation polarization, ionic polarization, and electronic polarization (see The Drude-Lorentz Model for an example of the latter type).

As demonstrated in Chapter 3, it is the polarization $\mathbf{P}(\omega,t) = \chi_e\varepsilon_0\mathbf{E}(\omega,t)$ (Equation 3.36) of a material that is ultimately responsible for the frequency dependency of the dielectric constant ε and of the refractive index *n* because all these parameters are linked. Thus, a general and realistic relation between **P** and **E** includes the fact that the relation between **P** and **E** is dynamic:

$$\mathbf{P}(\mathbf{k},\omega) = \chi_e(\mathbf{k},\omega)\varepsilon_0\mathbf{E}(\mathbf{k},\omega) \qquad (5.182)$$

The polarization of a material along the applied field direction results in an opposing response field and consequently a reduced total field. In other words, some of the external field applied over a material disappears and is replaced by polarization **P**. For a linear, homogeneous, and isotropic medium, the electric susceptibility, χ_e, is a scalar, but for an anisotropic material the electric susceptibility is a tensor, and the polarization **P** can point in a different direction from the field **E**; **D** and **E** are also not necessarily collinear. In this case, Equation 3.36 is modified to:

$$\mathbf{P}_i = \sum_j \chi_{ij}\varepsilon_0\mathbf{E}_j \qquad (5.183)$$

From the expression linking the dielectric constant with the polarization **P**: $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P} = (1 + \chi_e)\varepsilon_0\mathbf{E} = \varepsilon(\omega)\mathbf{E}$ (Equation 3.37), we can establish the link between the relative permittivity $\varepsilon_r(\omega)$ (also called *dielectric constant*) and the electric susceptibility tensor as:

$$\varepsilon_r = 1 + \chi_e(\omega) \qquad (5.184)$$

Because $n(\omega)^2 = \varepsilon\mu$ (Equation 5.22), the refractive index *n* can be derived from χ_e, assuming that μ = 1, as:

$$n = \sqrt{1 + \chi_e(\omega)} \qquad (5.185)$$

The Drude-Lorentz Model

Drude Model Recap

In the Drude physical model, we learned in Chapter 3, the only conduction mechanism is that resulting from a free electron gas (FEG). For a time harmonic electrical field $[\mathbf{E}(t) = \mathbf{E}_m e^{-\omega t}]$, we obtained the displacement *x* of the electron gas $x(t) = \dfrac{e\mathbf{E}(\omega)}{m_e(\omega^2 + i\gamma\omega)}$ (Equation 3.34) from the equation for the momentum per electron: $m_e\dfrac{d^2x}{dt^2} + m_e\gamma\dfrac{dx}{dt} = -e\mathbf{E}(t) = -e\mathbf{E}_{(\omega)}e^{-i\omega T}$ (Equation 3.33). The macroscopic polarization **P** was then calculated as the product of this displacement *x*(*t*), and the electron density ne (**P** = –nex), so that for the electric displacement we obtained $\mathbf{D} = \varepsilon(\omega,\mathbf{k})\mathbf{E} = \varepsilon_0\mathbf{E} + \mathbf{P} = \varepsilon_0\mathbf{E} - \dfrac{ne^2\mathbf{E}}{m_e(\omega^2 + i\gamma\omega)}$ (Equation 3.39). When electrons are free, as in a metal, the damping factor γ in this expression comes about because of electron scattering (γ = 1/τ, where τ is the electron scattering time). From the expression for the displacement field **D**, we then derived the complex dielectric constant as $\varepsilon(\omega,\mathbf{k}) = \varepsilon_0\left[1 - \dfrac{\omega_p^2}{(\omega^2 + i\gamma\omega)}\right]$

(Equation 3.40), where ω_p is the plasma frequency, defined as $\omega_p = \sqrt{\frac{ne^2}{m_e\varepsilon_0}}$ (Equation 3.41). If we consider a highly conductive system, where the damping factor γ tends to zero, the real component of the dielectric constant $\varepsilon_r'(\omega,\mathbf{k}) \approx 1 - \frac{\omega_p^2}{\omega^2}$ (Equation 3.44) dominates because the imaginary component $\varepsilon_r''(\omega,\mathbf{k}) \approx \frac{\omega_p^2\gamma}{\omega^3}$ (Equation 3.45) tends to zero. Based on Equation 5.118, the complex refractive index in this case can be written as $n_c(\omega) = \sqrt{\varepsilon_r'(\omega)} = \sqrt{1 - \frac{\omega_p^2}{\omega^2}}$. The Drude model predicts a monotonous decrease of $\varepsilon_r'(\omega,\mathbf{k})$ for decreasing frequency, and experiments confirm that $\varepsilon_r'(\omega,\mathbf{k})$ becomes zero when $\omega = \omega_p$. When $\varepsilon_r'(\omega)$ becomes zero at $\omega = \omega_p$, the material can support free longitudinal collective modes for which all electrons oscillate in phase. Such a collective charge oscillation is called a *plasmon*. Finally, in a metal where both bound electrons $[\varepsilon_B(\omega)]$ and conduction electrons $[\varepsilon_{r(\omega)}]$ contribute, an effective dielectric constant $\varepsilon_{Eff(\omega)}$ was defined as $\varepsilon_{Eff}(\omega) = \varepsilon_B(\omega) + \varepsilon_r(\omega)$ (Equation 3.48). As soon as there are bound charges involved, as in some metals and in all dielectrics, the Drude model fails, and the Drude-Lorentz model needs to be invoked.

The Drude-Lorentz Model

Introduction The Drude model advanced a first physical concept for $\chi(\omega)$, but as we just saw, it only accounts for the free electron contributions to the electric susceptibility. The Lorentz model takes into account the contribution stemming from bound electrons. In this physical microscopic model, atoms and molecules are replaced by a set of oscillators, resonant at some frequency:

$$\omega_0 = \sqrt{\frac{k}{m}} \tag{5.186}$$

These oscillators come with a restoring force $\mathbf{F}_r$ where:

$$\mathbf{F}_r = -k\mathbf{r} = -m\omega_0^2\mathbf{r} \tag{5.187}$$

with $\mathbf{r}$ being the oscillator's displacement. The model does not specify the nature of the oscillating dipole and makes only very basic assumptions about the restoring forces involved. For these reasons it applies to a wide variety of situations. For electronic polarization of atoms in the visible and near-infrared frequencies (10^{14}–10^{15} Hz), the mass m in the above equations is the mass of the electron. For molecular and lattice vibrations in the mid- and far-infrared frequencies (10^{12}–10^{13} Hz), m is the mass of the nucleus. Depending on the exact atom, the nucleus is three or four orders of magnitude heavier than the electron, and the interatomic force is similar to, or less than, the electron-nucleus force (k); thus, using $\omega_0 = \sqrt{\frac{k}{m}}$ one expects nuclear vibrations to be a factor of 50–1000 times slower than electronic resonances. As a consequence, it is infrared rather than visible light that strongly interacts with nuclear vibrations, and molecular rotations interact with even longer wavelengths in the far-infrared. Free electrons in metals and semiconductors also can be described by the Drude-Lorentz model as a special case. Because there is no restoring force ($\mathbf{F}_r$) pulling charges back to their equilibrium positions, the Drude-Lorentz model simplifies back to the Drude model.

Atomic Polarizability $\chi_{a,e}$ Imagine a nucleus and electrons connected by springs with spring constant k as illustrated in Figure 5.55. For electron oscillations bound to an atom, a restoring force related to the strength of the bond of the electrons to the ion cores is introduced in the electron momentum equation. The restoring force is assumed to be linear in the displacement $\mathbf{r}$, which remains true for sufficiently small amplitudes (Hooke's

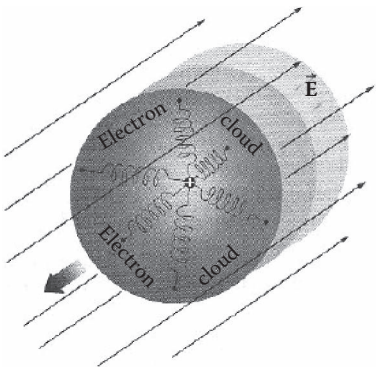


FIGURE 5.55 Oscillator model of electrons bound to an atom. The atoms and molecules are replaced by a set of harmonically bound electron oscillators, resonant at some frequency ω_0.

law). The damping force or friction loss, just as in the Drude model, is given by $\mathbf{F}_r = m_e\gamma\frac{d\mathbf{r}}{dt}$. This damping force is assumed to be linearly related to the charge velocity, and γ (in units of s^{-1}) is the damping coefficient. The resulting equation of motion is given as:

$$m_e\frac{d^2\mathbf{r}(\omega)}{dt^2} + m_e\gamma\frac{d\mathbf{r}(\omega)}{dt} + m_e\omega_0^2\mathbf{r}(\omega) = -e\mathbf{E}(t) = -e\mathbf{E}_m e^{-\omega t} \quad (5.188)$$

with a solution for the oscillation's amplitude given by:

$$\mathbf{r}(\omega) = -\frac{e}{m_e}\frac{\mathbf{E}(\omega)}{(\omega_0^2 - \omega^2 - i\omega\gamma)} \quad (5.189)$$

This expression is that of a driven damped oscillator where the motion matches the drive frequency. The maximum amplitude $\mathbf{r}(\omega)$ is reached when the denominator is at a minimum, i.e., when $\omega = \omega_0$ (resonance). For excitation at resonance ($\omega = \omega_0$), $\mathbf{r}(\omega)$ is entirely imaginary, and there is a 90° phase difference between $\mathbf{E}(\omega)$ and $\mathbf{r}(\omega)$. At high frequencies ($\omega >> \omega_0$) the amplitude $\mathbf{r}(\omega)$ vanishes, and at low frequencies ($\omega << \omega_0$) a finite amplitude is obtained.

For free electrons in a metal there is no restoring force ($\mathbf{F}_r = -k\mathbf{r} = -m\omega_0^2\mathbf{r} = 0$), and this expression reduces back to the Drude AC model we derived for the case of free electrons in a metal in Chapter 3 (see Equation 3.34).

In a driven harmonic oscillator, amplitude and phase depend on frequency as sketched in Figure 5.56. At low frequency, one obtains a medium amplitude, and the displacement is in phase with the driving force. At resonance, a large amplitude is observed, and the displacement is 90° out of phase with the driving force. Finally, at high frequency, the amplitude vanishes, and displacement is in antiphase with the driving force. Note that it is the loss term $i\omega\gamma$ in Equation 5.189 that is responsible for the phase shift.

Based on Equation 5.182, we can define the polarization of a single atom as the charge (e) times the separation [$\mathbf{r}(\omega)$] or $\mathbf{P}_a = -e\mathbf{r}(\omega)$:

$$\mathbf{P}_a = \chi_{a,e}\varepsilon_0\mathbf{E}(\omega) \quad (5.190)$$

The atom susceptibility or atomic polarizability $\chi_{a,e}$ is a proportionality constant relating the electrical field $\mathbf{E}$ to the induced dielectric atom polarization $\mathbf{P}_a$:

$$\chi_{a,e} = \frac{\mathbf{P}_a}{\varepsilon_0\mathbf{E}(\omega)} = \frac{e^2}{m_e}\frac{1}{(\omega_0^2 - \omega^2 - i\omega\gamma)} \quad (5.191)$$

(Resonance frequency) (Damping term)

This expression applies to a single atom only. The dielectric constant of a medium depends on the manner in which the atoms are assembled. Let N be

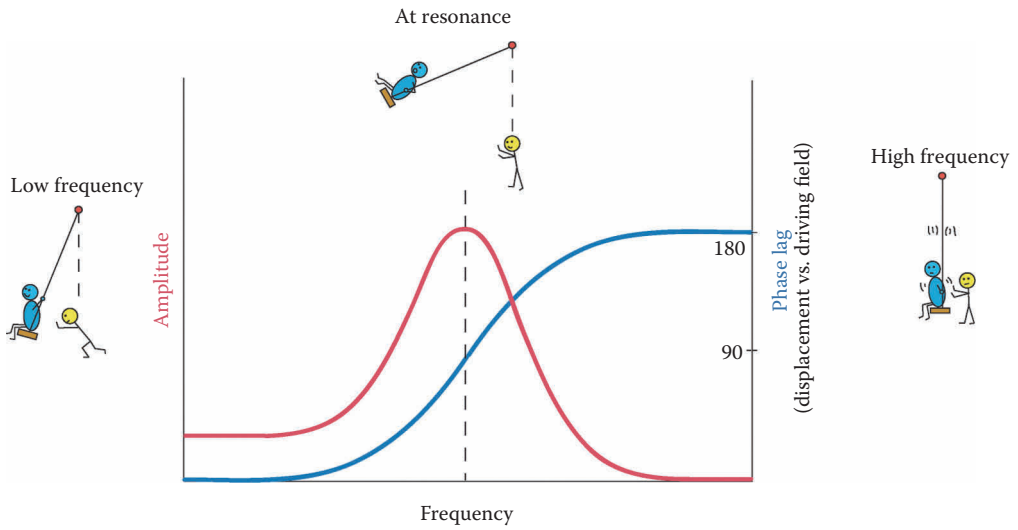


FIGURE 5.56 Amplitude and phase depend on frequency in a driven harmonic oscillator.

the number of atoms per unit volume. The polarization **P** of an assembly of atoms can be derived by summing over all the atom resonators N in a volume V and then dividing by V:

$$\mathbf{P}(\omega) = \frac{1}{V}\sum_{i=1}^{i=NV} \mathbf{P}_{a,i}(\omega) = N\langle \mathbf{P}_a(\omega)\rangle \qquad (5.192)$$

where NV is the total number of atoms and $<\mathbf{P}_a(\omega)>$ is the average dipole moment per atom. Additionally, because $\mathbf{P}(\omega) = \varepsilon_0 \chi_e \mathbf{E}(\omega)$:

$$\chi_e = \left(\frac{Ne^2}{\varepsilon_0 m_e}\right)\frac{1}{(\omega_0^2 - \omega^2 - i\omega\gamma)} \qquad (5.193)$$

where the prefactor is related to the plasma frequency $\omega_p = \dfrac{ne^2}{m_e\varepsilon_0}$ (Equation 3.41) and the density of electrons n is replaced by the density of atoms N. The assumptions here are a linear restoring force linear in **r**, a damping force linear in $d\mathbf{r}/dt$, averaged macroscopic and local fields are equal, and the dipoles are all free to point in the direction of the applied field (isotropic). Introducing the plasma frequency ω_p, we rewrite Equation 5.193 as:

$$\chi_e = \frac{\omega_p^2}{(\omega_0^2 - \omega^2 - i\omega\gamma)} \qquad (5.194)$$

For the complex refractive index n, we derived earlier that $n_c^2(\omega) = 1 + \chi_e(\omega)$ (Equation 5.113), or based on Equation 5.118:

$$\varepsilon_r = \varepsilon' + i\varepsilon'' = 1 + \chi' + i\chi'' = 1 + \frac{\omega_p^2}{(\omega_0^2 - \omega^2 - i\omega\gamma)} \qquad (5.195)$$

It follows that the real part of the complex dielectric constant is then given as:

$$\varepsilon' = 1 + \chi'(\omega) = 1 + \frac{\omega_p^2(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2} \qquad (5.196)$$

and the imaginary part equals:

$$\varepsilon'' = \chi''(\omega) = \frac{\omega_p^2\gamma\omega}{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2} \qquad (5.197)$$

These real and imaginary parts of the dielectric function as a function of ω are shown in Figure 5.57a. At very high frequencies the real part of the dielectric constant approaches 1 as χ' approaches zero. If we deal with gas molecules, the dielectric

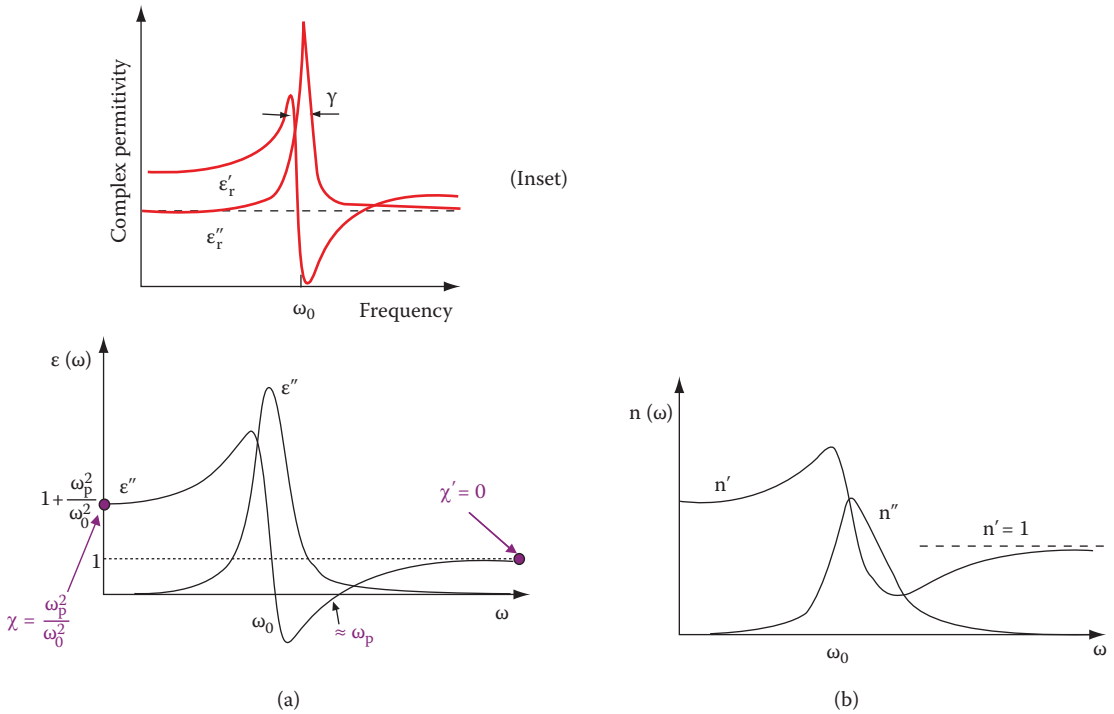


FIGURE 5.57 The real and imaginary parts of a dielectric function as a function of ω (a). The real (n) and imaginary (κ) parts of the complex refractive index n_c as a function of ω (b). (Inset) The full width at half maximum (FWHM) of the peak in ε″ is given by γ.

constant [ε(ω,k)] at those frequencies approaches ε_0 [one multiplies the relative dielectric constant $\varepsilon_r(\omega,k)$ with ε_0 to obtain ε(ω,k)]. In other words, the material becomes as transparent as a vacuum. With the molecule oscillators embedded in a solid, ε′ approaches ε_∞ at the highest frequencies (e.g., ε_∞ = 10.9 for GaAs). At very low frequencies, we get the full contribution of χ′ to the dielectric constant, i.e., $\varepsilon' = 1 + \chi'(\omega) = 1 + \frac{\omega_p^2}{\omega_0^2}$, and reach the so-called *relative static permittivity* (ε_{stat}), which is the same as the relative permittivity evaluated for a frequency of zero (e.g., ε_{stat} = 12.9 for GaAs). The real part of the dielectric constant steps from ε_{stat} to ε_∞ as the resonance is traversed from low to high frequencies. At $\omega = \omega_p$, with the damping term γ very small and $\omega_p \gg \omega_0$, the real part of the dielectric constant becomes zero (ε′ = 0). Also note that ε′ drops below zero between ω_0 and ω_p. When approaching ω_0 from below, ε′ starts rising to reach a maximum at $\Delta\omega = \omega - \omega_0 = -\gamma/2$ and a minimum at $\Delta\omega = +\gamma/2$. The peak (a Lorentzian line shape) in the imaginary part of the dielectric function (ε″) occurs at $\omega = \omega_0$ and has a full width at half maximum (FWHM) of γ (see Figure 5.57, inset). The quality factor (*Q*) of the resonance $Q = \omega_0/\gamma$ is a measure of the dissipation of the system. For an isolated molecule the damping parameter can be interpreted as the inverse of the lifetime of the excited quantum state. This is consistent with the Heisenberg uncertainty principle (HUP) ($\Delta E \Delta t \approx \frac{h}{2\pi}$ or $\Delta t \approx \frac{h}{2\pi . h \Delta\nu} \approx \frac{1}{2\pi\Delta\nu}$; Equation 3.107). If an absorption line is dampened solely by the natural lifetime of the state, one refers to natural broadening. This is also the peak that is so characteristic for bound electrons (see the example of Ag with bound d-electrons in Figure 5.64).

If one knows the ω-dependence of the dielectric function, then the ω-dependence of the refractive index can be calculated. From the relationship between the complex refractive index and the complex dielectric constant (Equations 5.115, 5.116, and 5.117), the real and imaginary parts of the complex refractive index n_c as a function of ω were plotted in Figure 5.57b. Comparing Figure 5.57a and b, one recognizes that the real part of the refractive index *n*(ω) approximately follows the frequency dependence of $\sqrt{\varepsilon_r'}$ (see Equation 5.118) and that the extinction coefficient κ more or less follows $\varepsilon_r''(\omega)/2n$ (see Equation 5.116). This correspondence is more and more exact if $\kappa \ll n$ (i.e., for weak absorption such as in gases). At high frequencies $\omega \gg \omega_0$, *n* gradually rises, and above the highest-frequency resonance (usually in the deep UV/soft x-ray region) tends toward 1, and the group velocity v_g becomes equal to the velocity of light *c*; in other words, the light moves through this medium as if it were moving through a vacuum. In the region between ω_0 and ω_p, *n* is small over the same region where ε is negative.

At low frequencies ($\omega \ll \omega_0$) the refractive index *n* is high, which means that the group velocity $v_g < c$ because:

$$V_g = \frac{c}{n + \frac{\omega dn}{d\omega}} = \frac{c/n}{1 + \frac{\omega dn}{n d\omega}} \qquad (5.89)$$

At frequencies $\omega \sim \omega_0$, the group velocity becomes very frequency-dependent as a result of the large absorption coefficient α (~κ because $\alpha = 2k\kappa$).

Multiple Resonances in Rarified Media In the treatment above we focused our attention mainly on electrons moving against an immobile nucleus, and Equation 5.194 was derived for idealized rarified media where a set of atoms featured only one electronic resonance per atom. Any real optical medium does not just have a single resonant frequency but many, as a result of different oscillators, including different "atomic resonances" (atomic transitions, interband transitions), molecular, lattice vibrations (phonons, vibrational resonances), and oscillations of the free electron plasma (plasmon excitation, free electron absorption). So, Equation 5.194 must be further corrected because realistic atoms in rarified media may have many resonances, and these resonances occur as a result of the various possible motions of electrons (high ω) and of atoms (low ω). Mathematically all these resonances are accounted for by summing over all the resonances in the material, leading to an expression for the rarified medium susceptibility given by:

$$\chi_e(\omega) = \frac{N_j e^2}{\varepsilon_0 m_e} \sum_j \frac{1}{(\omega_{0,j}^2 - \omega^2 - i\omega\gamma_j)} \qquad (5.198)$$

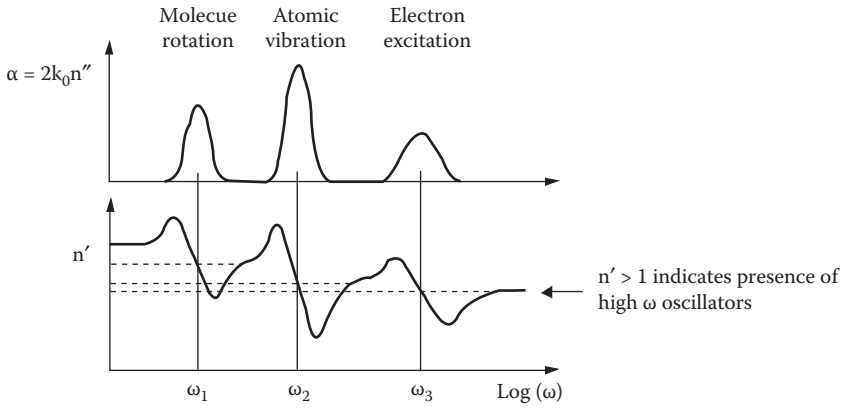


FIGURE 5.58 The real and imaginary parts of the complex refractive index n_c as a function of ω, where in the top we drew $\alpha = -2\frac{\kappa\omega}{c} = -2k\kappa$. The presence of an $n' > 1$ marks the presence of high ω oscillators.

with $\omega_{0,j}$ and γ_j the damping or decay rate of oscillator *j*. The total classical polarization **P** caused by all oscillators with resonant frequencies $\omega_{01},\omega_{02},\ldots\omega_{0j}$ and dissipation constants $\gamma_1,\gamma_2,\ldots\gamma_j$ is represented by:

$$\mathbf{P} = \chi_e(\omega) = \frac{N_j e^2}{m_e}\sum_j\left[\frac{1}{(\omega_{0,j}^2 - \omega^2 - i\omega\gamma_j)}\right]\mathbf{E} \quad (5.199)$$

and for the dielectric constant and refractive index we then find:

$$n_c(\omega)^2 = \varepsilon_r = 1 + \frac{N_j e^2}{\varepsilon_0 m_e}\sum_j\frac{1}{(\omega_{0,j}^2 - \omega^2 - i\omega\gamma_j)} \quad (5.200)$$

In Figure 5.58 we show an example of the dependence of the real and imaginary parts of the refractive index n_c on EM wave frequency for a rarified medium. A peak in the absorption (imaginary part) and a wiggle in the real part of refractive index signify each "resonance." The real part of the refractive index decreases step-wise as each resonance is traversed. At high frequencies $n' \to 1$ and $\varepsilon(\infty) \to 1$ (see Equation 5.118) because the frequency becomes too high for the material to respond. Typical spectral ranges for transitions are:

- Electronic: near infrared (about 1 μm) to UV (about 100 nm) (inner core electrons go up to x-ray energies)
- Vibrational: near infrared (about 1 μm) to long-wave infrared (about 50 μm)
- Rotational: far infrared (about 100–500 μm)

These three types of transitions are illustrated in Chapter 7 (Figure 7.83). The structure of a given molecule dictates the likelihood of the absorption of a photon to raise its energy state to an excited one.

In this classical picture each resonance has the same "strength," but when we treat the real situation properly with quantum theory each oscillator has a distinct transition matrix element, i.e., the oscillator strength f_j (see Chapter 3). This leads to a modification of Equation 5.200 as:

$$n_c(\omega)^2 = \varepsilon_r = 1 + \frac{N_j e^2}{\varepsilon_0 m_e}\sum_j\frac{f_j}{(\omega_{0,j}^2 - \omega^2 - i\omega\gamma_j)} \quad (5.201)$$

where f_j is the oscillator strength of each transition (~1 for an allowed transition) and $\omega_{0,j}$ is the resonance frequency of a transition between two states with an energy difference $\Delta E_j = \hbar\omega_{0,j}$ (Figure 5.59). The oscillator strength f_j in Equation 5.201 obeys the Thomas-Reich-Kuhn sum rule:

$$\sum_j f_j = Z \quad (5.202)$$

for an atom that has *Z* electrons. In other words, the total absorption integrated over all frequencies is only dependent on *Z*. However, usually one resonance dominates over all others.

The damping coefficient in Equation 5.201 corresponds to 1/lifetime (1/τ) of a state and can range from the kilohertz to the gigahertz (Figure 5.60). In an idealized case of purely radiative relaxation, the natural line width of an absorption or emission line, we explained above, is dictated by γ_j, which sets the

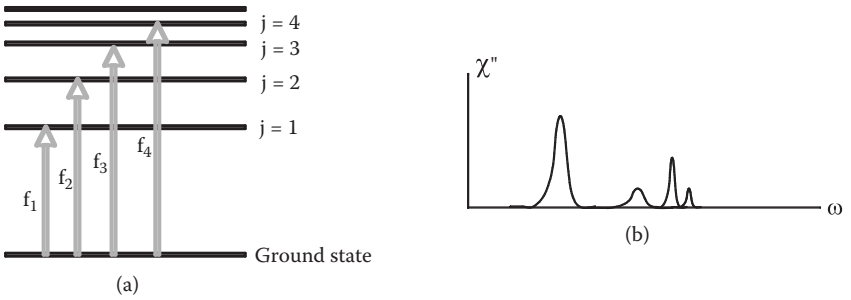


FIGURE 5.59 Real atoms have many electronic levels, so there can be many resonant transitions. Transitions between states with an energy difference $\Delta E_j = \hbar\omega_{0,j}$ (a). These transitions are reflected in absorption peaks $\alpha(\omega) \approx \frac{\omega}{c}\chi''(\omega)$, Equation 5.121 (b).

FWHM of the line (see Figure 5.57, inset). For a single simple electronic transition between two states in an atom, *fast decay* means generally high damping and leads to a broad resonance, i.e., a broad emission spectrum. Absorption associated with electronic transitions (electrons changing states) occurs in about 1 fs (10^{-15} s). The lifetime of an excited molecule depends then on how that molecule disposes of the extra energy. Because of the uncertainty principle (see Chapter 3), the more rapidly the energy is changing, the less precisely we can define the energy. Thus, long-lifetime excited states have narrow absorption peaks, and short-lifetime excited states have broad absorption peaks. In molecules, solution and solid absorption bands broaden. Electronic states in a molecule are split into many sub-bands, and in molecules each electronic state can exist in many possible vibrational and rotational states of the molecule. For an isolated molecule, the typical natural lifetime is about 10^{-8} s coming with a 5×10^{-4} cm^{-1} line width. Collisions between molecules can shorten this lifetime further, and the line width will depend on the number of collisions per second,

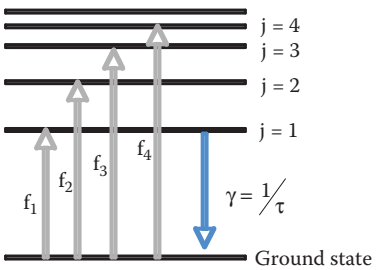


FIGURE 5.60 Idealized case—purely radiative relaxation: decay rate γ (units: s^{-1}) is related to the FWHM of an emission line: *fast decay* means a "high-damping" coefficient and a broad resonance/broad emission spectrum.

i.e., on the number density of the molecules (pressure, *P*) and the relative speed of the molecules (the square root of the temperature). A second source of line broadening in molecules is the so-called *Doppler broadening*. In the atmosphere, molecules move with velocities determined by the Maxwell-Boltzmann distribution, so molecules are in motion when they absorb, and this causes a change in the frequency of the incoming radiation as seen in the molecules' frame of reference. As a consequence, a broad emission/absorption line in a rarified medium does not necessarily mean a fast decay.

In Figure 5.61A we represent a set of calculated dispersion curves and compare a broad ($\gamma = 1$) and a narrow resonance ($\gamma = 0.3$) while keeping ω_0 and ω_p the same at 4 and 8, respectively. We compare the real and imaginary dielectric constants and the real and imaginary refractive indexes as a function of the frequency, as well as the reflectance *R* from air at normal incidence as a function of frequency. Noting that $\alpha = 2\omega\kappa/c$, we have plotted also a scaled version of α, by plotting $2\omega\kappa(\omega)$, scaled to ω_p. Note how α is skewed to higher frequencies. *R* is large in the region where *n* is small, which makes sense because in the limit of $n = 0$ then $R = 1$. In Figure 5.61B, we have colored in four distinct spectral regions, marked with *T* for transmissive ($\omega < \omega_0 - \gamma/2$), *A* for absorptive ($\omega_0 - \gamma/2 < \omega < \omega_0$), *R* for reflective ($\omega_0 + \gamma/2 < \omega < \omega_p$), and *T* for transmissive again ($\omega > \omega_p$), in other words, TART. These regions become more distinct the smaller γ and the larger ω_p. As expected for smaller damping γ, the curves are narrower and have larger maximum values. The reflectance is much higher and has sharper edges. The high *R* starts around ω_0 and falls near ω_p. This is because $\kappa >> n$ in

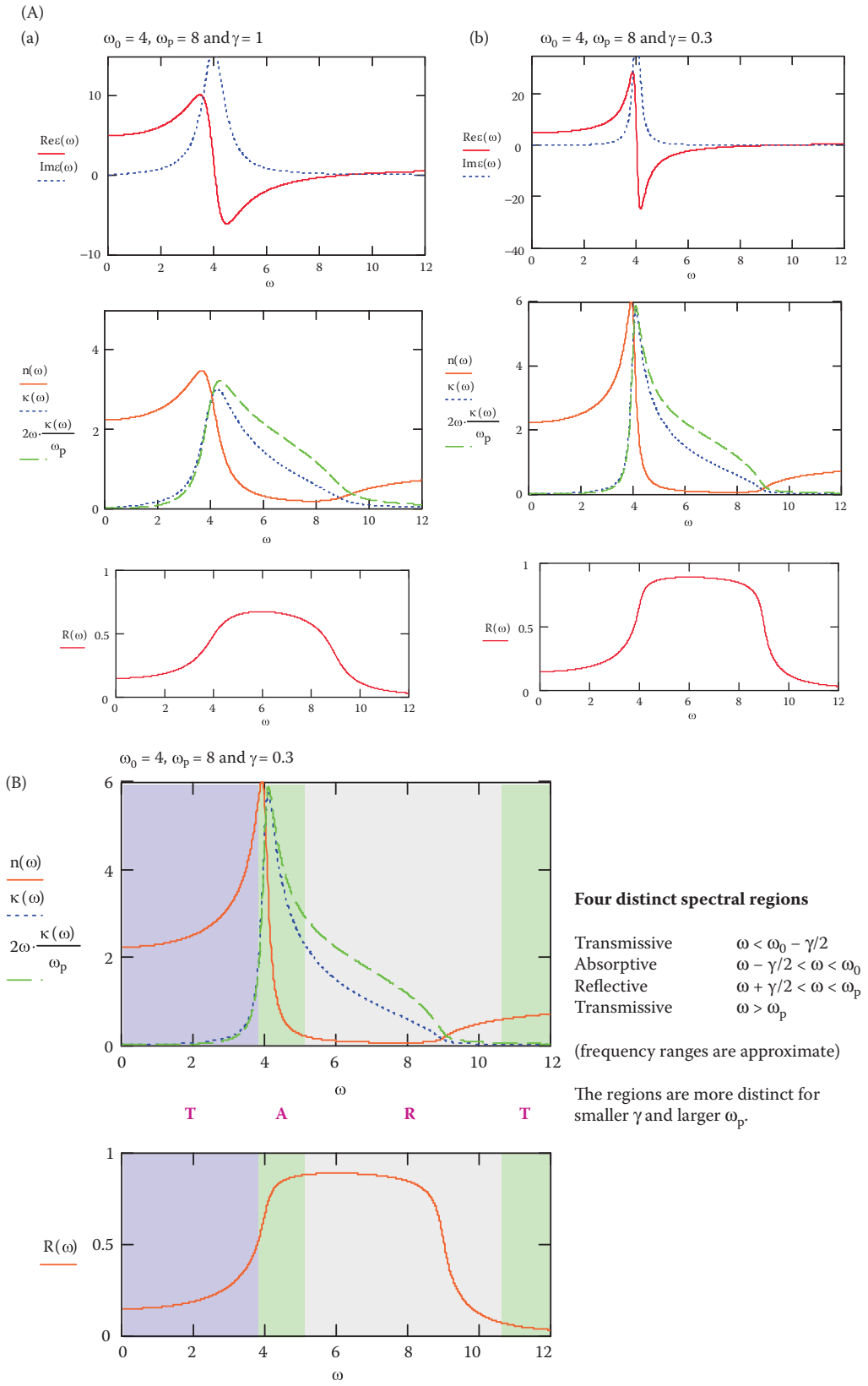


FIGURE 5.61 (A) Compare the real and imaginary parts of a dielectric function as a function of ω, the real and imaginary parts of the refractive index *n* as a function of ω, and the reflectivity from air at normal incidence for $\gamma = 1$ (broad resonance, a) and $\gamma = 0.3$ (narrow resonance, b). (B) We have colored in four distinct spectral regions, marked with *T* for transmissive, *A* for absorptive, *R* for reflective, and *T* for transmissive again—TART. We have plotted also a scaled version of α by plotting $2\omega\kappa$ scaled to ω_p.

the reflecting band. Above ω_0, n is well below unity. Above ω_p, n rises, but only in the high-frequency limit does n approach unity.

Local Field Corrections The calculations of the dielectric constant we have discussed until now are valid for a rarefied gas with a low density of atoms. In "dense optical media," such as solids, there are other factors that we have to consider. In a real solid, atoms may feel a field from an incident light beam, but they also experience an induced field from the surrounding atoms. In other words, the local field $\mathbf{E}_L$ is the sum of a field without matter $\mathbf{E}_0$ and an induced dipolar field $\mathbf{E}_{dipole}$ from all the other surrounding atoms, or:

$$\mathbf{E}_L = \mathbf{E}_0 + \mathbf{E}_{dipole} \tag{5.203}$$

In our discussions of dipole oscillators in solids, we should have been using $\mathbf{E}_L$ not $\mathbf{E}_0$! The degree of polarization in a solid is related to the structure of the material. As a consequence, dielectric behavior in electrostatic and alternating electric fields depends on both static and dynamical properties of the solid. The local electric field can be calculated according to the crystal structure by the method of Clausius and Mossotti. For example, for cubic structures the Clausius-Mossotti equation reads (e.g., see Ashcroft and Mermin, *Solid-State Physics*, p. 539):

$$\mathbf{E}_L = \mathbf{E}_0 + \frac{\mathbf{P}}{3\varepsilon_0}, \tag{5.204}$$

which is larger than $\mathbf{E}_0$. Similar relations can be derived for any solid. For a solid consisting of atoms of type j with a concentration N_j, we then derive for the polarization $\mathbf{P}$:

$$\mathbf{P} = \varepsilon_0 \sum_j N_j \mathbf{E}_L \chi_{j,e} = \varepsilon_0 \sum_j N_j \left(\mathbf{E}_0 + \frac{\mathbf{P}}{3\varepsilon_0} \right) \chi_{j,e}$$

$$= \varepsilon_0 \sum_j N_j \chi_{j,e} \mathbf{E}_0 + \sum_j N_j \chi_{j,e} \left(\frac{\mathbf{P}}{3} \right) = (\varepsilon_r - 1)\varepsilon_0 \mathbf{E}_0 \tag{5.205}$$

where $\chi_{j,e} = \frac{e^2}{\varepsilon_0 m_e} \sum_j \frac{f_j}{(\omega_{0,j}^2 - \omega^2 - i\omega\gamma_j)}$. But we know that $\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}$, $\mathbf{D} = \varepsilon_0 \varepsilon_r \mathbf{E}$, and $\mathbf{P} = \varepsilon_0(\varepsilon_r - 1)\mathbf{E}$, and this way we obtain the Clausius-Mossotti (CM) relation:

$$\frac{(\varepsilon^2 - 1)}{(\varepsilon^2 + 2)} = \frac{N\chi_{j,e}}{3} \tag{5.206}$$

This relation links the polarizability to the dielectric constant. Using $\varepsilon_r = n^2$ we find the so-called *Lorentz-Lorenz relationship*:

$$\frac{(n^2 - 1)}{(n^2 + 2)} = \frac{N\chi_{j,e}}{3} \tag{5.207}$$

And we find that the refractive index approximately follows the formula:

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{n_c^2 - 1}{n_c^2 + 2} = \frac{Ne^2}{3\varepsilon_0 m_e} \sum_j \frac{f_j}{(\omega_{0,j}^2 - \omega^2 - i\omega\gamma_j)} \tag{5.208}$$

We follow with some examples from the n, κ database (http://www.ioffe.ru/SVA/NSM/nk/index). In Figure 5.62a we show the refractive index n and absorption index κ versus photon energy for silicon. At visible frequencies, i.e., below the resonance, Si is "transmissive." But the high refractive index of Si still yields high reflectivity (30–50%), and all visible frequencies are reflected strongly, so Si appears as a grayish reflector throughout the visible spectrum (1.7–3.2 eV) "gray metallic."

In Figure 5.62b we show the refractive index for two common dielectrics (Si_3N_4 and SiO_2). Dielectrics typically exhibit one resonance only, and at low frequencies they are transmissive. The resonances are typically in the UV range associated with electron polarization. There is an additional resonance in the IR range for polar materials resulting from nuclear motion and orientation (see Figure 5.63). We come back to describe these polar materials in more detail further below. To design a material with a high dielectric constant, we choose ions with high polarizability and a high density of them.

Example 5.6: Estimation of $\varepsilon_r(\omega = 0)$ for Si

The reflectance R of Si increases sharply at about 3 eV, and we may take this as an estimate for ω_0. Hence $\omega_0 \approx 3 \times 1.6 \times 10^{-19} = 4.53 \times 10^{15}$ rad/s. Because $\varepsilon_r(\omega) = 1 + \frac{\omega_p^2}{\omega_0^2 - \omega^2 - i\gamma\omega}$, $\varepsilon_r(\omega = 0) = 1 + \frac{\omega_p^2}{\omega_0^2}$, so if

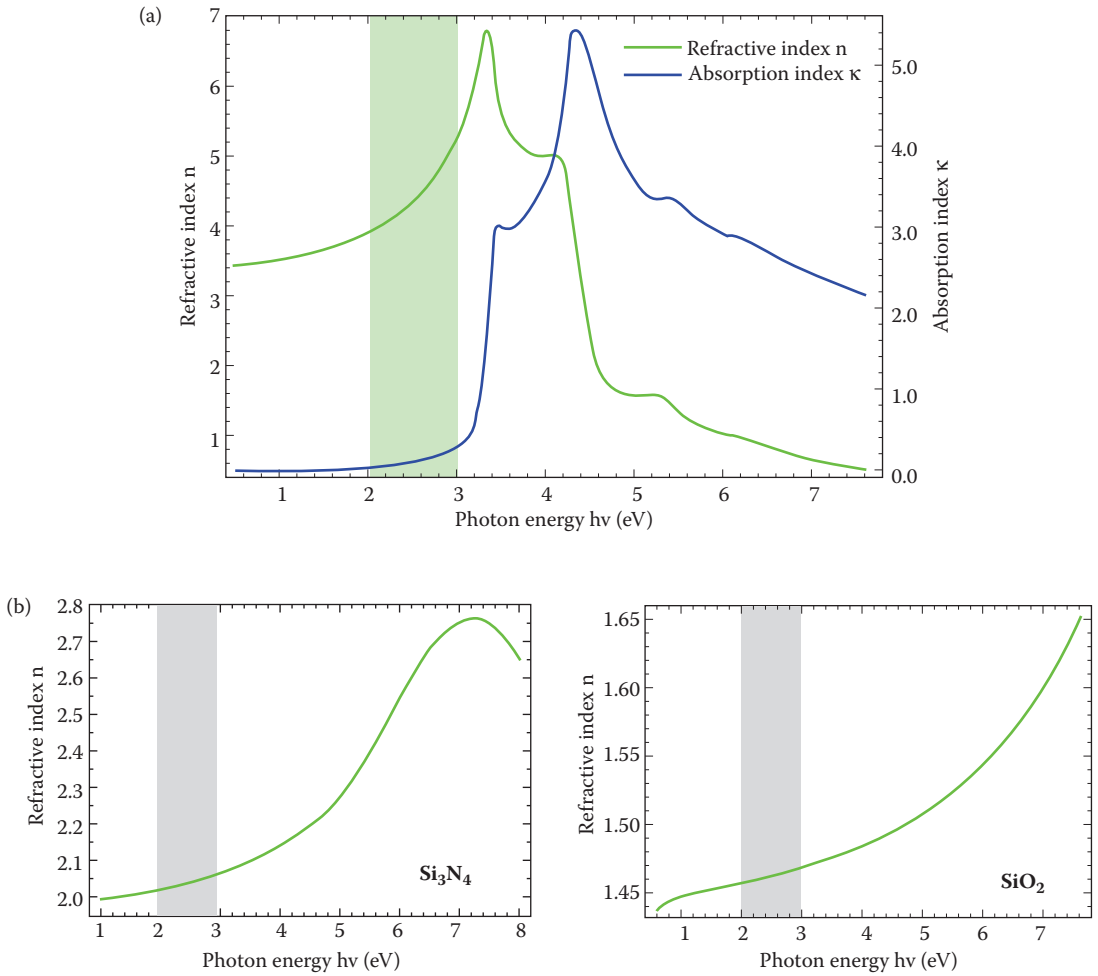


FIGURE 5.62 Examples from the *n*, κ database (http://www.ioffe.ru/SVA/NSM/nk/index.html#A3B6). (a) Refractive index *n* and absorption index κ versus photon energy for silicon. (b) The refractive index for two common dielectrics (Si_3N_4 and SiO_2). The color bands in the three figures roughly mark the visible spectrum (1.7–3.2 eV).

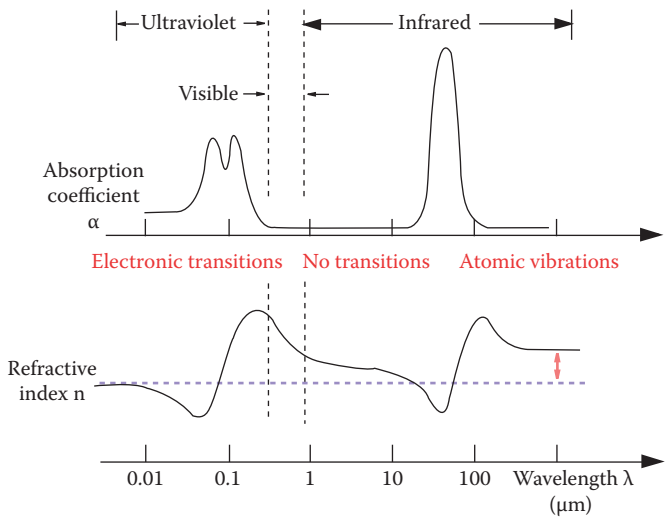


FIGURE 5.63 The absorption coefficient and the index of refraction versus wavelength for a generic polar insulator. As we derived earlier: $\alpha = 2\mathbf{k}\kappa$.

we can determine ω_p, we can estimate $\varepsilon_r(\omega = 0)$. Now we remember that $\omega_p = \sqrt{ne^2/\varepsilon_0 m}$, and because each Si atom has four valence electrons, $n = 4N_{Si} \approx 4 \times 2\ 10^{28}\ m^{-3}$. This gives us an estimate of $\omega_p \approx 1.6 \times 10^{16}$ rad/s (corresponding to about 10.5 eV), and hence $\varepsilon_r(\omega = 0) \approx 14$. This compares to a measured value for $\varepsilon_r(\omega = 0)$ of 12, so our approximations are reasonable.

Optical Properties of Metals

Introduction

The Drude model implies that only the plasma frequency ω_p dictates the appearance of metals. This works indeed for many metals, say for aluminum, but it does not explain why copper is red, gold is yellow, and silver is colorless. In fact, the appearance

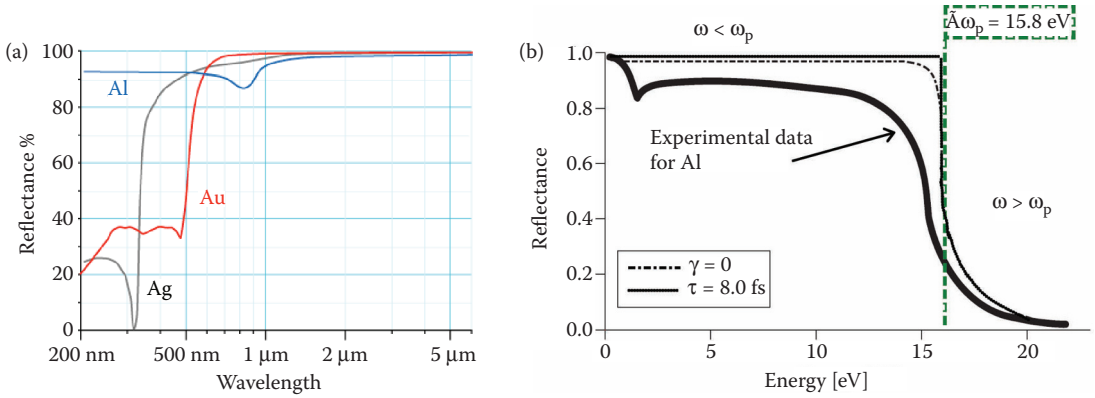


FIGURE 5.64 (a) Reflectance curves of Ag, Au, and Al. (b) Reflectance versus energy for an ideal metal with zero damping ($\gamma = 0$) and experimental data for aluminum. See also Example 5.8.

of these metals is characterized by a different edge in the reflectance spectrum (see Figure 5.64a), but the fact is that these three metals have the same number of valence electrons, so according to the Drude model they all should have been the same. Moreover, the calculated plasma frequency for all three should lie at about 9 eV, well outside the visible region, so the Drude plasma frequency cannot in itself account for the colors of Cu and Au. Colored metals have interband transitions in the visible that remove particular frequencies from the reflected spectrum. The combined effects of the free electrons and the bound d-electrons influence the reflectance properties of the metal in these cases. In Chapter 3, we pointed out that in a metal where both bound electrons $[\varepsilon_B(\omega)]$ and conduction electrons $[\varepsilon_r(\omega)]$ contribute to the dielectric constant, an effective dielectric constant $\varepsilon_{Eff}(\omega)$ must be defined as:

$$\varepsilon_{Eff}(\omega) = \varepsilon_B(\omega) + \varepsilon_r(\omega) \tag{3.48}$$

Whereas the Drude model explains the free electron contribution $[\varepsilon_r(\omega)]$ to $\varepsilon_{Eff}(\omega)$, the bound electron effect $[\varepsilon_B(\omega)]$ can only be treated by the Lorentz model. Based on the Drude-Lorentz model, we will now review optical properties of bulk metals (3D) and metals in reduced dimensionalities in plasmonics.

Reflectance and Absorption of Electromagnetic Waves in Metals

Reflectance of Conductors Metals can be treated as plasmas because they contain equal numbers of fixed positive ions and free electrons. In Figure 5.64a, we plot the reflectance R of Al, Au, and Ag as a function of wavelength. In Figure 5.64b, we show the reflectance of an ideal metal and that of aluminum versus the energy of the incoming radiation. A characteristic feature of all metals is that they are shiny as a result of the interaction of light with the free electrons (plasma excitations) and that their reflectance is close to 100% (R = 1) below the plasma frequency ω_p but much smaller at higher ω. At low frequencies, metals with plenty of free carriers are very good reflectors. The free carriers can respond rapidly, on a time scale of an AC cycle, so as to screen the external applied AC fields. Earlier we calculated the normal incidence reflectance for any material as:

$$R(\omega) = \frac{\left|E_m^r\right|^2}{\left|E_m^i\right|^2} = \left|\frac{n_c - 1}{n_c + 1}\right|^2 = \frac{(n-1)^2 + \kappa^2}{(n+1)^2 + \kappa^2} \tag{5.148}$$

If we consider a lightly damped (highly conductive) system, where the damping factor γ tends to zero (an ideal metal would come with zero damping $\gamma = 0$), the real components of the dielectric constant $\varepsilon_r'(\omega, \mathbf{k}) \approx 1 - \frac{\omega_p^2}{\omega^2}$ (Equation 3.44) dominate because the imaginary component $\varepsilon_r''(\omega, \mathbf{k}) \approx \frac{\omega_p^2 \gamma}{\omega^3}$ (Equation 3.45) tends to zero. For the complex refractive index in this case we can write:

$$n_c(\omega) = \sqrt{\varepsilon_r'(\omega)} = \sqrt{1 - \frac{\omega_p^2}{\omega^2}} \tag{5.209}$$

Substituting this expression in Equation 5.148 for the reflectance R, we see that when $\omega < \omega_p$, n_c is imaginary, R = 1, and the metal is shiny (say, up to

the near-infrared). At higher frequencies (visible and UV), with $\omega > \omega_p$, n_c turns real, the reflectance R suddenly falls (for $\omega \to \infty$, $R \to 0$), and metals are transparent in the UV. It follows from the expression for the plasma frequency of an FEG in a metal [$\omega_p = \sqrt{\frac{ne^2}{m_e\varepsilon_0}}$ (Equation 3.41)] that the higher the conductivity (n larger), the higher the plasma frequency. And indeed, plasma frequencies ω_p are in the optical range (UV) for metals and in the terahertz to the infrared region for semiconductors and insulators (see also Table 3.2).

Without any damping ($\gamma = 1/\tau = 0$), the reflectivity R equals 1 and then drops very steeply (see rectangular box in Figure 5.64b). Introducing some damping, R becomes less than 1, and the reflectance decrease is less severe (see Figure 5.64b where we used $\tau = 8.0$ fs, a value calculated for aluminum; see also Example 5.8). When the conduction electrons can no longer move rapidly enough to screen the external applied field on the time scale of an AC cycle, only the core electrons bound to the atoms can contribute to the dielectric response; the metal behaves now as a transparent dielectric. The frequency at which the metallic-to-dielectric transition occurs is the plasmon frequency ω_p. At $\omega = \omega_p$, the real part of the dielectric constant becomes zero. Hence $n_c(\omega_p) = 0$, which means the phase velocity tends to ∞. (This occurs in the anomalous dispersion regions shown in Figure 5.29.) Another way to describe this is that the wavelength, $\lambda = 2\pi c/n_c\omega \to \infty$ as $\omega \to \omega_p$. This means that all the electrons are oscillating in phase throughout the propagation length of the material.

Absorption and Penetration Depth of Electromagnetic Radiation in Conductors Optical measurements of the complex dielectric constant $\varepsilon_r(\omega,k)$ of a plasma are equivalent to AC conductivity measurements. This can be gleaned from a comparison of the dielectric function of a free electron plasma with its AC conductivity. In Chapter 3 we calculated for the AC conductivity of a plasma:

$$\sigma(\omega) = \frac{ne^2}{m_e}\left(\frac{1}{\frac{1}{\tau} - i\omega}\right) \quad \text{(based on Equation 3.30)}$$

and for the complex dielectric constant we derived:

$$\varepsilon_r(\omega,\mathbf{k}) = \frac{\varepsilon(\omega,\mathbf{k})}{\varepsilon_0} = \left[1 - \frac{ne^2}{m_e\varepsilon_0(\omega^2 + i\gamma\omega)}\right]$$

(based on Equation 3.40)

Comparing these expressions we note that:

$$\varepsilon_r(\omega,\mathbf{k}) = 1 + \frac{i\sigma(\omega)}{\varepsilon_0\omega} \tag{5.210}$$

At very low frequencies ($\omega \ll \tau^{-1}$) we can derive a useful relationship between the conductivity of the free electron gas and the absorption coefficient (α) of the metal. The real and imaginary parts of the complex dielectric constant are $\varepsilon_r'(\omega,\mathbf{k}) = 1 - \frac{\omega_p^2\tau^2}{1+\omega^2\tau^2}$ (Equation 3.42) and $\varepsilon_r''(\omega,\mathbf{k}) = \frac{\omega_p^2\tau}{\omega(1+\omega^2\tau^2)}$ (Equation 3.43), and one appreciates that, at low frequencies, $\varepsilon_r''(\omega,\mathbf{k}) \approx \frac{\omega_p^2\tau}{\omega}$ can become much larger than $\varepsilon_r'(\omega,\mathbf{k}) \approx 1 - \omega_p^2\tau^2$. Based on Equations 5.117 for n and κ, we then derive:

$$n \sim \kappa = \left(\frac{\varepsilon_r''}{2}\right)^{\frac{1}{2}} \tag{5.211}$$

Therefore, for the absorption coefficient we calculate:

$$\alpha = \frac{2\kappa\omega}{c} = \frac{2\omega\left(\frac{\varepsilon_r''}{2}\right)^{\frac{1}{2}}}{c} \tag{5.212}$$

and with $\varepsilon_r''(\omega,\mathbf{k}) \approx \frac{\omega_p^2\tau}{\omega}$ the absorption coefficient of a metal for $\omega \ll \omega_p$ and $\omega\tau \ll 1$ is given as:

$$\alpha = \left(\frac{2\omega_p^2\tau\omega}{c}\right)^{\frac{1}{2}} \tag{5.213}$$

or, finally, recognizing that $\omega_p^2\tau = \frac{\sigma_0}{\varepsilon_0}$ and $c^2 = \frac{1}{\varepsilon_0\mu_0}$, we obtain:

$$\alpha = \sqrt{2\sigma_0\omega\mu_0} \tag{5.214}$$

Thus, the optical absorption coefficient of a metal for $\omega \ll \omega_p$ and $\omega\tau \ll 1$ is proportional to the square root of the DC conductivity and the frequency.

We calculated earlier the decay of the intensity of an EM wave, propagating in the x-direction in a given material, as $I(x) = I_0 e^{-ax}$ (Equation 5.110). The intensity in this expression decreases to $1/e$ (or ~36%) of I_0 at a depth x for which $\alpha x = 1$. This depth is known as the *penetration depth* δ, and it is given by:

$$\delta = \left|\frac{1}{\alpha}\right| = \sqrt{\frac{1}{2\sigma_0 \omega \mu_0}} \qquad (5.215)$$

Because this depth is usually very small in metals (given that σ_0 in Equation 5.215 is very large), it is also referred to as *skin depth*. Horace Lamb first described the effect in a paper in 1883. Whereas at low frequencies most of a conductor's cross-section (say, a copper wire) carries the current, the current moves to the skin of the conductor as frequency increases. This penetration depth of an electromagnetic wave in a solid is illustrated in Figure 5.65 for Al, Cu, Fe, and Pb. Skin depth is dependent on the type of metal and the frequency of the applied field. At high frequencies the skin depth is very shallow, and the fields are often considered to fall to zero within a few millimeters. The skin effect predominates at gigahertz frequencies. The skin effect has practical consequences in the design of radiofrequency (RF) and microwave circuits and to some extent in AC electrical power transmission and distribution systems.

It is important to keep in mind that Equation 5.215 is only valid when $\omega\tau << 1$. At higher frequencies (mid-infrared, visible, UV) this is no longer valid, and we must use the whole expression for ε_r (Equation 3.40). We also did not include possible contributions of bound electrons (see Lorentz model below).

Example 5.7: Cu ($\sigma_0 = 6.5\ \Omega^{-1}\text{cm}^{-1}$) $\rightarrow \delta = 8.8$ mm at 50 Hz, 6 μm at 100 MHz.

Example 5.8: The DC conductivity of aluminum at room temperature is $\sigma_0 = 4.1 \times 10^7\ \Omega^{-1}\text{m}^{-1}$, and the free electron density is $n = 1.81 \times 10^{29}\ \text{m}^{-3}$. What is the reflectance R of the surface at $\lambda = 500$ nm under normal incidence?

1. We first calculate the damping ($\gamma = 1/\tau$ from the conductivity:

$$\sigma_0 = \frac{ne^2\tau}{m_e} \text{ (Equation 3.7)} \rightarrow \tau = 8.0\ 10^{-15}\ \text{s} = 8\ \text{fs}$$

2. Next we calculate the plasma frequency ω_p:

$$\omega_p = \sqrt{\frac{ne^2}{m_e \varepsilon_0}} \text{ (Equation 3.41)} \rightarrow \omega_p = 2.4\ 10^{16}\ \text{Hz}$$

3. A 500-nm wavelength corresponds to $\omega = 2\pi c/\lambda = 3.8 \times 10^{15}$ Hz. Now we can calculate the real and imaginary parts of the dielectric function:

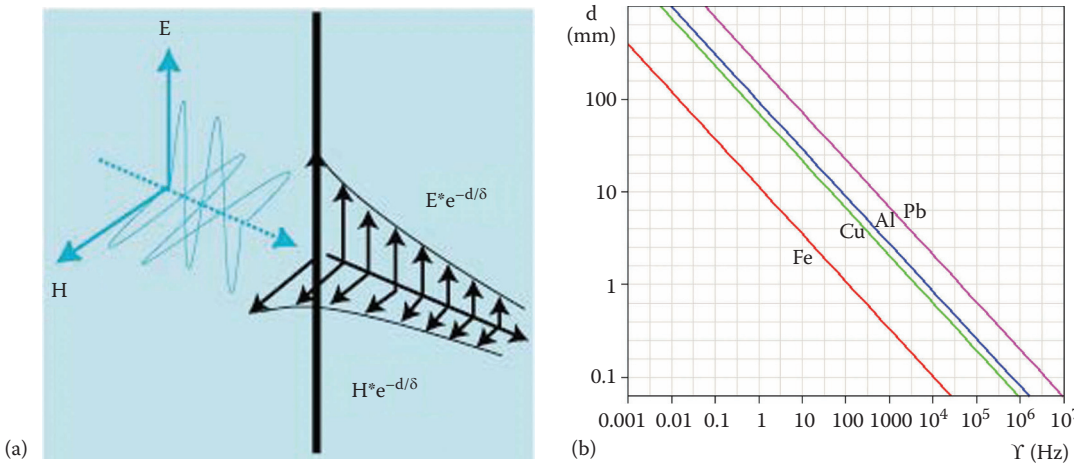


FIGURE 5.65 The depth of penetration δ of electromagnetic waves in a metal (a). Skin depth is dependent on the type of metal and the frequency of the applied field. At high frequencies the skin depth is very shallow, and the fields are often considered to be 0 in a few millimeters. Examples for Fe, Cu, Al, and Pb are shown (b).

$$\varepsilon_r'(\omega,k) = 1 - \frac{\omega_p^2\tau^2}{1+\omega^2\tau^2} \text{ (Equation 3.42)} \rightarrow -39$$

$$\varepsilon_r''(\omega,k) = \frac{\omega_p^2\tau}{\omega(1+\omega^2\tau^2)} \text{ (Equation 3.43)} \rightarrow 1.3$$

$$n(\omega) = \sqrt{\frac{(\varepsilon_r')^2 + (\varepsilon_r'')^2 + \varepsilon_r'}{2}}$$

(Equation 5.117) $\rightarrow n = 0.1$ and $\kappa = 6.2$

$$\kappa(\omega) = \sqrt{\frac{(\varepsilon_r')^2 + (\varepsilon_r'')^2 - \varepsilon_r'}{2}}$$

Finally, we calculate the reflectance R from:

$$R(\omega) = \frac{\left|E_m^r\right|^2}{\left|E_m^i\right|^2} = \left|\frac{n_c - 1}{n_c + 1}\right|^2 = \frac{(n-1)^2 + \kappa^2}{(n+1)^2 + \kappa^2}$$

(Equation 5.148) $\rightarrow$ 99%

Even though ω/ω_p is ~0.2, the reflectance remains very high.

To help interpret this example, in Figure 5.66 we look at some calculated plots of $\varepsilon_r(\omega)$, $n(\omega)$, $\alpha(\omega)$, and $R(\omega)$. These are plotted for $\omega_p = 10$ and for $\gamma \approx 0$ (Figure 5.66a) or $\gamma = 0.5$ (Figure 5.66b). In the limit of no damping, the refractive index $n = 0$ and $R = 1$ for $0 < \omega < \omega_p$. Above ω_p, κ is zero, and the reflectance R drops as n rises from zero toward unity. Note that even for $\varepsilon_r'' = 0$, κ and hence α are not zero. Introducing some damping causes R to be <1, and the reflectance drop at ω_p is less severe. The sharp edge in the reflectance seen at the plasma frequency is the predominant spectral feature in the optical properties of metals.

Drude-Lorentz Theory for Metals

The combined effects of the free electrons and the bound *d*-electrons determine the optical properties of some metals. In a metal where both bound electrons $[\varepsilon_B(\omega)]$ and conduction electrons $[\varepsilon_r(\omega)]$ contribute to the dielectric constant, an effective dielectric constant $\varepsilon_{Eff}(\omega)$ was earlier defined as:

$$\varepsilon_{Eff}(\omega) = \varepsilon_B(\omega) + \varepsilon_r(\omega) \qquad (3.48)$$

While the Drude model explains the free electron contribution $[\varepsilon_r(\omega)]$ to $\varepsilon_{Eff}(\omega)$, the Lorentz model explains the bound electron effect $[\varepsilon_B(\omega)]$ and clarifies why copper is red, gold is yellow, and silver is colorless. All three metals have filled *d*-shells: copper has the electronic configuration $[Ar].3d^{10}.4s^1$, silver $[Kr].4d^{10}.5s^1$, and gold $[Xe].4f^{14}.5d^{10}.6s^1$. These

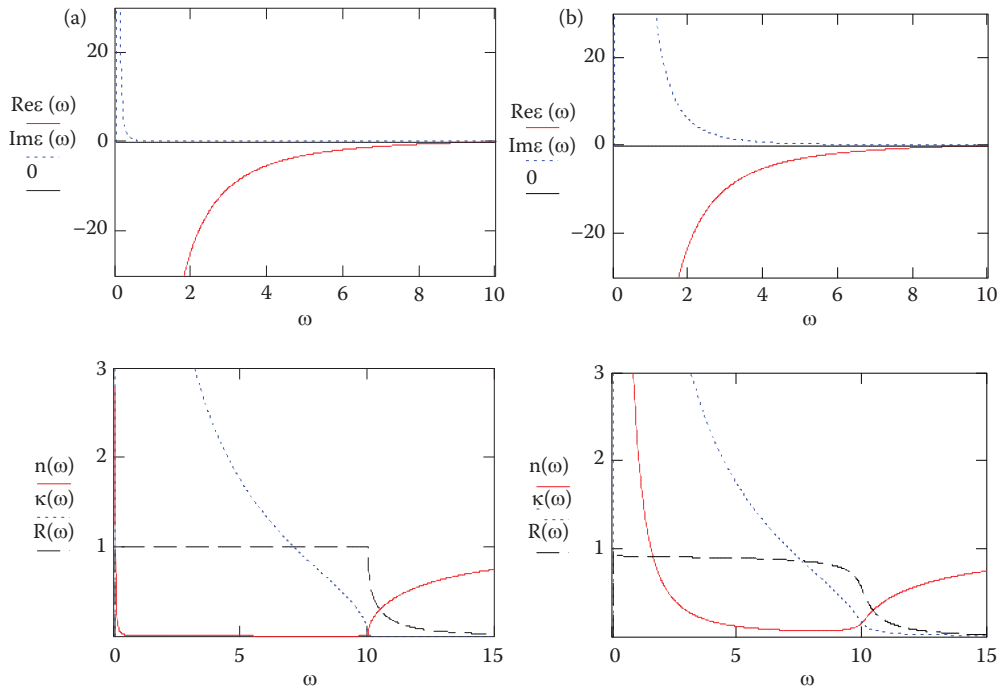


FIGURE 5.66 Calculated plots of $\varepsilon_r(\omega)$, $n(\omega)$, $\alpha(\omega)$, and $R(\omega)$. Plotted for $\omega_p = 10$ and for $\gamma \approx 0$ (a) or $\gamma = 0.5$ (b).

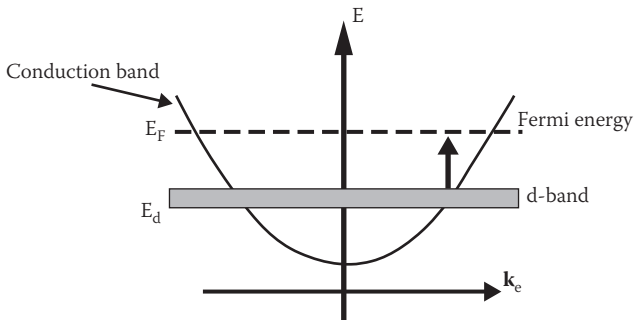


FIGURE 5.67 The combined effects of the free electrons (Drude model) and the interband transitions resulting from bound *d*-electrons (Lorentz model) influence the reflectance properties of the metal.

d-electron bands lie below the Fermi energy of the conduction band (Figure 5.67), and transitions from the *d*-band to empty states above the Fermi level occur over a fairly narrow band of energies ($\hbar\omega_0 = E_F - E_d$). These transitions are modeled as an additional Lorentz oscillator.

In Figure 5.68 we compare the optical properties of aluminum and silver in a plot of the real part and the imaginary part of the complex dielectric constant as a function of the photon energy. For most metals the only contribution to ε_{Eff} (Equation 3.48) stems from conduction electrons as illustrated by the experiments with aluminum where at $\omega = \omega_p$, $\varepsilon_{Eff}(\omega) \approx 1 - \frac{\omega_P^2}{\omega^2} \approx 0$ and $\varepsilon_B \approx 0$, with a monotonous decrease of $\varepsilon_{Eff}(\omega)$ for decreasing frequency as predicted (Figure 5.68a). In other words, for Al we observe the predicted $\varepsilon_{Eff} = 1 - \frac{\omega_P^2}{\omega^2} + i\frac{\omega_P^2}{\omega^3\tau}$ (see Equation 3.49).

For Ag, with a σ of $6.2 \times {}^7\Omega^{-1}m^{-1}$ and $\omega_p = 9.65 \times 10^{14}$ Hz (=311 nm or 4 eV), the drop in the $\varepsilon_{Eff}(\omega)$ curve toward lower frequencies is not monotonous as in the case of Al; a peak appears in the $\varepsilon''_{Eff}(\omega)$ at around 4 eV, which is reflected in the summation of the two terms [$\varepsilon_{Eff}(\omega)$; see Figure 5.68c]. Such a peak is more typical for the case of bound electrons in a dielectric. In silver, both conduction and bound electrons contribute to ε_{Eff}, and the bound electron contribution is caused by *d*-electrons (bound electrons in a *d* orbit) at 4 eV below the Fermi energy. The reflectance spectrum of silver, shown in Figure 5.68b, reveals a sharp decrease at the same 4 eV, well below the expected plasma frequency (at 9 eV). The reflectance increases again for frequencies just above 4 eV. This behavior comes about because of the *d*-band resonance at 4 eV in silver. The difference between the dielectric functions calculated from the Drude model and those observed from the experiment are a result of the d-band resonance, ε_B. The effect is to "pull" the $\varepsilon_{Eff} \rightarrow 0$ frequency in from 9 eV to about 3.9 eV. This shift in ω_p means that there is a shift in the free plasma oscillation in silver because of the *d*-electrons. This can be explained by noting that the highly polarizable *d*-electrons reduce the electric

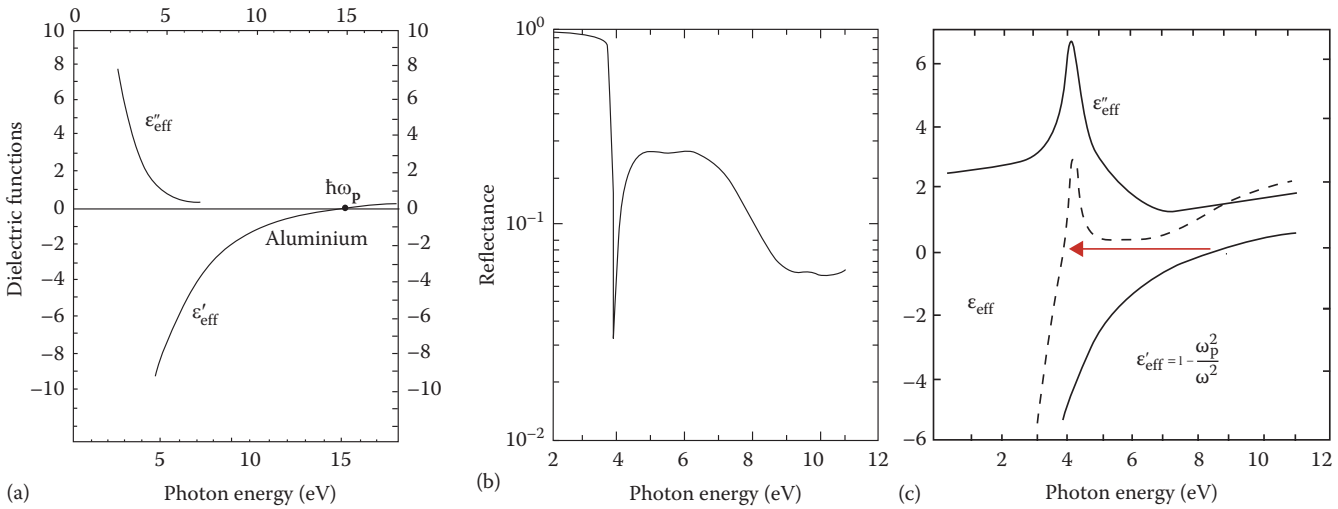


FIGURE 5.68 The real and the imaginary parts of the complex dielectric constant are plotted as a function of the photon energy for aluminum (a) and silver (c). In (b), we also show the reflectance *R* of silver versus photon energy.

field that provides the restoring force involved in these oscillations. A reduced restoring force results in a reduced oscillation frequency.

An interesting and unusual conductor in this regard is tin-doped indium oxide (ITO) with a very large EM wave penetration depth in the visible EM wave range. ITO is a semiconducting material that is transparent in the visible! The material is very useful in many low-current applications, such as liquid crystal displays. To obtain this transparency, the material combines a low electron density n with a high mobility μ_e, i.e., the electrons travel through the material with relatively few collisions. The plasma frequency of ITO depends on the Sn density and is typically around 0.7 eV ($\lambda \sim 1.7$ μm), making the material of lesser use in the near infrared ($\lambda >$ 1.7 μm) than it is in the visible. The properties of ITO, again, can be modeled by the Drude-Lorentz models.

Metal Nanoparticles

Introduction In the preceding section we used the Drude-Lorentz model to explain electrooptical properties of bulk metals; now we explore what happens with those properties when lower-dimensional metal structures are considered. We start with a comparison of metal nanoparticles with semiconductor nanoparticles, a comparison we will come back to after applying the Drude-Lorentz model to bulk and nanoparticle semiconductors (see Table 5.10). We finish this section adding some more details to plasmonics applications see the section Plasmonic Applications I, a discussion we started earlier.

Quantum Effects in Metal and Semiconductor Nanoparticles Compared In Chapter 3, we studied quantum-confined materials and found that for quantum effects to appear in an electronic conductor, the dimension along which there is quantum confinement must be smaller than the de Broglie wavelength of the electrons. The de Broglie wavelength for electrons at the Fermi energy corresponds to the size of an electron in a solid and is represented by its Fermi wavelength $\lambda_F = 2\pi/\mathbf{k}_F$ with $\mathbf{k}_F$, the Fermi wave vector, given by $(3n\pi^2)^{1/3}$ (Equation 3.238), where n is the electron density. In most metals, the electron density is very large ($n > 10^{21}\,\mathrm{cm}^{-3}$) and the Fermi wavelength is of the order of a few nanometers. In semiconductors, the lower carrier density ($n > 10^{21}\,\mathrm{cm}^{-3}$) gives rise to larger electron wavelengths of several tens of nanometers (see also Table 3.12). Therefore, confinement and quantization phenomena are visible in semiconductors already at dimensions of 200 nm, whereas in metals they typically are seen only at 1–2 nm. This is one important reason why quantum effects have been observed mainly in semiconductor nanostructures, which have large Fermi wavelengths, rather than in metals, which have Fermi wavelengths comparable to a few crystal lattice periods only. We can also approach the difference in quantization effects between a metal and semiconductor nanoparticle from the point of view of the separation between adjacent energy levels. The average level spacing at the Fermi level between subsequent levels can be estimated from Equation 3.237:

$$\Delta E_{n+1-n} = \frac{1}{G(E)_{3D}(E = E_F)} = \frac{2E_F}{3n_{3D}} \qquad (3.237)$$

Based on this expression for a spherical Au nanoparticle with a radius R = 2 nm, an average level spacing $\Delta E = E_{n+1} - E_n$ of 2 meV only is found, whereas in Volume III, Chapter 3 we calculate the energy separation between adjacent levels for a 2-nm semiconducting CdSe particle as 0.76 eV. Obviously energy-level quantization is far more pronounced in semiconductor dots than in metal dots. In addition, the appearance of the plasmon oscillations of metal particles often obscures quantization effects in metals.

Beyond the Fermi wavelength difference, there are two additional reasons for this increased energy separation in the case of semiconductors: 1) the energy-level spacing, i.e., the energy difference $E_{n+1} - E_n$, increases for low-lying states in a 3D potential well (see Figure 3.142); and 2) the effective mass for electrons is smaller in semiconductors than in metals. With a smaller effective, we saw in Chapter 3, the "blue" shift becomes even larger (e.g., see Equation 3.251).

FIGURE 5.69 (a) The Lycurgus Cup (glass; British Museum, fourth century AD). When illuminated from outside, it appears green. However, when illuminated from within the cup, it glows red. The red color is caused by very small amounts of gold powder (about 40 ppm), which have an absorption peak at around 520 nm. (b) Luster decorated 16th century Renaissance pottery, Gubbio, Italy. NPs = nanoparticles.[10] (Image and information from the National Gallery of Art, Washington, DC.)

Whereas in semiconductor structures with reduced dimensionalities electron confinement leads to quantization of energetic states, metallic nanostructures exhibit major changes in their optical spectra as a result of light absorption and the creation of plasmonics waves at the surface of these structures. As mentioned above, these effects often mask the metal nanoparticle quantum effects.

Optical Properties of Metal Nanoparticles The optical properties of nanoparticles are spectacular, and they have promoted a great deal of excitement during the past few decades in metal nanoparticles. Examples of the use of metal nanoparticles for decorative purposes go as far back as the Roman time, such as those contained in the glass of the famous Lycurgus Cup (fourth century AD) pictured in Figure 5.69a. The cup can still be seen at the British Museum and possesses the unique feature of changing color depending on the light in which it is viewed. It appears green when viewed in reflected light but looks red when a light is shone from inside and is transmitted through the glass. Analysis of the glass has revealed that it contains a very small amount of tiny metal crystals containing Ag and Au in an approximate molar ratio of 14:1. It is the presence of these nanocrystals that gives the Lycurgus Cup its special color display. Other examples include decorative stained glasses in cathedrals (see Figure 5.48) and the Renaissance pottery shown in Figure 5.69b. The beautiful Maya blue, a paint often used in Mesoamerica, involves simultaneously metal nanoparticles and superlattice organization effects (see below). These examples represent applications of nanotechnology before an understanding of mechanisms or design rules had emerged. It was not until 1857 that Michael Faraday reported a systematic study of the synthesis and colors of colloidal gold: "The mere variation in the size of [gold] particles gave rise to a variety of resultant colors."[9] Before that time, colloidal gold was known and used in the Middle Ages for its purported health restorative

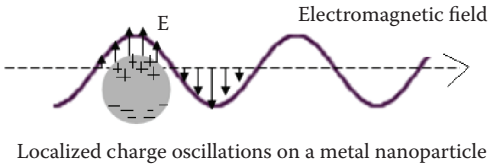


FIGURE 5.70 The interaction of an electromagnetic wave with a metal nanoparticle.

properties. Alexandria, Egypt was believed to have been the original founding place for the use of gold in medicine by a group of adepts known as *alchemists*. The alchemists developed an "elixir" made of liquid gold that purportedly had the ability to restore youth and perfect health. Paracelsus developed medicines from metallic minerals, including gold, to cure the sick.*

Earlier in this chapter we learned that 2D macroscopic metallic/dielectric interfaces require special geometries for plasmon excitation. However, in nanoparticles we saw the surface plasmon wave is localized (not traveling), and so no special excitation geometry is required (Figure 5.70). These "localized plasmons" of metallic nanoparticles may be excited directly by light absorption. For metallic nanoparticles significantly smaller than the wavelength of light, light absorption concentrates to a narrow wavelength range. Wavelengths that are absorbed by metal nanoparticles and produce such a wave are called *surface plasmon bands* or simply *plasmon bands*.

The interaction of light with metal nanoparticles is different than that for a bulk piece of metal. Bulk Au, for example, looks yellowish in reflected light, but thin films of Au look blue in transmission. The characteristic blue of gold particles (say 400–500 nm) changes to orange, via several tones of purple and red, as the particle size is reduced to ~3 nm (see Figure 5.71b). These color changes are the result of changes in the surface plasmon resonance ω_p, the frequency at which conduction electrons oscillate in response to the alternating electric field of incident electromagnetic radiation (Figure 5.71a). Only metals with free electrons (essentially Au, Ag, Cu, and the alkali metals) possess plasmon resonances in the visible spectrum, which give rise to intense colors. Gold and silver are often used in nanophotonics because of their excellent chemical stability and very good conductivity. The position and shape of the plasmon absorption band depend on the material, size, and shape of the nanoparticles, as well as on the surrounding medium, i.e., its dielectric constant.

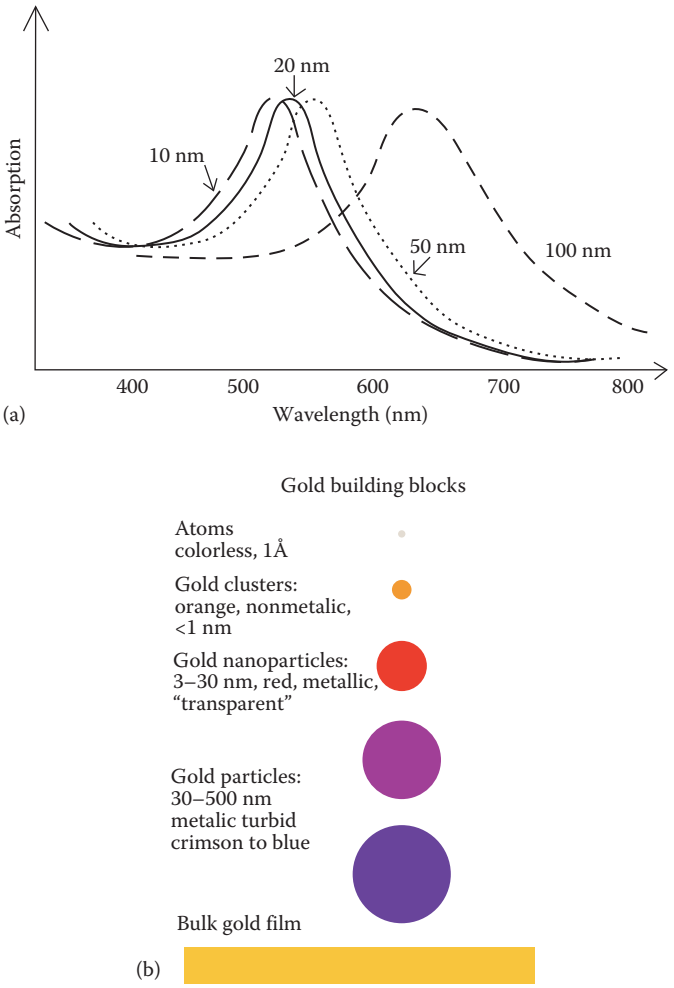


FIGURE 5.71 The surface plasmon resonance (SPR) dependence on size for gold nanoparticles (a). The properties of a material depend on the type of motion its electrons can execute, which depends on the space available for them. Thus, the properties of a material are characterized by a specific "length scale," usually in the nanometer range. If the physical size of the material is reduced below this length scale, its properties change and become sensitive to size and shape (b).

* Even today in China, remnants of the belief in the restorative properties of gold remain intact in rural villages, where peasants cook their rice with a gold coin to help replenish gold in their bodies. In India, a certain kind of fine candy is wrapped in extremely fine gold foil that is eaten along with the rest of the delicacy. It has been reported that in the early 1900s doctors would implant a $5 gold piece under the skin, such as a knee joint.

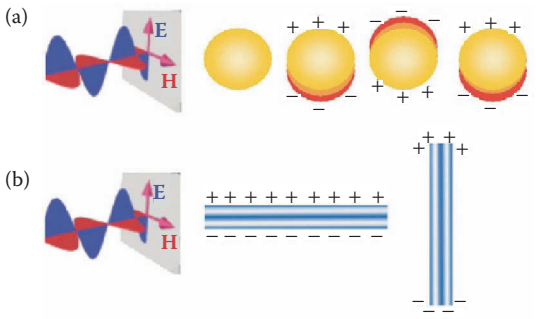


FIGURE 5.72 (a) Schematic drawing of the interaction of an electromagnetic radiation with a metal nanosphere. A dipole is induced, which oscillates in phase with the electric field of the incoming light. (b) Transverse and longitudinal oscillations of electrons in a metal nanorod.

For nonspherical particles, such as rods, the resonance wavelength depends also on the orientation of the electric field. In Figure 5.72 we illustrate the interaction of electromagnetic radiation with spherical and rod-shaped particles. Because for elongated particles the resonance depends on the particle orientation relative to the field, a collection of such particles can show transverse and longitudinal resonances. In addition, when nanoparticles are sufficiently close together, interactions between neighboring particles arise (see further below, Figure 5.76), so that the models for isolated particles do not further hold.

The electron density at the surface of a nanoparticle can induce large electric fields, and the localized surface plasmon resonance ω_p depends on a number of factors, among which particle size and shape, as well as the nature of the surrounding medium, are the most important. As described earlier, the light absorption by metallic nanoparticles is described by a size effect of the coherent oscillation of electrons, induced by an interaction with electromagnetic fields as described above.* Curiously, the major changes that metal nanoparticles exhibit in their optical spectra can be explained based simply on a further detailing of the classical Drude theory.

We use Equation 3.40 from the Drude model for the complex dielectric function and focus on the damping term γ:

$$\varepsilon_r(\omega,\mathbf{k}) = 1 - \frac{\omega_p^2}{\omega^2 + i\gamma\omega} \qquad (5.216)$$

The damping term γ comes about because of scattering ($1/\gamma = \tau$). In the bulk metal, the main contributions to γ are from electron-electron scattering and electron-phonon scattering. In small particles—small compared with the wavelength of the incident light—scattering from the particle boundaries (surfaces) becomes important, and this scattering produces a damping term that is inversely proportional to the particle radius. Thus, the effect of size on plasmon resonance in metal nanospheres is contained in γ.

The real part of the dielectric constant (Equation 3.42) $\varepsilon'(\omega,\mathbf{k})$ is negative for $\omega < \omega_p$. So, for example, the calculated plasma wavelength ω_p for silver is 137 nm and for copper it is 114 nm; therefore, in both cases, $\omega < \omega_p$ and $\varepsilon_r'(\omega,\mathbf{k})$ are negative in the visible region for these particles.

Consider now a dilute dispersion of metal nanoparticles embedded in a dielectric. When shining light on such an assembly, one observes scattering of the incoming radiation, i.e., particles in the path of the EM waves continuously abstract energy from the waves and reradiate it in all directions. To classify these optical phenomena, as we will see later in this chapter, it is handy to introduce a dimensionless size parameter α, which is a measure of the size of the particles compared with the wavelength of the incoming radiation:

$$\alpha = \frac{2\pi r}{\lambda} \qquad (5.217)$$

where r is the radius of the particle and λ is the wavelength of the incoming radiation. With $\alpha << 1$, one obtains Rayleigh scattering, which is inversely proportional to λ^4 (e.g., scattering of visible light by atmospheric molecules; see Equation 5.256). For $\alpha \geq 1$, one obtains Mie scattering, inversely proportional to λ (e.g., scattering of visible light by clouds). Absorption of energy by particles and molecules leads to emission at different frequencies, and the combined effects of absorption and scattering in attenuating the incoming radiant energy are quantified by the extinction, C_{ext}. For very small spherical particles with a frequency-dependent complex

* The collective dipolar oscillation at the plasma frequency is also called the *Fröhlich mode*.

dielectric function, $\varepsilon'_{m(\omega,k)} = \varepsilon' + i\varepsilon''$, embedded in a medium of dielectric constant ε_d, the Mie theory predicts[11]:

$$C_{ext} = \frac{18\pi NV\varepsilon_d^{3/2}}{\lambda} \frac{\varepsilon''}{\left[\varepsilon' + 2\varepsilon_d\right]^2 + \varepsilon''^2} \tag{5.218}$$

where λ = wavelength of the incoming light
ε_d = dielectric constant of the surrounding medium
N = number density of the metal nanospheres
V = volume of a nanoparticle
ε' and ε'' = real and imaginary parts of the metal dielectric constant ε_m

If ε'' is small or weakly dependent on ω, we expect large extinction when the resonant condition is fulfilled (Fröhlich mode):

$$\varepsilon' + 2\varepsilon_d = 0 \tag{5.219}$$

For metal nanospheres in air the plasmon resonance condition is then obtained when $\varepsilon' = -2$ (because ε_d for air = 1). Nanoparticle resonance, from the Drude model, then occurs at:

$$\varepsilon' \approx 1 - \frac{\omega_p^2}{\omega^2} \text{ or } \omega = \frac{1}{\sqrt{3}}\omega_p \approx 0.58\omega_p \tag{5.220}$$

In contrast for a thin film we found that the plasma frequency was given as $\omega = 2^{-1/2}\omega_p$ scaled relative to the bulk plasmon frequency (Equation 5.163). When the shape of the nanoparticles changes from a nanosphere to a nanorod, the surface plasmon absorption spectrum also changes as illustrated in Figure 5.73. For example, the surface plasmon absorption of gold nanorods features two bands: a strong long wavelength band in the near-infrared region resulting from the longitudinal oscillation of electrons and a weak short wavelength band in the visible region around 520 nm caused by the transverse electronic oscillation. To calculate the spectra for elongated particles (nanorods), the orientation with respect to the oscillating field must be taken into account. The absorption spectrum of the nanorods is very sensitive to the aspect ratio (R = length/width = A/B, with the length along A > B = C; see Figure 5.73b). When the aspect ratio increases, the absorption maximum of the longitudinal band is greatly red-shifted. For ellipsoidal particles under illumination, the Mie theory gives the following expression for the extinction coefficient, C_{ext}:

$$C_{ext} = \frac{2\pi V\varepsilon_d^{3/2}}{3\lambda} \sum_j \frac{(1/L_j^2)\varepsilon''}{\left[\varepsilon' + \frac{1 - L_j}{L_j}\varepsilon_d\right]^2 + \varepsilon''^2} \tag{5.221}$$

(see the Gans theory for SPR of ellipsoidal nanoparticles[12,13]), where L_j is a geometry factor, with values between 0 and 1, different along each axis (j = A, B, and C):

$$L_A = \frac{1 - e^2}{e^2}\left[\frac{1}{2e}\ln\left(\frac{1+e}{1-e}\right) - 1\right]$$
$$L_B = L_C = \frac{1 - L_A}{2} \tag{5.222}$$

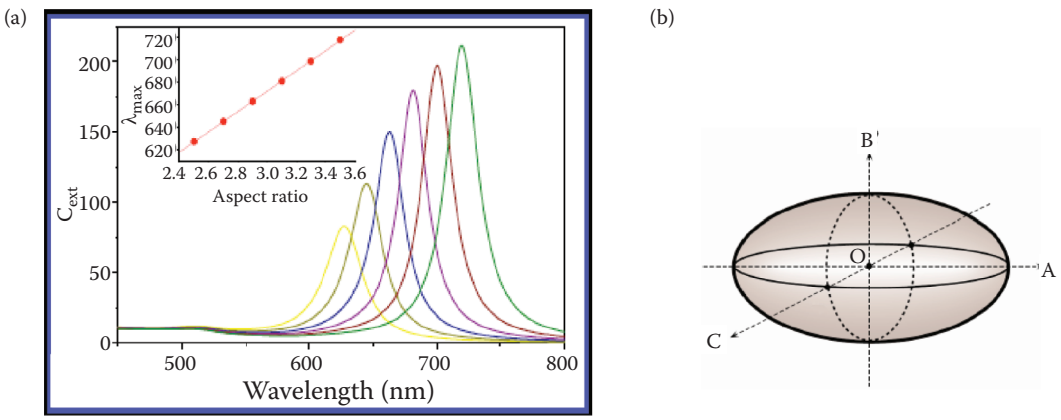


FIGURE 5.73 (a) When the aspect ratio *R* (*A*/*B*) increases, the absorption maximum of the longitudinal band is greatly red-shifted. (b) Ellipsoidal particle A > B = C.

The geometry factors L_A and L_B are connected to the aspect ratio R = A/B via:

$$e = \sqrt{1 - \left(\frac{B}{A}\right)^2} \qquad (5.223)$$

From Equation 5.221, the resonance condition is given as:

$$\varepsilon' = \varepsilon_d \left(1 - \frac{1}{L}\right) \qquad (5.224)$$

so that, in air ($\varepsilon_d = 1$), the resonance frequency is $\omega_p L$.

Another type of nanoparticle that is studied heavily is a dielectric sphere coated with a nanoshell of a metal like gold. Core materials (e.g., silica) have a radius between 30 and 250 nm, and the shell thickness is in the 10–30-nm range. In 1998, Naomi Halas at Rice University demonstrated that one could create such nanostructures by growing a thin layer of gold onto a spherical silica nanoparticle.[14] The thinness of the metal shell results in a substantial redshift of the plasmon resonance. This is caused by increased electron scattering and thus an increase in the damping constant in the metal dielectric constant. Nanoshells are more sensitive than simple nanoparticles with a gain of x = a/b, where *a* is the radius of the particle without shell and *b* is the radius of the particle with shell (Figure 5.74).

The resonance frequency of a nanoshell is a strong function of geometry, and more resonant frequencies are possible. Without going into details, the condition for resonance in this case is given as:

$$\varepsilon'(\text{shell}) + 2\varepsilon_d = 0 \qquad (5.225)$$

and the resonance frequency is:

$$\omega_{\text{nanoshell},+/-} = \frac{\omega_p}{\sqrt{2}}\left[1 \pm \frac{1}{2L+1}\sqrt{1 + 4L(L+1)x^{2L+1}}\right]^{\frac{1}{2}}$$

in the case of a sphere L = 1 and we obtain

$$\omega_{\text{nanoshell},+/-} = \frac{\omega_p}{\sqrt{2}}\left[1 \pm \frac{1}{3}\sqrt{1 + 8x^3}\right]^{\frac{1}{2}} \qquad (5.226)$$

The plasmon energies of a nanoshell depend on the aspect ratio x (= a/b), so the plasmon energies of nanoshells can be tuned (Figure 5.74b)!

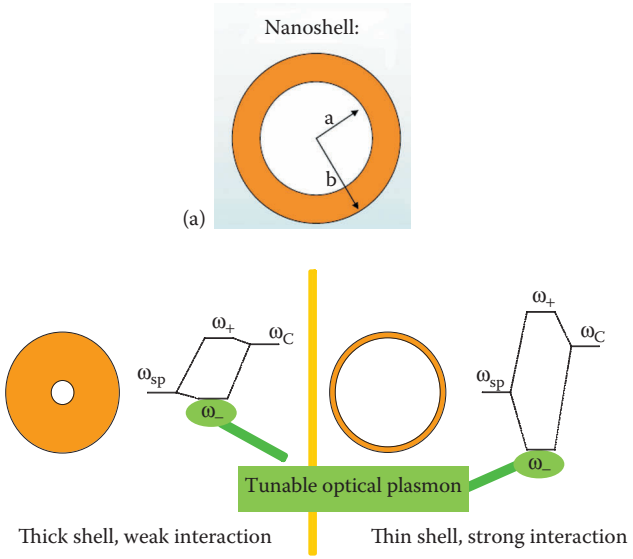


FIGURE 5.74 The resonance frequency of a nanoshell is a strong function of geometry, and more resonant frequencies are possible. The ratio of a/b = x in Equation 5.226 (a). The plasmon energies of a nanoshell depend on this aspect ratio, so the plasmon energies of nanoshells can be tuned (b).

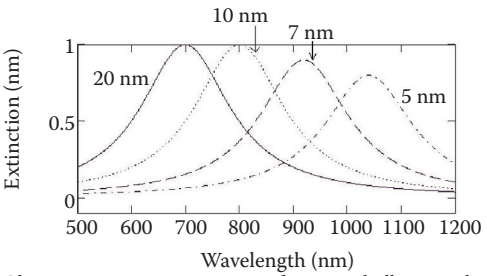
For fabrication, a dielectric sphere is coated with a layer of amines, which binds 1–2-nm gold colloids from suspension. This is followed by a chemical treatment with $HAuCl_4$ in the presence of formaldehyde. This results in an additional layer of gold (Figure 5.75).

Younan Xia of the University of Washington (Seattle, WA) described the synthesis of gold nanocages.[15] Like nanoshells, these nanoparticles have tunable plasmon resonances because of the presence of a dielectric core. In nanocages, however, the core is empty rather than composed of silica. Gold nanocages are grown over a silver nanocube template. The monodisperse nanocubes are made by a polyol process, such as reduction of silver nitrate in ethylene glycol in the presence of poly(vinylpyrrolidone). These template particles are then exposed to a solution of gold ions, which replace the more reactive silver on the nanoparticle surface. Gold grows around the silver nanocube surface, and the silver is dissolved into solution, leaving gold nanocages.

In Table 5.6 we have compiled values for the resonance frequency and resonance condition for a variety of configurations.

Surface Plasmon Applications II

We reviewed some plasmonics applications under "Surface Plasmon Applications I." Here, having



Plasmon resonance extinction for core-shell nanospheres of various shell thicknesses. Core radius is 60 nm.

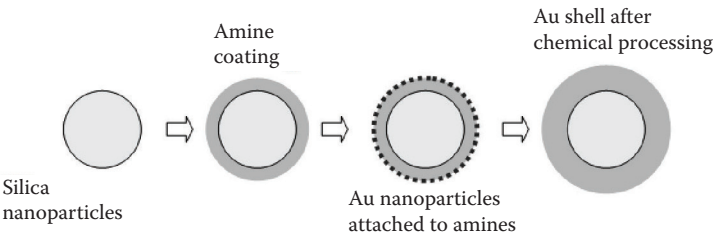


FIGURE 5.75 In general, as the shell thickness decreases, the resonance shifts to longer wavelengths. Vial on the left has solid gold colloids. Others have colloids with metal nanoshells with decreasing thickness. Vial on left absorbs IR. (From Dr. Anup Sharma.)

TABLE 5.6 Resonance Frequency and Resonance Condition for a Variety of Configurations

Material	Resonance Condition	Resonance Frequency
Bulk metal	$\varepsilon' = 0$	$\varepsilon' = 1 - \frac{\omega_p^2}{\omega^2} = 0,\ \omega_{\text{Bulk plasmon}} = \omega_p$
Planar surface	$\varepsilon' = -\varepsilon_d$ ($\varepsilon_d = 1$ in air)	$\varepsilon' = 1 - \frac{\omega_p^2}{\omega^2} = -\varepsilon_d$ or $\omega_{\text{Surface plasmon}} = \frac{\omega_p}{\sqrt{2}}$
Sphere	$\varepsilon' + 2\varepsilon_d = 0$ ($\varepsilon_d = 1$ in air)	$\varepsilon' \approx 1 - \frac{\omega_p^2}{\omega^2} = -2\varepsilon_d$ or $\omega_{\text{Nanosphere}} = \frac{1}{\sqrt{3}}\omega_p$
Ellipsoid	$\varepsilon' = \varepsilon_d\left(1 - \frac{1}{L}\right)$	$\varepsilon' \approx 1 - \frac{\omega_p^2}{\omega^2} = \varepsilon_d\left(1 - \frac{1}{L}\right)$ or $\omega_{\text{Ellipsoid}} = L\omega_p$
Cavity		$\omega_{\text{cavity}} = \omega_p\sqrt{\frac{L+1}{2L+1}}$ In the case of a spherical cavity: $\omega_{\text{cavity}} = \omega_p\sqrt{\frac{2}{3}}$
Nanoshell	$\varepsilon'(\text{shell}) + 2\varepsilon_d = 0$	$\omega_{\text{nanoshell},+/-} = \frac{\omega_p}{\sqrt{2}}\left[1 \pm \frac{1}{2L+1}\sqrt{1 + 4L(L+1)x^{2L+1}}\right]^{\frac{1}{2}}$ In the case of a sphere: $\omega_{\text{nanoshell},+/-} = \frac{\omega_p}{\sqrt{2}}\left[1 \pm \frac{1}{3}\sqrt{1 + 8x^3}\right]^{\frac{1}{2}}$

ε' is the real part of the dielectric constant of the metal; ε_d is the dielectric constant of the contacting dielectric (= 1 in the case of air as the dielectric); *L* is the geometrical factor between 0 and 1, for a sphere L = 1 and x = a/b.

acquired some additional insights in metal nanoparticle behavior, we come back to two nanoparticle-based applications, i.e., resonant particle detection and plasmon-based circuits.

Several properties make resonant nanoparticles unique for sensor applications:

- They are very bright, making them easy to detect. The reason for this brightness is their large scattering cross-section.
- They are small enough to diffuse through biological tissues.
- They can easily be derivatized (linking to organic molecules).
- They suffer no photobleaching, making long illumination possible.
- Most of the sensor applications of localized plasmon waves are based on the EM field enhancement in the vicinity of the metal nanoparticle surface.
- The resonance can be tuned (multilayer/core-shell) for parallel detection.

Today the performance of electronic circuits is becoming rather limited when digital information needs to be sent from one point to another. With photons traveling much faster than electrons, photonics potentially offers an effective solution to this problem by implementing optical communication systems based on optical fibers and photonic circuits. Unfortunately, as we saw earlier in this chapter, traditional optical waveguides involve structures that cannot be made much smaller than $\lambda/2$ because of diffraction. Furthermore, waveguiding along bent guides is very lossy (see Figure 5.49). The latter problem is overcome in photonic bandgap structures (see below). However, the dimensions of the structures are still limited by the wavelength of light. Surface plasmon-based circuits, which merge electronics and photonics at the nanoscale by waveguiding of plasmonic excitation in closely placed metal nanoparticles, may offer a solution to this size-compatibility problem. A chain of closely spaced metallic nanoparticles allows waves of surface plasmons to propagate down the chain as illustrated in Figure 5.76. The typical size of the coupling metal nanospheres is 50 nm. The waves can be either transverse (T) or longitudinal (L) modes, and can have group velocities up to 0.1 c or higher. This approach might potentially be useful for propagating energy along effectively very narrow waveguides, controlling energy flow around corners, and so on. The individual particles in the chain have a specific resonance frequency, and light propagates through these "coupled resonators." The particle center frequency is tunable by material, shape, and size, and the group velocity is tunable by shape and spacing.

These surface excitations can be at any frequency between UV and IR, and so these waveguides combine the less bulky nature of metal waveguides with the high bandwidth of optical waveguides. A weakness of this technique is its high loss (6 dB/μm), and transmission has only been shown over short (~1-μm) distances. It is an active area of research for future applications.

Plasmonics has revived the general enthusiasm for photonics, and one might say that optics has evolved from lenses to fibers to metal nanoparticles

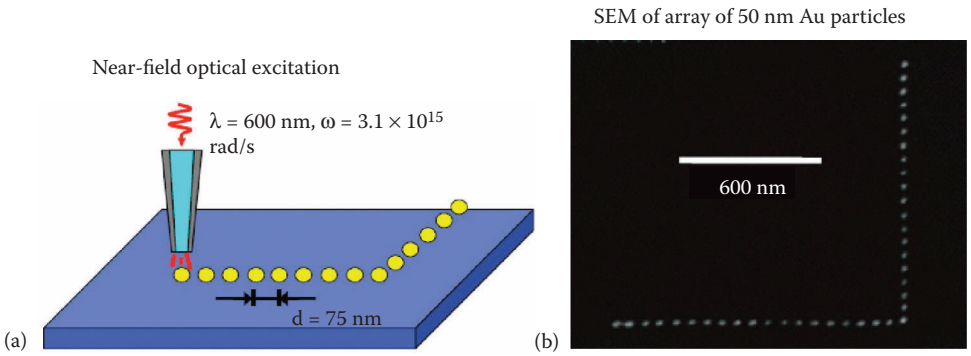


FIGURE 5.76 (a) Propagation of high-frequency EM waves along subwavelength-wide metal waveguides. Plasmonic waveguides. (b) e-beam lithography with liftoff was used to fabricate 50-nm Au nanoparticles on ITO-coated glass.[16]

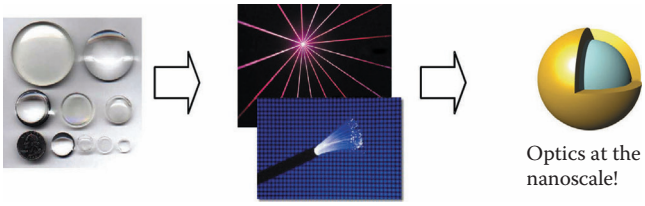


FIGURE 5.77 From lenses to fiberoptics to metal nanoparticles.

(see Figure 5.77). Once we have introduced Raman scattering later in this chapter, we will come back one more time to plasmonics applications. Specifically, we will detail its application in near-field scanning microscopy (NSOM) for subwavelength resolution and in surface enhanced Raman spectroscopy (SERS; see "Surface Plasmon Applications III").

Lorentz Model for Dielectrics

Nonpolar Dielectrics

Whereas isolated atoms have well-defined electronic states representative of the atom, as individual atoms are brought together the atoms start to interact, leading to shifts in the electronic energy levels. At the densities common in solids (~10^{22} atoms/cm^3), significant interaction occurs, causing levels to broaden out into bands. Remembering that real materials correspond to a collection of Drude-Lorentz oscillators all operating at different frequencies, the Drude-Lorentz model works surprisingly well for nonpolar dielectrics. The outer, or valence, electrons of the constituent atoms predominantly determine the characteristics of the optical properties of any solid. The tighter the valence electrons are bound, the higher the resonance frequency, so that these materials may have a transparency range that extends far into the UV. These dielectrics typically show only one resonance in the UV, and at low frequencies they are transparent (see Figure 5.62b for two common dielectrics: Si_3N_4 and SiO_2). Their bandgap, $E_g > 4$ eV, results in high resistivity and low absorption throughout the visible spectrum ($E < E_g$, $\alpha \sim 0$, and $\kappa \lll n$, so that n is real). At the higher frequencies, the refractive index n approaches 1 ($v_g = c$) as there are usually no high ω oscillators in this case, and light then passes through the material as if it were a vacuum.

Besides for their optical properties, dielectrics are also used as insulators or high-k dielectrics.* In Chapter 4 we emphasized the need to find loss-free high-k dielectrics for the next-generation CMOS. Loss is characterized through the quality factor Q. In dielectrics, a low-quality factor Q indicates that a material heavily dissipates energy from the alternating electric field by different types of adsorbing mechanisms. The quality factor (Q) of a resonance $Q = \omega_0/\gamma$ is a measure of the dissipation of the system. The loss tangent $tan\delta$ and quality factor Q of a material are figures of merit for a dielectric and are defined as:

$$\tan\delta = \frac{\varepsilon_r''}{\varepsilon_r'} \tag{5.227}$$

and:

$$Q = \frac{1}{\tan\delta} \tag{5.228}$$

Polar Dielectric Polaritons

In polar ionic materials, e.g., alkali halides such as KCl, the valence electrons are very strongly localized at the negative ion (for KCl, this would be the Cl atom), and hence the optical spectrum contains some atomic-like features, with many resonances. So, whereas the reflectance for a single transition leads to one broad reflection band, as shown in Figure 5.78a the reflectance for the polar solid KCl has many high-energy transitions (Figure 5.78b). At less than 7 eV, KCl is transmissive (KCl is transparent at visible frequencies) and displays a normal dispersion.

In Chapter 3, four types of phonons—LA, longitudinal acoustic; TA, transverse acoustic; LO, longitudinal optical; and TO, transverse optical—were illustrated in Figure 3.159. The transverse and longitudinal "optical phonons" are lattice vibrations that are called *optical phonons* because they are easily excited by light (IR radiation) (e.g., for GaAs, $\hbar\omega_{OP} = 36$ meV). Such phonons can only arise in crystals that have more than one atom in the smallest unit cell. In a polar solid, mechanical

* High-k is another name and symbol used in this context for a high relative dielectric constant.

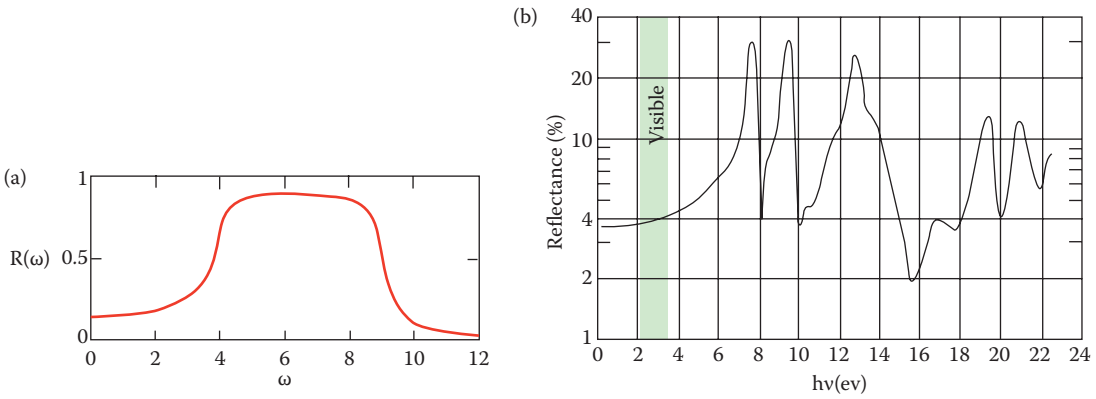


FIGURE 5.78 (a) Calculated reflectance *R* for a single transition leads to one broad reflection band. (b) The reflectance for the polar solid KCl has many high-energy transitions. KCl is transparent at visible frequencies.

vibrations may correspond to a mode where positive and negative ions at adjacent lattice sites swing against each other, creating a time-varying electrical dipole moment **P** that EM radiation can interact with. (EM longitudinal waves do not interact with lattice vibrations.)

The polarization $\mathbf{P}_{\text{lattice}}$ was derived in Chapter 3 as $\mathbf{P}_{\text{lattice}} = \text{Ne}(u_0 - v_0)$ (Equation 3.326), with *e* the effective charge on the atoms in the lattice and u_0 and v_0 the amplitudes of the excursion of atom *M* and atom *m*, respectively. The solutions for amplitudes u_0 and v_0 were calculated as $u_0 = \dfrac{+qE_0/M}{\omega_T^2 - \omega^2}$ and $v_0 = \dfrac{-qE_0/m}{\omega_T^2 - \omega^2}$ (Equation 3.324), where $\omega_T = \sqrt{\dfrac{2K}{\mu}}$ (Equation 3.325) is the resonant frequency of the TO mode at wave vector $\mathbf{k} = 0$, and μ is the reduced mass (see Figures 3.159 and 5.79).

The excitation near ω_T (at $\mathbf{k} = 0$) features the largest amplitude. When in this resonance condition

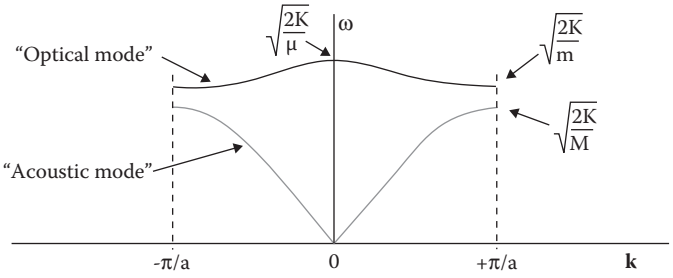


FIGURE 5.79 Two separate "phonon branches" are observed, a low-frequency or "acoustic branch" and a high-frequency "optical branch." The optical branch has a maximum at low wave vector limit: $\mathbf{k} \cong 0$ (or $ka \ll 1$) and the acoustic branch at $\mathbf{k} \cong \pm\pi/a$.

the frequencies and wave vectors of optical and mechanical waves are approximately equal, then the phonon-photon coupling can change the propagation of the waves dramatically. The quantum of the coupled phonon-photon transverse wave field is called a *polariton*. In Chapter 3 we also introduced the total polarization of a material as the sum of the lattice contribution ($\mathbf{P}_{\text{lattice}}$), a bound electronic contribution ($\mathbf{P}_B$), and a free electron contribution ($\mathbf{P}_E$) as:

$$\mathbf{P} = \mathbf{P}_{\text{lattice}} + \mathbf{P}_B + \mathbf{P}_E \tag{3.327}$$

From the expression linking the dielectric constant with the polarization $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P} = (1 + \chi_e)\varepsilon_0\mathbf{E} = \varepsilon(\omega,\mathbf{k})\mathbf{E}$ (Equation 3.37), we derive:

$$\varepsilon_r(\omega) = 1 + \frac{\mathbf{P}}{\varepsilon_0\mathbf{E}} = 1 + \frac{\mathbf{P}_{\text{lattice}}}{\varepsilon_0\mathbf{E}} + \frac{\mathbf{P}_B}{\varepsilon_0\mathbf{E}} + \frac{\mathbf{P}_E}{\varepsilon_0\mathbf{E}} \tag{5.229}$$

and ignoring, for the moment, the free electron contribution:

$$\varepsilon_r(\omega) = 1 + \frac{\mathbf{P}_B}{\varepsilon_0\mathbf{E}} + \frac{\mathbf{P}_{\text{lattice}}}{\varepsilon_0\mathbf{E}} = 1 + \frac{\mathbf{P}_B}{\varepsilon_0\mathbf{E}} + \frac{\text{Ne}(u_0 - v_0)}{\varepsilon_0\mathbf{E}} \tag{5.230}$$

Because $u_0 - v_0 = \dfrac{eE_0}{\mu(\omega_T^2 - \omega^2)}$ and replacing $1 + \dfrac{\mathbf{P}_B}{\varepsilon_0\mathbf{E}}$ (for the bound electrons) with $\varepsilon_r = 1 + \chi_B(\omega)$ (Equation 5.115; we ignore the prime symbol here, but we are talking about the real part of the dielectric constant), we rewrite Equation 5.230 as:

$$\varepsilon_r(\omega) = 1 + \chi_B + \frac{\text{Ne}^2}{\varepsilon_0\mu(\omega_T^2 - \omega^2)} \tag{5.231}$$

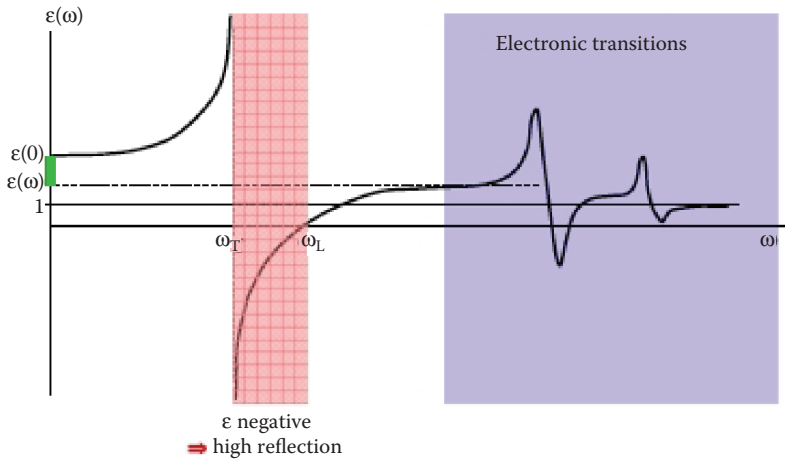


FIGURE 5.80 Dielectric constant as a function of ω for a polar material.

When considering low frequencies, this equation simplifies to:

$$\varepsilon_r(\omega) - \varepsilon_r(\infty) = \frac{Ne^2}{\varepsilon_0 \mu(\omega_T^2 - \omega^2)} \qquad (5.232)$$

where $\varepsilon_r(\infty)$ is the high-frequency dielectric constant. Thus, the difference between the low- and high-frequency dielectric constant $\varepsilon_r(\omega) - \varepsilon_r(\infty) = \frac{Ne^2}{\varepsilon_0 E\mu(\omega_T^2 - \omega^2)}$ is related to charge density and mass. It is important to recognize that $\varepsilon_r(\infty)$ is not truly the value of the dielectric constant at an infinite frequency because that value is always 1. What it means is a value at frequencies well above ω_T, yet well below any electronic transitions (see Figure 5.80).

Thus, the behavior sketched of this type of resonance is similar to that of the Lorentz resonator but without the damping term (see Equation 5.195).

We can use Equation 5.232 to find a value for the low-frequency dielectric constant as $\varepsilon_r(0) = \varepsilon_r(\infty) + \frac{Ne^2}{\varepsilon_0 \mu \omega_T^2}$. The same equation can be used to find the frequency ω_L, at which $\varepsilon_r(\omega) = \varepsilon_r(\infty) + \frac{Ne^2}{\varepsilon_0 E\mu(\omega_T^2 - \omega^2)} = 0$

The latter expression with $\frac{Ne^2}{\varepsilon_0 \mu} = [\varepsilon_r(0) - \varepsilon_r(\infty)]\omega_T^2$ can be modified as:

$$\varepsilon_r(\omega) = \varepsilon_r(\infty) + \frac{[\varepsilon_r(0) - \varepsilon_r(\omega)]\omega_T^2}{(\omega_T^2 - \omega^2)} \qquad (5.233)$$

This can be rewritten as:

$$\varepsilon_r(\omega) = \frac{\varepsilon_r(0)\omega_T^2 - \varepsilon_r(\infty)\omega^2}{(\omega_T^2 - \omega^2)} \qquad (5.234)$$

so that we find that $\varepsilon_r(\omega) = 0$ when $\varepsilon_r(0)\omega_T^2 = \varepsilon_r(\infty)\omega^2$. The particular frequency ω where this happens is labeled ω_L, for longitudinal optical (LO) phonons. Similar to the case of plasma oscillations with a plasma oscillation frequency ω_p, ω_L is the frequency for longitudinal wavelength lattice oscillations (LO phonon frequency). We can rewrite $\varepsilon_r(0)\omega_T^2 = \varepsilon_r(\infty)\omega_L^2$ at ω_L also as:

$$\frac{\omega_T^2}{\omega_L^2} = \frac{\varepsilon_r(\infty)}{\varepsilon_r(0)} \qquad (5.235)$$

This is known as the *Lyddane-Sachs-Teller relationship*, which works well for polar diatomic solids.

Example 5.9:

For GaAs: $\frac{\omega_L}{\omega_T} = 1.07$ and $\sqrt{\frac{\varepsilon_r(0)}{\varepsilon_r(\infty)}} = 1.08$

And for KBr: $\frac{\omega_L}{\omega_T} = 1.39$ and $\sqrt{\frac{\varepsilon_r(0)}{\varepsilon_r(\infty)}} = 1.38$

The resonant frequency ω_T is called the *Restrahlen frequency*, and the frequency at which the dielectric constant vanishes is ω_L [notice that ω_L is always larger than ω_T; hence $\varepsilon_r(0) > \varepsilon_r(\infty)$].

In Figure 5.81a we show the real and imaginary parts of the refractive index of AlSb, a polar material.

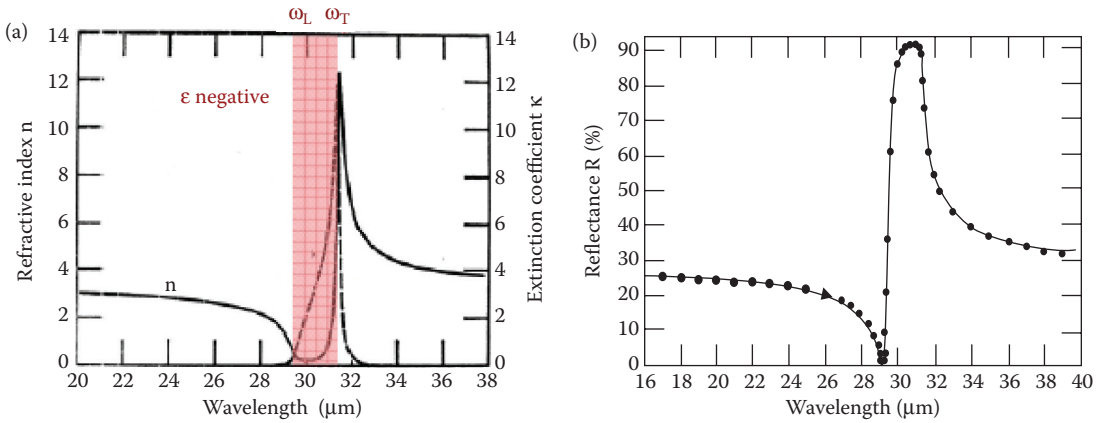


FIGURE 5.81 (a) The real (*n*) and imaginary (κ) parts of the refractive index of AlSb, a polar material. (b) Reflectance spectrum of AlSb with the Restrahlen band. (*R*, in the absence of damping, must have unity in the region $\omega_L < \omega < \omega_T$.)

Although in Equation 5.232 we have assumed that damping was not important, the damping in a real case cannot be too small as κ appears large here. Just like for electronic transitions, the reflectance *R*, in the absence of damping, must be one in the region $\omega_L < \omega < \omega_T$. This region is called the *Restrahlen band*, where light cannot propagate, so it must be reflected as shown in Figure 5.81b.

Summary: Generalized Dispersion Relation

Substituting the time-varying EM wave $\mathbf{E}(\mathbf{r},t) = \mathbf{E}_m e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$ in Maxwell's wave equation, expressed in terms of the velocity of light *c* and the dielectric constant $\varepsilon_r(\omega,\mathbf{k})$ $c^2\nabla^2\mathbf{E}(\mathbf{r},t) = \varepsilon_r(\omega,\mathbf{k})\dfrac{\partial^2\mathbf{E}(\mathbf{r},t)}{\partial^2 t}$ (because most materials have $\mu \approx 1$ and are nonmagnetic), one obtains an expression for the dispersion of EM waves in terms of the dielectric constant:

$$\varepsilon_r(\omega,\mathbf{k})\omega^2 = c^2\mathbf{k}^2 \qquad (5.88)$$

where we have ignored the μ term because most materials have $\mu \approx 1$ and are nonmagnetic.

In free space $\varepsilon_r(\omega,\mathbf{k}) = 1$, and Equation 5.88 then simply reduces to $\mathbf{k} = \omega/c$, the so-called *light line*.

For the case that $\varepsilon_r(\omega,\mathbf{k})$ is real and >0, **k** is also real, and transverse EM waves propagate with a group velocity of $v_g = \dfrac{d\omega}{d\mathbf{k}} = \dfrac{c}{\sqrt{\varepsilon_r}}$ (Equation 5.87 with $\mu = 1$). With $\varepsilon_r(\omega,\mathbf{k})$ real but <0, **k** is imaginary, and the wave is damped with a characteristic skin depth δ (Equation 5.215). With $\varepsilon_r(\omega,\mathbf{k})$ complex, **k** is complex as well, and waves are damped again. At $\varepsilon_r(\omega,\mathbf{k}) = 0$, longitudinally polarized waves are possible (free electron plasmons). In longitudinal oscillations the displacement is in the same direction as the wave motion, and Drude's free electron gas (FEG), with an AC field applied, can exhibit such collective longitudinal oscillations because the displacement of all the electrons against the ion bodies of the material induces a dipole moment and an electric field opposing that displacement, as was shown in Figure 3.12.

At optical frequencies $\omega^2\tau^2 >>> 1$ (so that $1/\tau = \gamma$ tends to zero) and $\varepsilon_r(\omega) \cong \varepsilon_r'(\omega) = 1 - \dfrac{\omega_p^2}{\omega^2}$ (see Equation 3.42). This expression is drawn in Figure 5.82 (see also Figure 3.13) with a region (1) where the metal is transparent ($\omega > \omega_p$ so that **k** is real and there is no attenuation), and a region (2) where the waves are totally reflected ($\omega < \omega_p$ so that **k** contains an imaginary component and the wave is reflected) to give us the typical shine of a metal.

Substituting $\varepsilon_r'(\omega,\mathbf{k}) \approx 1 - \dfrac{\omega_p^2}{\omega^2}$ (Equation 3.44) in the dispersion expression (Equation 5.88), we find:

$$\omega^2 - \omega_p^2 = c^2\mathbf{k}^2 \qquad (5.150)$$

This expression is plotted in Figure 5.83, where the reflection [corresponding to the region (2) of Figure 5.82] is shown as a forbidden frequency gap. This forbidden frequency gap falls in the region of $0 < \omega \leq \omega_p$. At those frequencies light incident on the medium does not propagate but is totally reflected, giving us shiny metals (imaginary **k**).

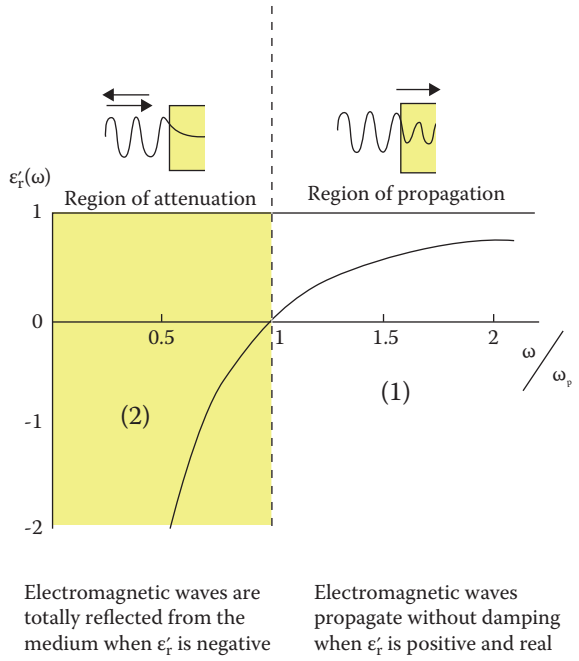


FIGURE 5.82 The dielectric function of a free electron gas versus frequency in units of the plasma frequency ω_p. Waves propagate without damping through the material when ε_r is positive and real. (Based on Kittel.)

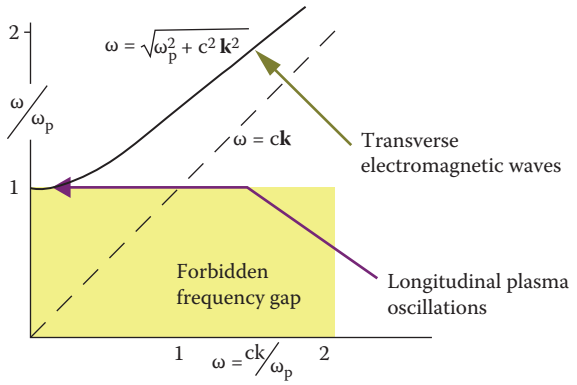


FIGURE 5.83 Dispersion relation for transverse electromagnetic waves in a plasma. The group velocity $v_g = d\omega/d\mathbf{k}$ is the slope of the dispersion curve. The straight line is for $\omega/\mathbf{k} = c$: propagation in free space. The group velocity v_g is less than the velocity of light in vacuum. The phase velocity, $v_p = \omega/\mathbf{k} > c$, is faster than the speed of light, but the latter does not correspond to the velocity of a real physical propagation of any physical quantity. A wave that propagates in the medium with the dispersion relation $\omega^2 = \omega_P^2 + c^2\mathbf{k}^2$ is also shown, as is the region of longitudinal plasma oscillation. The yellow area is the forbidden frequency gap. (Based on Kittel.)

Below the plasma frequency ($\omega < \omega_p$) there is a rapid drop in **E** field that leads to a thin penetration depth of the EM waves—the skin effect.

For $\omega > \omega_p$, the dielectric function is positive and real, and waves propagate through the transparent metal, with the dispersion relation $\omega = \sqrt{c^2\mathbf{k}^2 + \omega_P^2}$. Longitudinal plasma oscillations are shown in Figure 5.83 at $\varepsilon_r' = 0$ ($\omega = \omega_p$). For longitudinal plasma oscillations $\omega_L = \omega_P$ and from Equation 3.44, we then obtain:

$$\varepsilon_r'(\omega,\mathbf{k}) \approx 1 - \frac{\omega_p^2}{\omega^2} = 0 \qquad (3.44)$$

In other words, at the plasmon frequency sustained electron oscillations are possible. For completeness sake, transverse EM waves are also shown for the vacuum situation ($\omega = \mathbf{k}c$—the light line).

In a polar dielectric, the coupling of phonons and photons in a polariton is responsible for the dispersion curves shown as solid curves in Figure 5.84. In Chapter 3, Figure 3.158b, we started by drawing the dispersion of light (ω versus **k**) in a polar solid such as GaAs assuming no interaction of light with phonons $\left(\omega = \frac{c\mathbf{k}}{\sqrt{\varepsilon(\infty)}}\right)$. However, in polar materials, light, close to the vibrational resonance frequencies ($\omega \sim 10^{14}$ rad/s), does interact with the transverse optical phonon modes, and ε_r starts to vary strongly so that ω versus **k** is not a straight line anymore as evident from Figure 5.84.

At large **k** wave vectors ($\mathbf{k} > 10^5$ cm^{-1}), we derived a positive photon-like solution with $\omega = \frac{c\mathbf{k}}{\sqrt{\varepsilon(\infty)}}$ (just like in Figure 3.158b) and a negative phonon-like mode with a singularity at $\omega = \omega_T$. With ω approaching ω_T, the index of refraction increases, and the group and phase velocities of the wave slow down (decreasing slope of the dispersion curve) to a value close to that of the lattice modes (AT mode). The wave now propagates as a strongly coupled polarization-electric field wave, with a velocity characteristic of the lattice transverse acoustic mode. At small **k** vectors ($\mathbf{k} \ll 10^4$ cm^{-1}), we obtain again two solutions: a positive phonon-like mode with $\omega = \omega_L$ (the optical phonon branch—OL), where the dielectric function vanishes, and a negative photon-like mode with $\omega = \frac{c\mathbf{k}}{\sqrt{\varepsilon(0)}}$. There is a forbidden gap established

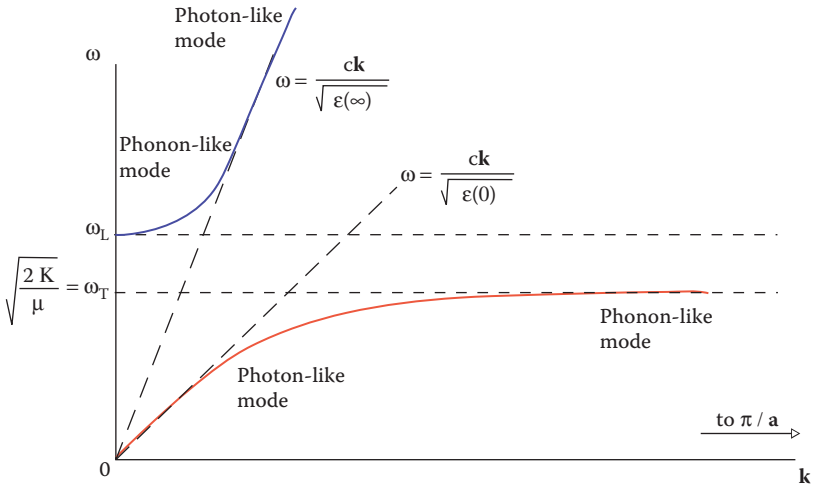


FIGURE 5.84 Dispersion relationship of light in a polar solid illustrating coupled modes of photons and transverse optical phonons. The lines labeled $\omega = \frac{ck}{\sqrt{\varepsilon(\infty)}}$ and $\omega = \frac{ck}{\sqrt{\varepsilon(0)}}$ represent electromagnetic waves in the crystal that are not coupled to the lattice vibration modes.

between the two horizontal lines at ω_T and ω_L. In this gap the EM wave attenuates as $\exp[-|\kappa| x]$, but the absorption is much stronger near ω_T than near ω_L as only the transverse waves couple and resonate strongly.

Lorentz Model for Semiconductors

Introduction

We saw earlier that valence electrons in insulators are relatively strongly bound to the individual atoms in the material, resulting in high resistivity and low absorption throughout the visible spectrum. In contrast for metals, where valence electrons can be considered free, we obtain a high conductivity, a negative dielectric constant at low frequencies, and high absorption and reflection. Semiconductors form an intermediate case between metals and insulators with valence electrons that are generally "somewhat" bound to the atoms in the material. Like insulators, semiconductors come with a very characteristic transmission spectrum with a transparency window for energies below the bandgap: $E < E_g$ (with $\alpha \sim 0$, $\kappa \ll n$, and n is real) and a fundamental absorption edge for $E \sim E_g$ (with $\alpha \neq 0$ and n is complex). The bandgap for semiconductors (<4 eV) is smaller than that of insulators ($E_g > 4$ eV). As a result, at room temperature these materials contain a small concentration of electrons that have broken free from the atoms, resulting in a small conductivity. The example of silicon was shown in Figure 5.62a. Optical spectroscopy is an important experimental tool for the elucidation of band structure. Even the color of a semiconductor reveals something of its band structure. Because electronic transitions from the valence to the conduction band span a fairly large range of energies, semiconductors act as a sort of long-pass filter-only reflecting light with less energy than the bandgap; this gives rise to colors as shown in Table 5.7. As we discovered already, much more is to be gained from optical absorption and reflection spectroscopy of solids than this simple table.

The band model for semiconductors was derived from quantum mechanics in Chapter 3. In this section we detail the various optical processes that can take place in a semiconductor.

Free Carrier Absorption (E < E_g)

EM wave absorption by free carriers in semiconductors results from the excitation of free carriers into available states higher in their respective bands (e.g.,

TABLE 5.7 Color and the Bandgap

Bandgap (eV)	Color	Examples
>3.0	White	ZnO
3.0–2.5	Yellow	CdS
2.3–2.5	Orange	GaP
1.8–2.3	Red	HgS
<1.8	Black	CdS

electrons higher into the conduction band). Free carrier absorption may be quite different for electrons than for holes; this is mainly because there are multiple valence bands available for holes, but there is only one conduction band for electrons. Free electrons located at wave vector $\mathbf{k} = 0$, at the bottom of a single parabolic conduction band, can reach higher-energy states only if their momentum is increased. As a consequence, in n-type semiconductors, conduction band electrons may only be excited by the simultaneous absorption of a photon and the absorption or the emission of a phonon, to conserve momentum (see also direct and indirect bandgap transitions, discussed next). With p-type semiconductors, on the other hand, one can have direct (i.e., no phonons needed) transitions between the different valence bands (heavy- and light-hole bands). As a result, holes produce stronger free carrier absorption than conduction band electrons. The Drude model is sufficient to describe the free electron absorption process. The free electron density in semiconductors is often small, say $10^{14}/\text{cm}^3$, which is about 10^8 times smaller than in a metal, so the plasma frequency ω_p of semiconductors is also much smaller because $\omega_p = \sqrt{\frac{ne^2}{m_e\varepsilon_0}}$ (Equation 3.41), about 10^4 times lower than for a metal! For weakly absorbing materials such as semiconductors (relatively few conduction electrons), the electric susceptibility tensor χ' and χ'' <<<< 1, and one obtains for frequencies well above the plasma frequency:

$$\varepsilon_r'(\omega, \mathbf{k}) \approx 1 - \frac{\omega_P^2}{\omega^2} \tag{3.44}$$

$$\varepsilon_r''(\omega, \mathbf{k}) \approx \frac{\omega_P^2 \gamma}{\omega^3} \tag{3.45}$$

and because the absorption coefficient is given as $\alpha(\omega) \approx \frac{\omega}{c}\varepsilon_r''(\omega)$ (see Equations 5.115 and 5.121), we derive:

$$\alpha(\omega) \approx \frac{\omega_P^2 \gamma}{\omega^3} \approx \frac{\lambda^2 \gamma}{c\lambda_P^2} \tag{5.236}$$

This describes a λ^2 dependence for free carrier absorption as a result of electrons. This type of dependence of α on λ is commonly seen in semiconductors, where dopant densities are typically 10^{16}–10^{19} cm^{-3} (ω_p in the 10^{13} rad/s range and λ_p > ~10 μm) as compared with ~10^{22} cm^{-3} (with ω_p in the 10^{16} rad/s range and λ_p at UV wavelengths) in metals. Infrared absorption in Si is approximately proportional to λ^2. This type of absorption is also seen with metals at very high frequencies, frequencies well above the plasmon frequency. At these high frequencies ($\omega > \omega_p$) the dielectric constant n_c becomes real, and the metal acts now like a dielectric.

The form of the free-hole absorption in p-type semiconductors is more difficult to describe because it depends on the valence band structures and the hole density. In semiconductors with equal numbers of electrons and holes, the free-hole absorption dominates. As we saw above and in Chapter 3 this is because there are multiple valence bands available for holes, but there is only one conduction band for electrons. Free electrons located at wave vector $\mathbf{k} = 0$, at the bottom of a single parabolic conduction band, may reach higher-energy states only if their momentum is increased. With p-type semiconductors, one can have such direct (i.e., no phonons needed) transitions between the different degenerate valence bands (heavy and light hole bands) because they occur at the same $\mathbf{k}$. Because no momentum or $\mathbf{k}$ change is required, holes produce stronger free carrier absorption than conduction band electrons.

Light Absorption for Direct and Indirect Bandgap Semiconductors ($E > E_g$)

The bandgap for direct bandgap and indirect bandgap semiconductors in $\mathbf{k}$-space was first shown in Figure 3.116; in the case of silicon, the valence band maximum (VBM) occurs at $\mathbf{k} = 0$ (Γ point). However, the conduction band minimum (CBM) occurs at *L*. Thus, the VBM and the CBM occur at different points in the Brillouin zone, and this makes silicon an indirect bandgap material. GaAs is a direct bandgap semiconductor because the VBM and CBM occur at the same point in the Brillouin zone. The dominant feature of the absorption spectrum of a semiconductor is the band-to-band or interband transition: photon absorption promotes an electron from the valence into the conduction in so-called

band-to-band absorption. In such a transition both energy and momentum must be conserved. In other words, for the energy balance we write:

$$E_i + h\nu = E_f \tag{5.237}$$

with E_i the initial and E_f the final energy and $h\nu$ the photon energy, and similarly for the momentum balance:

$$\mathbf{k}_i + \mathbf{k}_\nu = \mathbf{k}_f \tag{5.238}$$

with $\mathbf{k}_\nu$ a photon wave vector. For a bandgap E_g of 1 eV, one calculates for this photon wave vector a value of:

$$\mathbf{k}_\nu = \frac{\mathbf{P}}{\hbar} = \frac{E/c}{\hbar}$$

$$\approx \frac{1\,ev\,(1.60 \times 10^{-19}\ J/ev)}{(3.00 \times 10^{8}\ m/s)(1.05 \times 10^{-34}\ Js)} \approx 5 \times 10^{6}\ m^{-1}$$

a quantity very small compared with a typical vertical (or direct transition) because $\mathbf{k}_i$ is of the order:

$$\mathbf{k}_i \approx \frac{2\pi}{a} \approx \frac{2\pi}{3.00 \times 10^{-10} m} \approx 2 \times 10^{10}\ m^{-1}$$

In other words, a photon in the visible range of the spectrum comes with a momentum vector much smaller than that of an electron. As a consequence, direct bandgap transitions (with an absorption coefficient α typically 10^6–10^8 m^{-1}) appear as vertical lines in an E-$\mathbf{k}$ diagram because Equation 5.238 reduces to:

$$\mathbf{k}_i = \mathbf{k}_f \text{ or } \Delta\mathbf{k} = 0 \tag{5.239}$$

For an indirect bandgap transition, momentum can only be conserved by absorption or emission of a phonon (lattice vibration), and we have:

$$\mathbf{k}_i \neq \mathbf{k}_f \text{ or } \Delta\mathbf{k} \neq 0 \tag{5.240}$$

and for the energy balance:

$$E_i + h\nu \pm h\nu_{ph} = E_f \tag{5.241}$$

with $h\nu_p$ the phonon energy, + for absorption, and – for emission. For the momentum conservation, we write:

$$\mathbf{k}_i + \mathbf{k}_{ph} = \mathbf{k}_f \tag{5.242}$$

Estimating a typical phonon energy, we obtain:

$$h\nu_p \approx 0.07\ eV$$

The calculated value, although generally much less than the bandgap energy, is significant enough to change the electron's wave vector. Phonons have a momentum that is four to five orders of magnitude higher than for photons, so they can help in indirect transitions where photons cannot. Indirect absorption is a second-order process as a phonon intervention is required, so the process has a much lower transition probability (10^{-5}–10^{-6} s^{-1}) than direct absorption. The absorption α at a given frequency is an integral over all the transitions in the Brillouin zone that satisfy either Equations 5.237 and 5.238 for a direct transition or Equations 5.241 and 5.242 for an indirect transition. The need of phonons to assist in absorption and emission of light in indirect bandgap semiconductors like Si makes for weak absorption and emission of light, and most optoelectronic applications are easier implemented in direct bandgap semiconductors like GaAs. We encountered phonons first in Chapter 3, where we saw that they represent mechanical deformation waves traveling through the lattice. Compression waves do constitute a periodic deformation potential of the bandgap, and it is this periodic narrowing and widening of the bandgap that causes electron scattering and momentum transfer to photons. Electron-phonon interaction is called *Frölich interaction*.

A direct gap semiconductor features a sharp onset of absorption when the photon energy is equal to the bandgap. Indirect gap semiconductors like silicon and germanium do not have such a sharp onset of absorption as illustrated in Figure 5.85, where we compare the absorption α for Si and GaAs together with that of some other important semiconductors. Compared with atomic absorption peaks, absorption bands in crystals are broad and seem quite featureless when the photon energy is greater than the bandgap. As we discovered earlier in this chapter, despite this seemingly bland appearance, significant information on the detailed band structure of materials can be obtained from absorption and reflection spectroscopy (e.g., by taking the derivative of

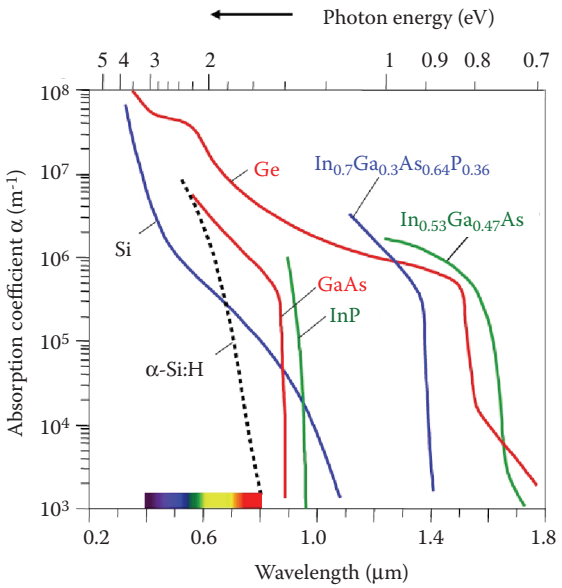


FIGURE 5.85 Absorption α of Si and GaAs and some other important semiconductors are compared.

the reflectance with respect to wavelength, temperature, electric field, pressure, etc., in so-called *modulation spectroscopy*).

A very general effect of quantum confinement in semiconductors, we saw in Chapter 3, is a widening of their optical bandgap, leading to characteristic blue shifts. Quantum confinement also changes the density of states (DOS) dramatically, and absorption and emission of light by a semiconductor depends on both the bandgap width and the availability of states. As the energy levels are packed in narrower and narrower energy ranges near the bandgap of quantum-confined structures, the optical transition strengths for absorption and emission improve the stronger the confinement. In addition to the interband transition, quantization of semiconductors also leads to new transitions between different sub-bands within the conduction band (minibands) (Figure 5.86). These transitions are called *intraband*, *inter-sub-band transitions*, or *Cerenkov radiation*, and they are finding more and more important applications, for example, in the quantum cascade lasers discussed at the end of this chapter. We also saw how the transition probability in indirect gap semiconductors increases with increasing confinement: confinement relaxes some selection rules, thus allowing emission (see next section)

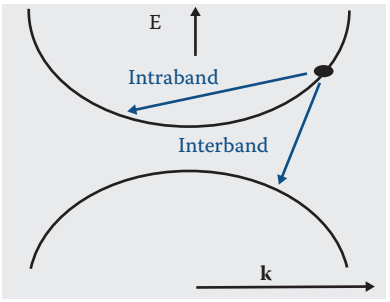


FIGURE 5.86 Interband and intraband absorption. The latter is also known as *Cerenkov radiation*.

from nanoparticles that are strongly forbidden in large devices.

Emission of Photons

Luminescence comprises the generation of any light that is not induced by heat (as in incandescence) (see also Chapter 7). In photoluminescence, emission of light from a semiconductor is stimulated by using light (say, from a laser) with a photon energy $E > E_g$. In other words, the pumping wavelength is shorter than the bandgap wavelength. Hot electrons and holes are generated in this process and relax back to the bottom and the top of their bands, respectively, in a process called *thermalization*, as illustrated in Figure 5.87a for hot electrons. In this process the carriers take on a thermal distribution over the available energy levels according to statistical mechanics (Fermi-Dirac statistics because they are fermions). Subsequently, emission of light occurs: in the case of a direct bandgap, emission is strong at the bandgap energy E_g. For an indirect bandgap material, the emission (like the absorption) is a low-probability process because a transition from Γ to the X or L point requires a three-particle interaction (one photon, one electron, and one phonon) as shown in Figure 5.87b. Although the radiative lifetime is long for an indirect bandgap material, the quantum efficiency is low. Because the photon has a very small momentum $\mathbf{k}$ (~0), as a result of its long wavelength compared with that of an electron, the optically induced electronic transition basically requires that $\Delta\mathbf{k} = 0$. As a consequence, emission from an indirect bandgap material like Si is forbidden by this selection rule. Therefore, bulk Si is not a luminescent emitter, but the direct bandgap semiconductor GaAs is.

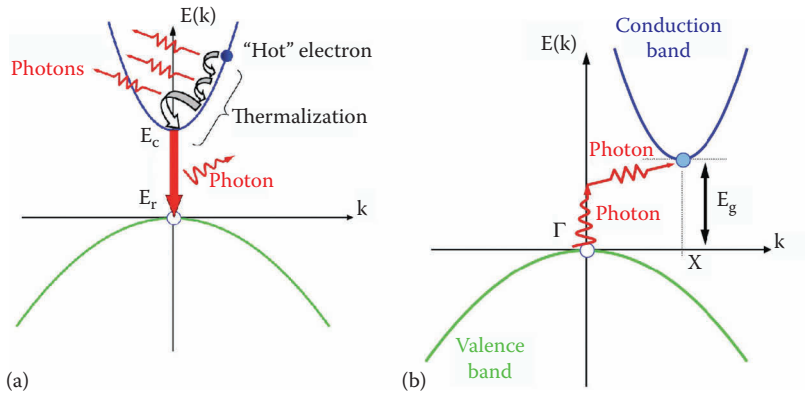


FIGURE 5.87 (a) Thermalization of hot electrons in a direct bandgap semiconductor and subsequent emission of a photon in photoluminescence. (b) For an indirect bandgap material the emission is low probability because a transition from Γ to the *X* or *L* point requires a three-particle interaction (one photon, one electron, and one phonon).

Excitons

Excitons in Large Structures We have assumed so far that when an electron-hole pair is created there are no interactions between these charge carriers in a solid. In reality, electrons and holes interact with each other through the Coulomb interaction, and an electrostatically bound electron-hole pair forms a neutral entity called an *exciton*. The latter may be compared with an electron bound to a proton to form a neutral hydrogen atom. An exciton represents an elementary excitation of a semiconductor and can be created by either a photon or an electron. Although overall neutral, an exciton carries a dipole moment and can move through the crystal and transport energy. In an exciton, the electron and hole move together as a single uncharged particle. All excitons are unstable with respect to their ultimate recombination process, in which the electron drops into the hole (Figure 5.88).

The exciton wave vector $\mathbf{k}_{ex}$ is given by:

$$\mathbf{k}_{ex} = \mathbf{k}_e - \mathbf{k}_h \tag{5.243}$$

with $\mathbf{k}_e$ the electron vector and $\mathbf{k}_h$ the hole vector. Absorption occurs only at $\mathbf{k}_{ex} = 0$, to preserve momentum and because of energy conservation:

$$hv = E_{ex} \tag{5.244}$$

We can treat excitons as chargeless particles capable of diffusion in a solid, but in the ultimate confinement in a molecule they can also be regarded as excited states of a single molecule. The exciton binding energy $E_n(\mathbf{k})$ is the energy required to separate the pair. In a molecule, excitons come with a large binding energy in the 1-eV range and feature a small radius of 10 Å. These are called *Frenkel excitons*, typical for organic materials. The less confined, wandering excitons in a solid are called *Mott-Wannier excitons*; they have a smaller binding energy $E_n(\mathbf{k})$ in the 5–20-meV range and a radius of about 100 Å (>>> the typical lattice constant *a*). This binding energy of Mott-Wannier excitons is of the order of kT at room temperature, and as a result exciton lines in an absorption spectrum are generally not well resolved from the band-edge absorption unless the sample is cooled. The two types of excitons are compared in Figure 5.89. Frenkel developed the theory of an exciton as a small and tightly bound entity in 1931.[17,18] E.F. Gross in 1951 discovered the Mott-Wannier excitons—large and weakly bound—in the

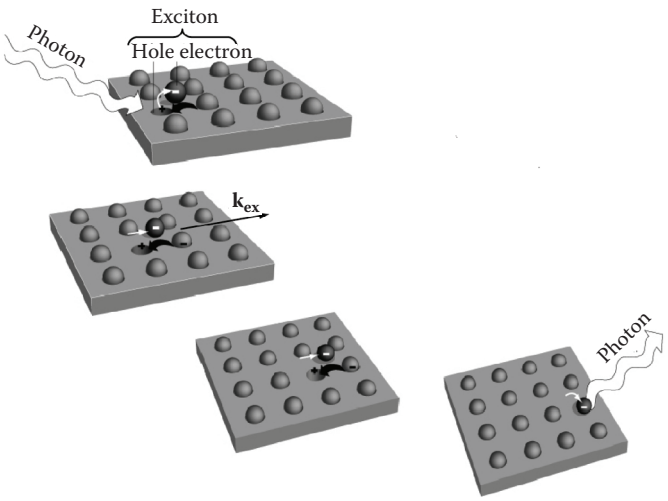


FIGURE 5.88 Life and death of an exciton.

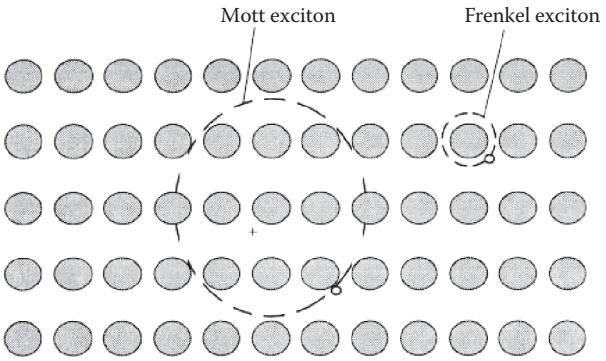


FIGURE 5.89 Frenkel excitons, typical for organic materials, and the wandering excitons in an inorganic solid, Mott-Wannier excitons, are compared.

optical spectrum semiconductor crystals of cuprous oxide.[19]

Because there is an attraction between the electron and the hole in an exciton, the photon energy required to create an exciton is lower than the bandgap energy, as illustrated in Figure 5.90. This results in absorbance and reflectance spectra revealing some structure just below the energy gap.

The exciton energy and radius can be derived from the Bohr model for hydrogen (a method we also used in Chapter 4 to calculate dopant energy levels). The Bohr model for the hydrogen atom gave us the following expression for the ground state energy (n = 1), also sometimes called the *Rydberg energy* R_y:

$$R_y = E_0 = -\frac{e^2}{8\pi\varepsilon_0 a_0} = -13.6 \text{ eV or } 21.8\ 10^{-19} \text{ J} \quad (3.91)$$

where a_0 is the Bohr radius.

The Rydberg energy is the energy required to remove an electron from the hydrogen atom. Applying the Bohr model to excitons, we modify Equation 3.91 by introducing the reduced mass $m_r^* = \frac{m_h^* m_e^*}{m_h^* + m_e^*}$,

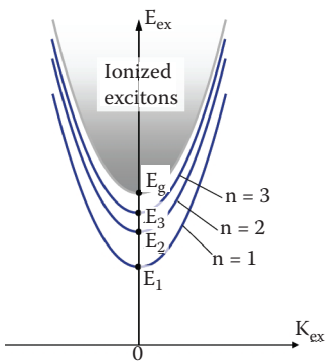


FIGURE 5.90 Exciton energy levels.

as well as ε_r, the dielectric constant of the semiconductor, to obtain:

$$R_y = 13.6 \frac{m_r^*}{n^2 m_e} \left(\frac{1}{\varepsilon_r}\right)^2 \quad (5.245)$$

A bound exciton can only be found when the thermal energy $kT < R_y$. If the thermal energy is larger than this quantity, all holes and electrons behave as separate quantities. The energy levels referred to the top of the valence band are then given (for the kinetic energy term in this expression, see Chapter 3 on quantum mechanics):

$$E_{ex,n} = -\frac{R_y}{n^2} + \underbrace{\frac{\hbar^2 k_{ex}^2}{2m_r^*}}_{\text{kinetic energy of exciton}} + E_g \quad (5.246)$$

The exciton ground state is found by setting n = 1; this corresponds to the ionization energy of the exciton.

In the Bohr model we derived the radii for the hydrogen atom's energy states as:

$$r_n = \frac{n^2 \hbar^2 4\pi\varepsilon_0}{m_e e^2} = n^2 a_0 \quad (3.89)$$

where a_0 is the Bohr radius, and permitted orbits have radii $r_1 = a_0$, $r_2 = 4a_0$, $r_3 = 9a_0$ … for n = 1, 2, 3, …. The Bohr radius, $a_0 = \frac{\hbar^2 4\pi\varepsilon_0}{m_e e^2}$, has a value of 0.53×10^{-10} m. Applying this to an exciton, we again introduce the reduced mass m_r^* to obtain (see Figure 5.91):

$$r_{ex,n} = 0.53\, n^2 \frac{m_e}{m_r^*} \varepsilon_r \quad (5.247)$$

Some calculated values for $E_{ex,1}$ and $r_{ex,1}$ for a variety of materials are summarized in Table 5.8. The exciton Bohr radius gives an estimate of the most probable distance of the electron from the hole. Like

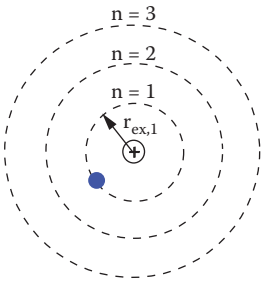


FIGURE 5.91 Bohr radii of an exciton.

TABLE 5.8 Values for $E_{ex,1}$ and $r_{ex,1}$ for Some Selected Bulk Materials

Semiconductor	E_g	$E_{ex,1}$ (meV)	$r_{ex,1}$ (nm)
Si	1.11	14.7	4.9
Ge	0.67	4.15	17.7
GaAs	1.42	4.2	11.3
CdSe	1.74	15	5.2
Bi	0	Small	>50
ZnO	3.4	59	3
GaN	3.4	25	11

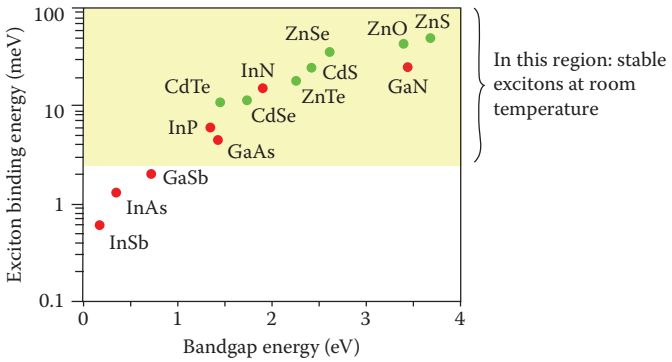


FIGURE 5.93 Stable excitons at room temperature.

the Bohr radius, the "radius" of the exciton increases with n^2. Excitonic absorption peaks appear in the absorption spectrum below the bandgap energy (Figure 5.92). Note that the absorption α decreases when $E_{ex,n}$ increases (higher n value).

Based on Equation 5.245, one can evaluate which materials will lead to stable excitons at room temperature as shown in Figure 5.93.

Excitons in Reduced Dimensionality Structures The above discussion concerned excitons in bulk semiconductors only. Let us consider what happens with their absorption spectra when we confine electron-hole pairs into semiconductor structures of the size of the Bohr radius $r_{ex,1}$ or smaller ($r_{ex,1}$ for a variety of semiconductors was introduced in Table 5.8). In Chapter 3 we similarly considered what happens with the bandgap absorption when confining the semiconductor size (answer: blue shift). Quantum confinement in semiconductor nanoparticles produces quantization of the electron and hole energy states, producing major changes in their optical spectra. The confinement of the oppositely charged carriers at reduced separations also has dramatic effects on the electron-hole Coulomb energy, and thus on exciton formation. For example, in quantum dots

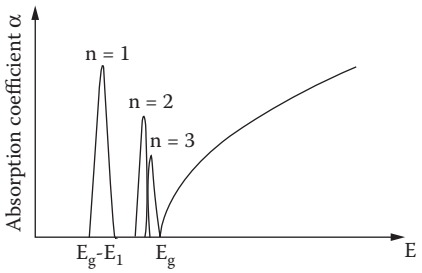


FIGURE 5.92 The hydrogen-like absorption of excitons.

(QDs), excitons are generated inside the dot with a degree of confinement determined by the dot size. The physical boundary of the QD (0D) confines excitons in a well with dimensions on the order of the exciton Bohr radius (~10 nm). One expects that in these confined nanostructures the electron-hole pair interactions will be enhanced with a resulting binding energy increase. The enhanced binding between electron-hole pairs produces increased exciton energy compared with the exciton binding energies we listed for bulk semiconductors in Table 5.8. As a consequence, in confined semiconductors, just as in bulk semiconductors, the excitonic excitations are found below the conduction band edge, but in nanostructures these excitonic peaks are much more pronounced and easier to resolve, even at room temperature. In other words, the probability for exciton formation is much higher in a confined semiconductor such as in quantum wells, quantum wires, and QDs. The binding energy of a 2D exciton, for example, can be deduced from the 2D hydrogen atom problem. The exciton Rydberg energy then changes from $E_{ex,n} = -R_y/n^2$ in the 3D case to $E_{ex,n} = -R_y/(n - 1/2)^2$ in the 2D case (exciton in a well). Thus, the ionization energy for the ground state (n = 1) in the 2D case is four times larger than for the 3D case. Therefore, excitons are four times more stable in quantum wells.

Summary of Semiconductor Absorption Processes

In the preceding discussion we considered EM wave absorption by free carriers, direct and indirect band transitions, and excitons. In a real semiconductor there are also transitions of high-lying energy bands

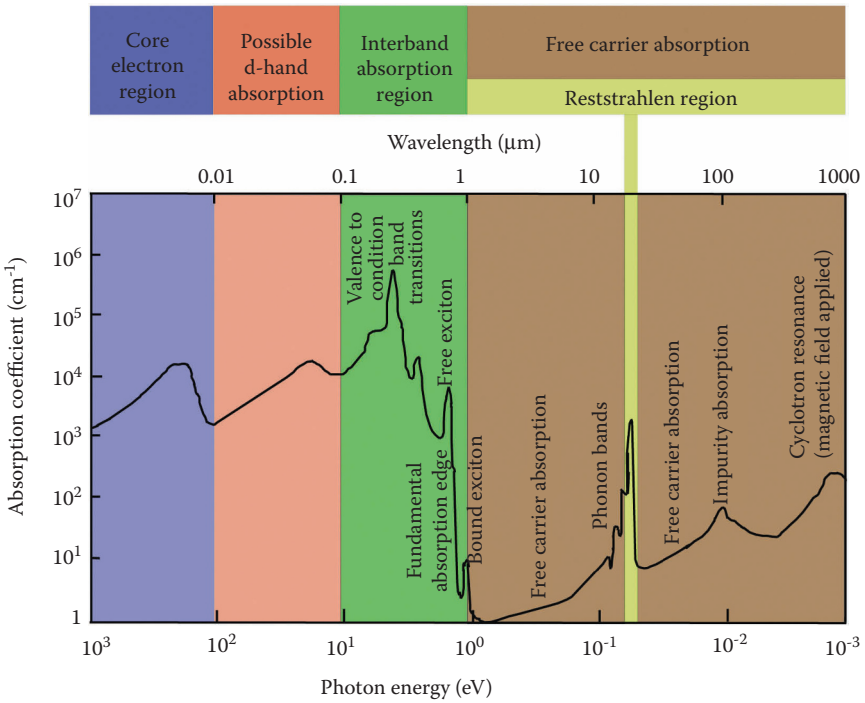


FIGURE 5.94 Different absorption processes in a semiconductor.

and from impurities (including from semiconductor dopant atoms).

Impurities create energy levels in the bandgap (= defect states). Donors and acceptors are situated in the bandgap close to the band edges. Transitions between defect states are at the origin of the impurity absorption. Deep levels are defect states close or in the middle of the bandgap (often created by transition metals such as Fe, Ni, Cu, and Au). Some deep levels generate radiation-less recombinations and should be avoided in optical components. In general, impurities in the bandgap behave like the isolated molecules in a gas as treated above. Doping of Si with donor and acceptor impurities in

TABLE 5.9 Fundamental Excitations and Energy Scales in Solids

Process	Energy
Resulting from "free" valence electrons	
Intraband transitions	≤ $E_{bandwidth}$ (few eV)
Single-electron oscillations	≤ 0.5 meV
Collective electron oscillations (plasmons)	≈ 2–17 eV
Resulting from bound electrons	
Inner (core and valence) electron shells	≈ 4 eV–98.5 keV
Inner-shell excitations (d-d, f-f)	≈ 100 meV–2 eV
Interband transitions (VB → CB)	≈ 10 meV–11 eV
Absorption from localized states:	
- Excitons	≈ E_{Bind} (few meV up to few eV)
- Defects (impurities, vacancies, or interstitials):	≈ 10–300 meV
- Spin-orbit splitting	≈ few eV (for d and f electrons)
- Spin waves (magnons)	≈ 10–40 meV
- Charge transfer excitations	≈ up to several eV
- Correlation effects	≈ order of eV
Resulting from coupling of photons to phonons (lattice vibrations)	
Photons: acoustic phonons (Brillouin scattering)	≈ 10–6–0.1 meV
Photons: optical phonons (polariton scattering)	≈ 40–70 meV

TABLE 5.10 Comparison of Semiconductor Nanoparticles with Metal Nanoparticles

Semiconductor Nanoparticles	Metal Nanoparticles
Synthesis with a variety of chemical methods to produce nanoparticles with desired size, shape, and structure	Synthesis with a sol method by using capping agents to prevent aggregation and control the size
Quantum confinement occurs when the radius of the nanoparticle is comparable to the exciton's Bohr radius	Quantum confinement occurs when the electron's motion is limited by the size of the nanoparticle
The electronic energy levels of the particle around the Fermi level are strongly affected for sizes in the 10–100-nm range	The electronic energy levels of the particle around the Fermi level are affected only for very small sizes (a few nanometers)
The UV-visible absorption is determined by the band edge transition, which is strongly size- and shape-dependent (bandgap energy increases as the size decreases) and hence highly tunable	The UV-visible absorption is determined by the surface plasmon resonance, which is size- and shape-dependent
Can exhibit band edge or defect (red-shifted) luminescence; in both cases luminescence can be very intense and can be tuned	Exhibit a very weak luminescence
The I-V behavior is characterized by the Coulomb blockade	The I-V behavior is characterized by the Coulomb blockade
Show valuable catalytic properties, owing to the large surface to mass ratio	Show valuable catalytic properties, owing to the large surface to mass ratio

semiconductors is covered in detail in Chapter 4. As for excitons, the Bohr model was used to calculate the energy levels associated with those impurities. The different absorption processes are summarized in Figure 5.94 and Table 5.9.

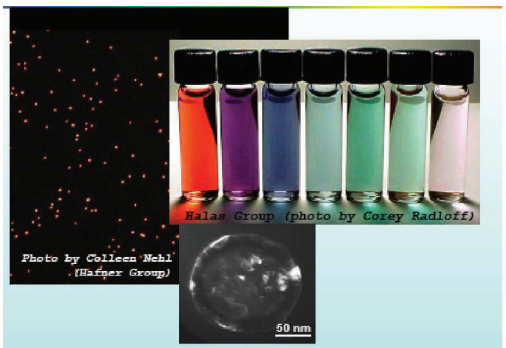
Comparing Semiconductor Nanoparticles (Quantum Dots) with Metal Nanoparticles

In Table 5.10 we compare semiconductor nanoparticles with metal nanoparticles (see also Figure 5.95).

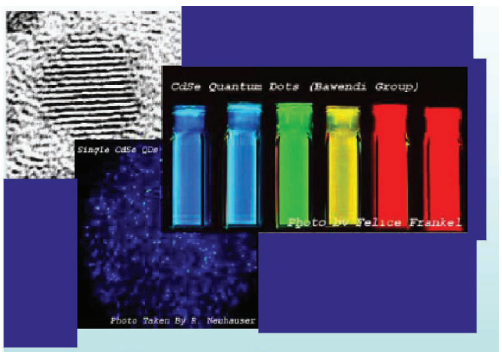
Light Interaction with Small Particles

Introduction: Scattering of Light by Small Particles

When light passes through a medium with suspended particles, some of it is scattered, i.e., directed away from its original direction of travel. In the scattering process, the photon flux from the incident beam (I_0) is not lost but is redistributed over the total solid angle centered around the scattering



Nanoshells:
Tunable plasmonic nanoparticles
~ 10–300 nm diameter
Quantum efficiencies ~ 10^{-4}
Spectral range (extinction):
500(Ag)–9000 nm
Cross-sections:
~10^{-13} m^2



Quantum dots:
Tunable excitonic nanoparticles
~ 1–10 nm diameter
Quantum efficiencies ~ 0.1–0.5
Spectral range (extinction):
400–2000 nm
Cross-sections:
~10^{-19} m^2

FIGURE 5.95 Plasmonic and excitonic nanoparticles are compared.

particles in the medium, and it does not change the internal energy states of the particles (elastic process) (Figure 5.96).

In contrast, absorption changes the internal energy states of the particles (inelastic process). Absorption is also very spectrally selective, whereas scattering is not. Absorption and scattering both attenuate the incoming radiant energy, and this combined effect is quantified by the so-called *extinction coefficient* (C_{ext}; see Equation 5.218 for the case of suspended spherical metal nanoparticles). To classify the interaction of light with small particles, we introduce a dimensionless size parameter α, which is a measure of the size of the particles compared with the wavelength of the incoming radiation:

$$\alpha = \frac{2\pi r}{\lambda} \tag{5.217}$$

In Table 5.11 we list the various scattering regimes of visible light on particles of various sizes. This table is very general, when applied, say, to atmospheric sciences; it illustrates that how solar radiation scatters when it impacts gases, aerosols, clouds, or the ocean surface depends on the ratio of the scatterer's size to the incoming wavelength. A large α in this application may correspond to phenomena such as sunlight on a flat ocean or sunlight on raindrops, and a small α may correspond to IR scattering of air and aerosols or microwave scattering of clouds.

When scatterers are very small compared with the wavelength of the incident radiation (radius of scatterer r is $<\lambda/10$), the scattered intensity in both forward and backward directions is equal (Figure 5.97). This is representative of a type of scattering called *Rayleigh scattering*. A strong dependence of the scattering on the frequency of the light is important to distinguish it from Mie scattering: in Rayleigh scattering, short visible waves (blue) get scattered much more than longer visible wavelengths (yellows and reds). Mie or nonmolecular scattering with $r > \lambda/10$ is much less wavelength-dependent.

Rayleigh scattering is predominant in the presence of molecules and the tiniest particles ($r < 1/10\ \lambda$). An analysis for scattering of light in this regime was first performed by Lord Rayleigh (1871), who undertook this task to explain why the sky is blue. Rayleigh scattering of light is mostly elastic, so no energy is gained or lost.

Rayleigh scattering is not always 100% elastic: in inelastic Raman scattering, 1–10% of the incoming light interacts with molecules in such a way that energy is actually gained or lost, and as a result scattered photons are shifted in frequency. Scattered light under these circumstances reveals satellite lines below and above the Rayleigh scattering peak at the incident frequency. Just like Rayleigh scattering, Raman scattering depends on the polarizability of the scattering molecules.

For particles larger than molecules and more similar in size to the wavelength of the incoming light, the angular distribution of scattered intensity becomes more complex, with more energy scattered in the forward direction than in the backward direction (see Figure 5.97). Scattering in this case is much less dependent on the wavelength of the incoming light, and this is basically why clouds are white. In this regime ($r > 1/10\ \lambda$) the scattering of light is called *Mie* or *nonmolecular scattering*. The German physicist Gustav Mie (1908) was the first to present a complete analytical solution of Maxwell's equations for the scattering of EM radiation from spherical particles (also called *Mie scattering*).[20] Mie's theory of light scattering was actually developed to explain the color of colloidal gold. In general, particles might also absorb, and Mie gave solutions for the absorption, scattering, and extinction cross-sections as a function of the scattering angle. In contrast to Rayleigh scattering, Mie solutions to scattering embrace all possible ratios of diameter to wavelength. It is so far the only "simple" way to study light scattering of any type of particle. Mie's solutions are very important in meteorological

Table 5.11 Various Types of Scattering of Visible Particles on Different Radii r

Type of Particle	Particle Diameter (μm)	Type of Scattering	Phenomena
Air molecules	0.0001–0.001	Rayleigh	Blue sky, red sunsets
Aerosols (pollutants)	0.01–1.0	Mie	Brownish smog
Cloud droplets	10–100	Geometric	White clouds

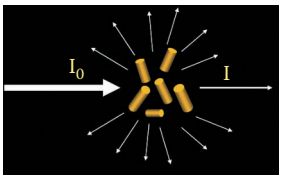


FIGURE 5.96 Incoming photon flux with intensity I_0 is scattered on particles in a medium.

optics, where diameter-to-wavelength ratios of the order of unity and larger are characteristic of many problems regarding white fog and cloud scattering, scattering of radar by raindrops (weather radar), coronas around the sun and the moon, and so on (see Figures 5.112 and 5.113). Other important Mie applications are optical particle characterization for understanding the appearance of common materials like milk, biological tissue, and latex paint and in graphic software for rendering outdoor scenes for movies.

Then finally, for particles much larger than the wavelength of the light, the ray pictures of light become appropriate: for example, when sunlight strikes raindrops between 0.1 and 2 mm in diameter, geometric optics applies, producing such effects as the rainbow illustrated in Figure 5.18.

Rayleigh Scattering (α <<< 1)—Elastic Scattering

In the Lorentz model we saw how an electric field can drive the harmonic motion of bound and free electrons in molecules (see Figure 5.98).

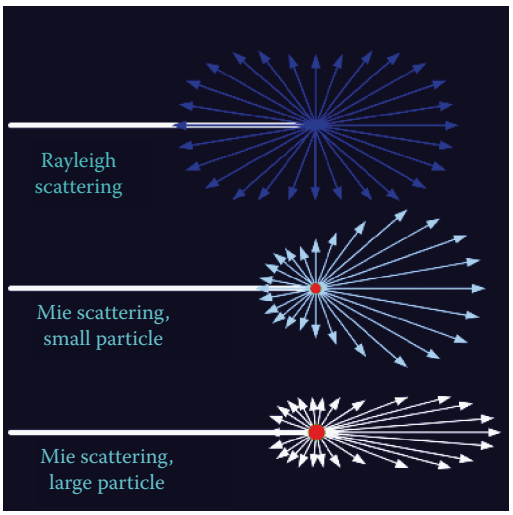


FIGURE 5.97 Rayleigh and Mie scattering compared.

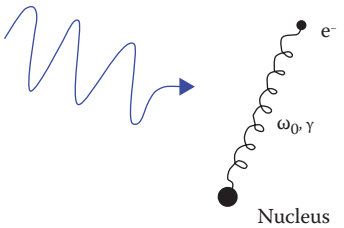


FIGURE 5.98 In the Lorentz model, an electric field drives the harmonic motion of electrons, with ω_0 the resonant frequency and γ the damping factor.

As the incident light's EM field varies, the atoms/molecules that are hit by the light are continuously redistributing their charges. When equal positive and negative charges are displaced in opposite directions, the charge distribution is called a *dipole*. The dipole moment **P** is the product of the quantity of displaced charge and the displaced distance with the vector **P** pointing in the direction of the charge displacement. The oscillating field causes the charges to oscillate at the frequency of the incoming radiation. Earlier in this chapter we explained how oscillating charges radiate EM waves (see Figure 5.3). Oscillating charges reradiate the incoming light elastically at the same frequency as the forcing field: no energy is gained or lost. However, the emitted radiation is not propagated in the same direction as the forcing field and will not necessarily have the same polarization. An isolated dipole oscillator radiates predominantly normal to the direction of the oscillation as illustrated in Figure 5.99.

A small point charge with a different refractive index n than its surrounding medium oscillates as a

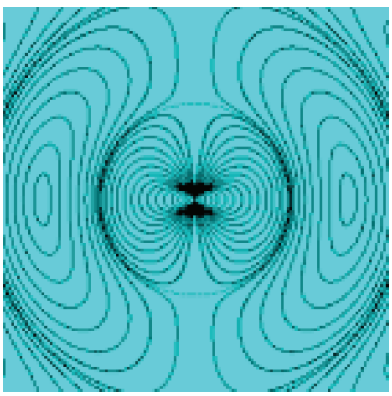


FIGURE 5.99 The electric field from an oscillating dipole (see also Figure 5.3). Formation of electromagnetic waves on a vertically oriented dipole: an oscillating dipole generates EM in a direction mainly perpendicular to the dipole axis. No waves come about along the axis of the dipole.

dipole under the influence of an incident electrical field and reradiates the incoming energy. The angle-dependent total scattered power per unit surface I_s, or the intensity of the light scattered from such an oscillating dipole at a given angle and distance r, is calculated as (see, e.g., the Feynman lectures):

$$I_s = \frac{P^2\omega^4}{32\pi^2\varepsilon_0 c^3 r^2}\sin^2\theta \qquad (5.248)$$

where θ is the polar angle measured from the axis in a spherical coordinate system (Figure 5.100), c is the speed of light, and r is a distance large enough away to be in the far field of the dipole. Notice that in Equation 5.248 the radiation intensity is inversely proportional to the distance squared from the dipole (r^2). The magnitude of the polarization **P** induced in a molecule or small particle by an incident EM field [$E_i(\omega) = E_M e^{i\omega t}$] was calculated earlier as $\mathbf{P} = \chi\varepsilon_0\mathbf{E}_i(\omega)$ (Equation 5.190). This radiation, with $r \ll \lambda$, constitutes Rayleigh radiation, and we detail its characteristics here.

To introduce any type of scattering properly, one also must introduce a phase function $p(\Theta)$ (in sr^{-1}) to describe the distribution of scattered radiation, be it for one particle or for a set of particles. The phase function is a dimensionless and normalized version of the scattering function, such that the integral of the phase function over 4π steradians equals 4π. The phase function for a given wavelength is a property of the medium, not of the incident radiation, provided the incident radiation is unpolarized. As the size parameter of a scatterer increases, its phase function becomes more anisotropic, with progressively more of the scattered radiance concentrated into a diffraction peak. Let the direction of light incidence in Cartesian coordinates be Ω′ and the direction of observation be Ω (Figure 5.101). The angle between these directions is Θ, the scattering angle, i.e., the angle between the incident beam and the scattered light. If Θ is $<\pi/2$, we define forward scattering, and if Θ is $>\pi/2$, we have backward scattering. Converting $\cos\Theta = \Omega'\cdot\Omega$ to polar coordinates, one derives $\cos\Theta = \cos\theta'\cos\theta + \sin\theta'\sin\theta\cos(\Phi - \varphi')$. The phase function $p(\Theta)$ is then mathematically defined as:

$$p(\Theta) = \frac{I_s(\Theta)}{\int_{4\pi} I_s(\Theta)d\omega} \qquad (5.249)$$

To normalize this phase function $p(\Theta)$, we write:

$$\frac{1}{4\pi}\int_{4\pi} p(\Theta)d\omega = \int_0^{2\pi} d\phi \int_0^{\pi} p(\theta',\phi',\theta,\phi)\sin\theta d\theta = 1 \qquad (5.250)$$

The angular distribution of Rayleigh scattering is governed by the classical radiation pattern for the far field of a classical dipole, i.e., a dipole radiation proportional to $\prod\sin^2\theta$, where θ is the polar angle measured from the axis, and ∏ is the induced dipole moment (Equation 5.248). As illustrated in Figure 5.101, we can take the incoming radiation and break it up into two linearly polarized incident waves, one with the electric vector parallel to the scattering

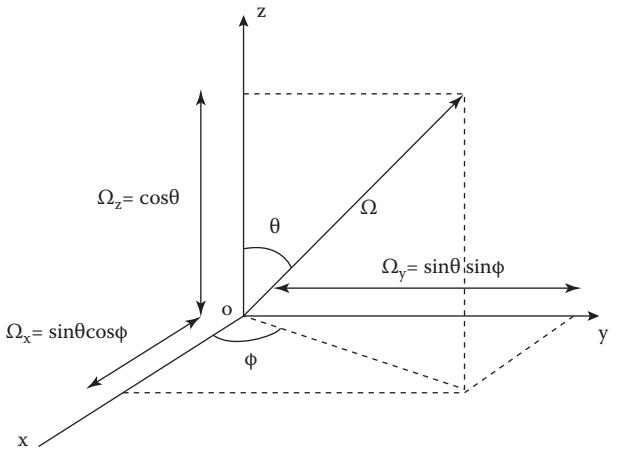


FIGURE 5.100 Relation between Cartesian and spherical coordinates. The rectangular components of the vector Ω are $\Omega_x = \sin\theta\cos\phi$, $\Omega_y = \sin\theta\sin\phi$, $\Omega_z = \cos\theta$.

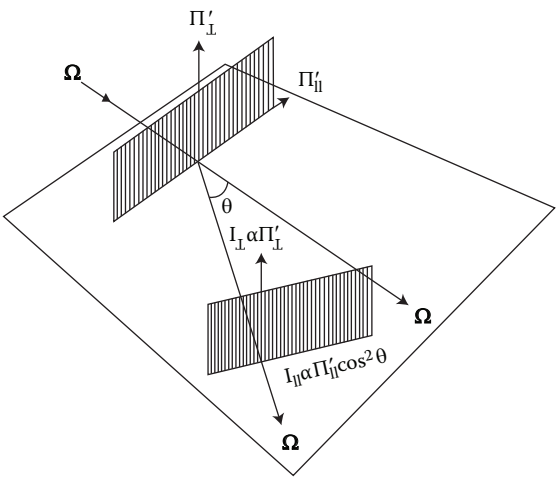


FIGURE 5.101 Scattering in the planes of polarization.

plane and the other perpendicular to the scattering plane. These waves give rise to induced dipoles. If the incident electric field lies in the scattering plane, then the scattering angle is $\pi/2 + \Theta$; if it is perpendicular to the scattering plane, the angle is $\pi/2$. Hence:

$$I_s(\Theta) \propto (I_\perp + I_\parallel) \propto \Pi_\perp \sin^2(\pi/2) + \Pi_\parallel \sin^2(\pi/2 + \Theta) \tag{5.251}$$

given that the parallel and perpendicular intensities are equal:

$$I_s(\Theta) \propto I(1 + \cos^2\Theta) \tag{5.252}$$

For Rayleigh scattering this leads to the following phase function $p(\Theta)$:

$$\frac{1}{4\pi}\int_{4\pi}(1 + \cos^2\Theta)d\omega = \int_0^{2\pi} d\phi \int_0^{\pi}(1 + \cos^2\theta)\sin\theta d\theta = \frac{3}{4} \tag{5.253}$$

or:

$$p(\Theta) = \frac{3}{4}(1 + \cos^2\Theta) \tag{5.254}$$

Figure 5.102 illustrates the phase function for Rayleigh scattering from a dipole in both a 2D and 3D diagram. Let us look first at the 2D diagram. The phase function is symmetric in the plane normal to the incident direction, and forward scatter equals backward scatter. The maxima occur in the direction of $\theta = 0°$ and 180°, and there is minimum radiation along the dipole axis ($\theta = 90°$). The angular distribution of the scattered intensity varies from 1.5 (6/4) in the forward and backward directions to 0.75 (3/4) at 90°, where the parallel component vanishes and the scattered radiation is plane polarized transverse to the scattering plane. In the 3D diagram we simplified things by dropping the factor ¾ in Equation 5.254 so that we obtain 2 at $\theta = 0°$ and 180° and 1 at 90°.

Using $\mathbf{P} = \chi\varepsilon_0\mathbf{E}_i(\omega)$ (Equation 5.190), the link between the intensity I_0 and the incoming electrical field $\left[I_0 = \frac{\varepsilon_0 c E_i(\omega)^2}{2}\right]$, and relying on the Lorentz model (Equation 5.191), we can rewrite Equation 5.248 as:

$$I = I_0 \frac{e^4\omega^4}{16\pi^2 m_e^2 c^4 r^2}\left(\frac{1}{\omega_0^2 - \omega^2 - i\omega\gamma}\right)^2 \sin^2\theta \tag{5.255}$$

where I_0 is the incoming EM radiation intensity. This equation reveals that scattering is strongest near a resonance and that it is also stronger for a higher ω (or shorter λ). In the visible range, oxygen and nitrogen molecules come with $\omega_0 >>> \omega$; in

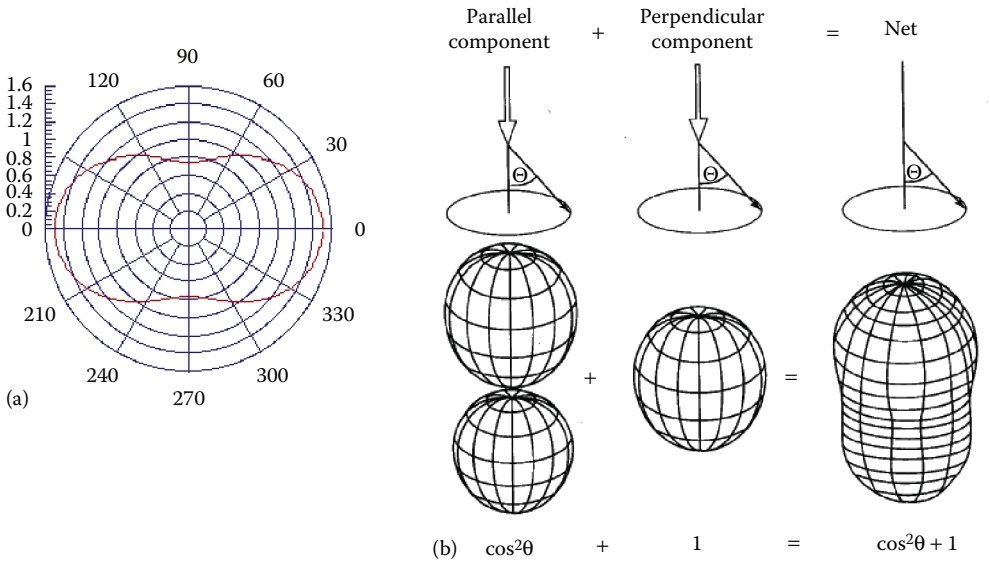


FIGURE 5.102 Rayleigh phase function $p(\theta)$: (a) 2D and (b) 3D. Light scattering from a driven dipole with $r \ll \lambda$. We can take the incoming radiation and break it up into two linearly polarized incident waves, one with the electric vector parallel to the scattering plane and the other perpendicular to the scattering plane. Of the 3D shapes, the shape on the left represents the parallel scattering component, and the sphere in the center represents the intensity from polarization out of the plane. The final peanut shape shows the total scattering pattern for a random incident polarization.

other words, the driving frequency is much less than the natural frequency, and we obtain nonresonant or Rayleigh scattering (also elastic scattering):

$$I = I_0 \frac{e^4\omega^4}{16\pi^2 m_e^2 c^4 r^2}\left(\frac{1}{\omega_0^2}\right)^2 \sin^2\theta \qquad (5.256)$$

Incoming sunlight comes in a wide range of wavelengths, but the longer wavelengths hardly get scattered, whereas based on this equation, a wavelength that is two times shorter gets scattered 16 times more ($2^4 = 16$).

An example of this is observed in the blue sky, where the radiating dipoles are oxygen and nitrogen molecules. The shorter wavelengths (i.e., blue) scatter much more efficiently than longer ones and that is the reason why the sky is blue (blue light ~ 4000 Å is much larger than the scattering particles ~1 Å) [we all should be able to explain why the sky is blue (Figure 5.103)]. During sunrise and sunset, blue and yellow light is efficiently removed from the direct beam over the very long atmospheric paths at the high solar zenith angles, leading to spectacular red skies. You might wonder: because violet light has a shorter wavelength than blue, why then does the sky not appear violet? The reason is that the energy contained in the violet spectrum is much smaller than that contained in the blue spectrum and the fact that the human eye has a much lower response to violet. The Rayleigh mechanism is also of great technological importance because it governs losses in optical fibers for communication. The Rayleigh scattering results from minute local density variations that are present in the liquid glass as a result of Brownian motion and become frozen into the solid.

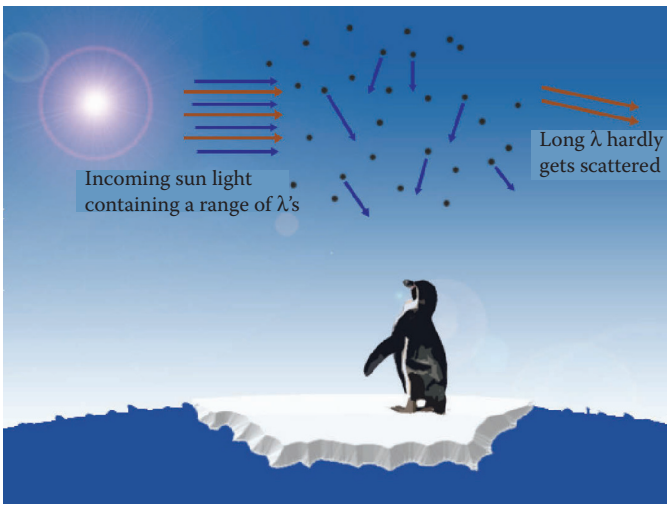


FIGURE 5.103 Why the sky is blue. (Drawing by Mary Amasia, University of California, Irvine.)

The situation just sketched for scattering molecules is equally applicable to scattering from nanoparticles (see the section on metal nanoparticles, metal optics, or plasmonics). If the scatterer is a sphere of radius a and a refractive index n and the incoming light is unpolarized, the particle's polarizability **P** can be written as a function of the refractive index n and diameter d. The intensity of the scattered light in this case is given as:

$$I = I_0 \frac{(1+\cos^2\theta)}{2r^2}\left(\frac{2\pi}{\lambda}\right)^4\left(\frac{n^2-1}{n^2+1}\right)^2\left(\frac{d}{2}\right)^6 \qquad (5.257)$$

where r is the distance to the particle, θ is the scattering angle, n is the refractive index of the particle, and d is the diameter of the particle. Thus, the scattering intensity and energy removed from the incident beam are proportional to the sixth power of the radius (or for nearly spherical particles to the square of the volume). The scattered intensity and energy removed from the incident beam remain proportional to ω^4 or to λ^{-4}. Because the intensity of Rayleigh scattering is proportional to the sixth power of the radius but absorption, however, is proportional to the volume, scattering falls off more rapidly than absorption.

The total scattering power is given by the integral over a sphere surrounding the radiating dipole:

$$I_{total} = \int_A I_s dA = \frac{P^2\omega^4}{12\pi\varepsilon_0 c^3} = \frac{4P^2\pi^3 c}{3\varepsilon_0\lambda^4} I_0 \qquad (5.258)$$

From Equation 5.207 (the Lorentz-Lorenz relationship), we remember that polarizability can be written in terms of refractive index n or also $\frac{\mathbf{P}}{E\varepsilon_0} = \frac{3}{N}\frac{n^2-1}{n^2+2} = \chi$, with the $3/(n^2+2)$ factor added to account for the local field correction, or:

$$I_{total} = \int_A I_s dA = \frac{P^2\omega^4}{12\pi\varepsilon_0 c^3} = \frac{24\pi^3}{N^2\lambda^4}\left[\frac{n^2-1}{n^2+2}\right]^2 I_0 \qquad (5.259)$$

Raman Scattering ($\alpha << 1$)—Inelastic Scattering

Raman Scattering Physics

In 1928, the Indian physicist Chandrasekhara Venkata Raman (1888–1970) from Calcutta University (Figure 5.104) discovered that the wavelength of a small fraction of radiation (1–10%) scattered by certain molecules differs from that of the incident beam. The shifts depend on the molecular structure of the molecules responsible for the scattering. Raman received the Nobel Prize in 1930. Rayleigh scattering of light (90–99%) is elastic, so no energy is gained or lost. In inelastic Raman scattering (1–10%), light interacts with molecules in such a way that energy is gained or lost, and as a result scattered photons are shifted in frequency.

Just like Rayleigh scattering, Raman scattering depends on the polarizability of the scattering molecules. IR bands, on the other hand, arise from a change in the dipole moment. In many cases, transitions that are allowed in Raman are forbidden in IR, so these techniques are often complementary. With polarizable molecules, incident light can excite vibrational modes, leading to scattered light diminished in energy by the amount of the vibrational transition energies (same as in fluorescence). Scattered light under these circumstances reveals satellite lines below the Rayleigh scattering peak at the incident frequency. These are called the *Stokes lines*. If there is enough vibrational energy pumped into the scattering molecules, it is also possible to see anti-Stokes lines (generally weaker than Stokes line) at frequencies above the incident frequency because now vibrational energy is added to the incident photon energy. Because anti-Stokes lines are generally much weaker, it is the Stokes part of the spectrum that is typically used. Summarizing, Raman spectroscopy is a method of determining modes of molecular motions, especially vibrations. It is predominantly applicable to the qualitative and quantitative analyses of covalently bonded molecules. A Raman spectrum is a plot of the intensity of Raman scattered radiation as a function of its frequency difference from the incident radiation (usually in units of wave numbers, cm^{-1}). This difference is called the *Raman shift*. A Raman spectrum of CCl_4 is illustrated in Figure 5.105.

FIGURE 5.104 Indian physicist Chandrasekhara Venkata Raman (1888–1970), who received the Nobel Prize in 1930.

The magnitude of the Raman shifts is independent of the wavelength of the excitation. Raman scattering results from the same quantized vibrational changes that are associated with IR absorption; therefore, Raman scattering spectra and IR absorption spectra of compounds often closely resemble each other. In Raman spectroscopy one typically uses a powerful monochromatic source (a laser), and during irradiation the spectrum of scattered light is measured at an angle (usually 90°) with a spectrometer (Figure 5.106). One important advantage of Raman over IR is that in Raman spectroscopy water does not cause any interference; thus, Raman spectra can be obtained from aqueous solutions. Also, glass or quartz cells can be used rather than NaCl or KBr disks. Raman can also involve rotational or electronic transitions of the molecules from which the scattering occurs. The molecular vibrational

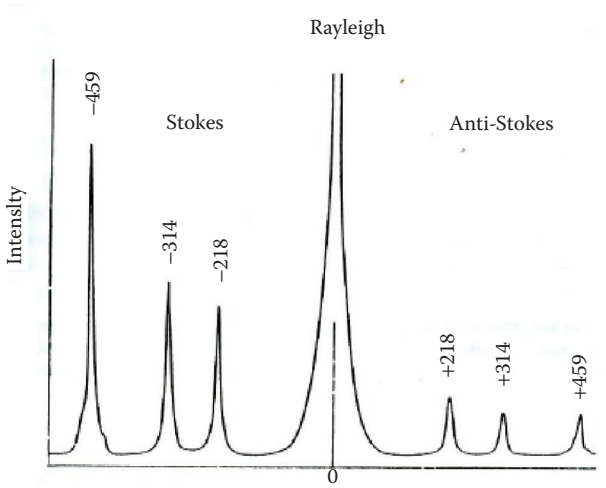


FIGURE 5.105 Raman spectrum of CCl_4. A Raman spectrum is a plot of the intensity of Raman scattered radiation as a function of its frequency difference from the incident radiation (usually in units of wave numbers, cm^{-1}). This difference is called the *Raman shift*.

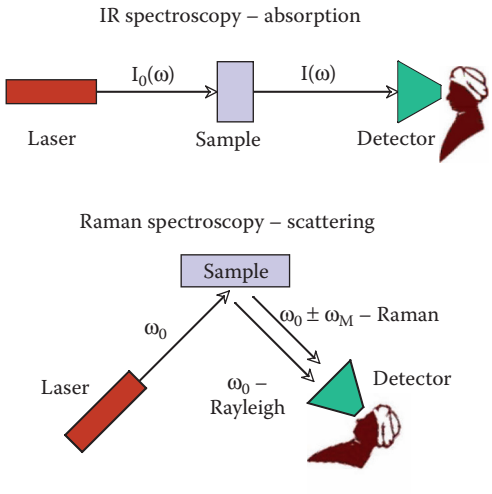


FIGURE 5.106 Raman and IR spectroscopy compared.

energies that may show up in Raman spectra were earlier calculated as $E_n = (n+\frac{1}{2})\hbar\omega$, $n = 0,1,2,3,\ldots$ (Equation 3.191) and the rotational energies as:

$$E_j = \frac{j(j+1)\hbar^2}{2I}, \; j = 0,1,2,\ldots \qquad (5.260)$$

where J is the rotational quantum number and I is the moment of inertia.

Summarizing the advantages of Raman spectroscopy:

1. It provides extremely rich information on molecular vibrations and material structure.
2. There are no special requirements for sample preparations in contrast with IR spectroscopy.
3. It can also be applied noninvasively under almost any ambient condition.

And the disadvantages:

1. It is a low-sensitivity technique because of the low cross-section involved (~10^{-30} cm^2/molecule).
2. Poor reproducibility.
3. A theoretical understanding of underlying mechanisms is still incomplete.
4. A limited number of suitable substrates with Cu, Ag, and Au being the best choices (noble metals with roughness <λ of the light).

Surface-enhanced Raman scattering (SERS) strongly increases Raman signals from molecules attached to metal nanostructures. In SERS, surface plasmons of Ag or Au particles with a size of about 10–100 nm (e.g., colloidal Ag, Au) are excited by a laser. These metals have a large negative real component and a small imaginary component, e.g., the dielectric constant ε'_m of Ag is –2. The local field for a spherical noble metal nanoparticle, as we learned above (see Table 5.6), is enhanced dramatically near the SP resonance with ε'_m (= –2 for a metal like Ag) + $2\varepsilon_d$ (= 1 for air) = 0 (Equation 5.219) and is given by $E_{Local} = \frac{3\varepsilon_d}{\varepsilon'_m + 2\varepsilon_d} E_0$; so E_{local} is very large close to resonance. If the particle is elliptical, then the local field is $E_{Local} = \frac{3\varepsilon_d}{\varepsilon_d\left(1 - \frac{1}{L}\right)} E_0$; In air ($\varepsilon_d$ = 1) the resonance frequency is $\omega_p L$ and the more elliptical the further the resonance is shifted to longer λ*s* (see Figure 5.73).

The resulting electric fields cause nearby molecules to become Raman active, resulting in an amplification of their detection by up to 10^{12}–10^{14} compared with nonresonant Raman. This effect was originally observed by Fleishman but explained by Van Duyne (Northwestern University) in 1977.[21] Two signal enhancement effects play a role in SERS: EM field and chemical adsorption enhancement. If we can increase the local fields, we can obtain a larger effective scattering cross-section (higher signal), and lower detection limits result. Plasmon excitation at metal nanoparticles produces induced fields, and these fields are localized and very intense. The reason that the SERS effect is so intense is because the field enhancement occurs twice. First, the field enhancement magnifies the intensity of the incident light, which excites the Raman modes, thereby increasing the signal of the Raman scattering. The Raman signal is then further magnified by the surface just like the incident light was, resulting in a greater increase in the total output signal. At each stage, the electric field is enhanced as $\mathbf{E}^2$, for a total enhancement of $\mathbf{E}^4$. Thus, because the laser field is enhanced along with the Stokes fields, the total enhancement factor, *G*, is given by:

$$G = \mathbf{E}^2_{laser}\,\mathbf{E}^2_{Raman}$$

$$G = \left[\frac{\varepsilon_m(\omega_{laser}) - \varepsilon_d(\omega_{laser})}{\varepsilon_m(\omega_{laser}) + 2\varepsilon_d(\omega_{laser})}\right]^2 \left[\frac{\varepsilon_m(\omega'_{Stokes}) - \varepsilon_d(\omega'_{Stokes})}{\varepsilon_m(\omega'_{Stokes}) + 2\varepsilon_d(\omega'_{Stokes})}\right]^2 \qquad (5.261)$$

where $\varepsilon_m(\omega)$ is the dielectric constant of the metal, and $\varepsilon_d(\omega)$ is the dielectric constant of the medium in which the nanoparticle is embedded. The laser frequency is ω_{laser}, and ω'_{Stokes} is the Raman shifted frequency. From Equation 5.261 we want $\varepsilon_m + 2\varepsilon_d$ to be small, i.e., near the resonance, so we choose metals (with a negative ε_m). Also, because the gain depends on both E^2 laser(ω) and E^2 Raman(ω'), as the difference $\omega - \omega'$ gets larger, G will get smaller since the resonance condition is *not* met at both ω and ω'. For 30-nm gold particles, the gain is plotted for various laser and Raman wavelengths in Figure 5.107. The field enhancement of nanoparticles also increases with an increase in the radius of curvature of the nanoparticle.

If we move away from the surface of the metal nanoparticle (see Figure 5.108), the dipole decay has to be considered; for a single molecule, this gives:

$$G(d) = \left[\frac{r}{r+d}\right]^{12} G_0 \tag{5.262}$$

with r the radius of the particle, d the distance away from the particle, and G_0 as given by Equation 5.261. And for the case of a monolayer, we obtain:

$$G(d) = \left[\frac{r}{r+d}\right]^{10} G_0 \tag{5.263}$$

The intensity away from the nanoparticle, for a single molecule, is then written out in full as:

$$G(d) = \left[\frac{\varepsilon_m(\omega_{laser}) - \varepsilon_d(\omega_{laser})}{\varepsilon_m(\omega_{laser}) + 2\varepsilon_d(\omega_{laser})}\right]^2 \left[\frac{\varepsilon_m(\omega'_{Stokes}) - \varepsilon_d(\omega'_{Stokes})}{\varepsilon_m(\omega'_{Stokes}) + 2\varepsilon_d(\omega'_{Stokes})}\right]^2 \left[\frac{r}{r+d}\right]^{12} \tag{5.264}$$

The decay curves for a particle and monolayer film are plotted in Figure 5.108. A typical gain G_0 for a single-metal nanoparticle is 10^6–10^7. However, single nanospheres yield relatively small SERS signals compared with a situation where many particles are close to each other (e.g., in a fractal-like aggregate); the additional average enhancement for these aggregates is ~10^5–10^6.[22] Theoretical and experimental studies indicate that the precise control of gaps between nanostructures in the sub-10-nm range (one refers to these sites as "hot junctions") is likely to be critical for the fabrication of SERS-active substrates with uniformly high Raman gain factors.

Besides the EM field, chemical adsorption also plays a role in SERS. When molecules are adsorbed to the metal surface, their electronic states can interact with the states in the metal and produce new transitions. The true nature of the chemical enhancement effect is still not fully understood.

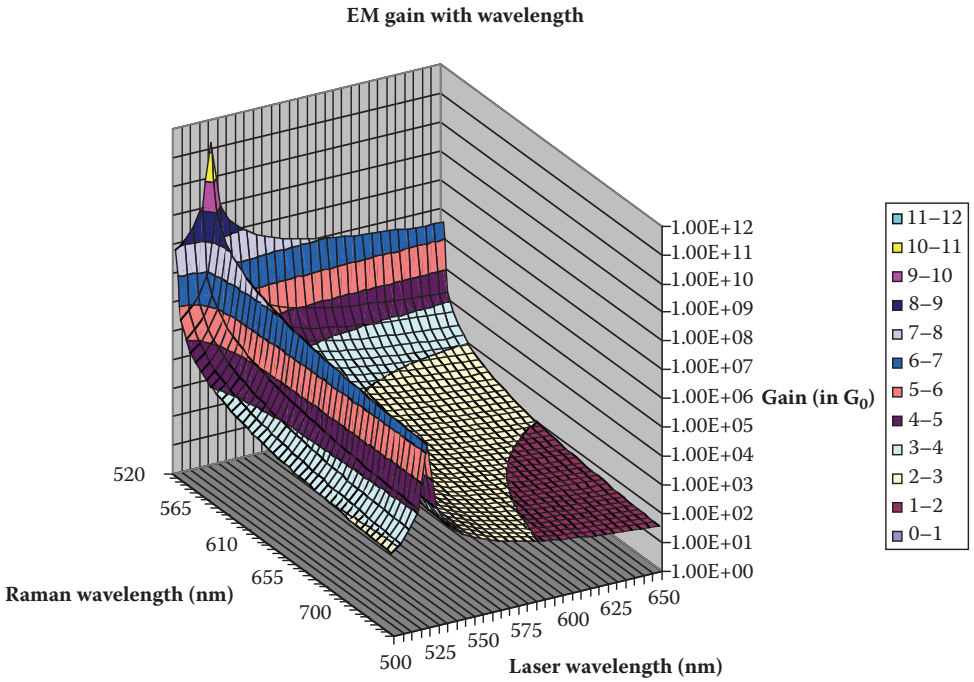


FIGURE 5.107 Gain G expressed in G_0 on 30-nm gold particles for various laser and Raman wavelengths.

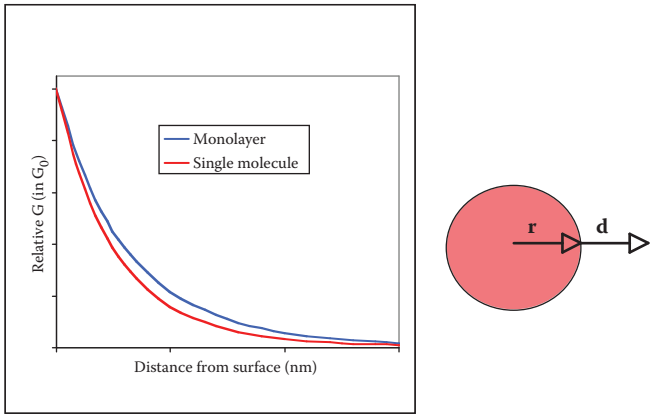


FIGURE 5.108 Decay curves $G(d)$ for a single molecule and for a monolayer (r = 30 nm).

Mie Scattering ($1 < \alpha < 10$)—Elastic Scattering

For particles that are small compared with the incoming wavelength but larger than Rayleigh scatterers (i.e., $r > 1/10\ \lambda$), one has to deal with multiple waves from different molecules/atoms within the particle, and the angular distribution of elastic scattered intensity becomes more complex with more energy scattered in the forward direction. The front and the back of the particle do not experience the same incident electric field, and there is interference of the scattered waves from different parts of the particle. Forward moving waves tend to be in phase, and this gives a large resultant amplitude, whereas backward moving waves tend to be out of phase, and this results in a small resultant amplitude (Figure 5.109). Hence the scattering phase function $[p(\Theta)]$ for such a particle has a much larger forward component (forward peak) than the backward component.

We show Mie's results only pictorially here. In Figure 5.110, polar plots of the Mie-derived scattering phase function $p(\Theta)$ for selected values of α are

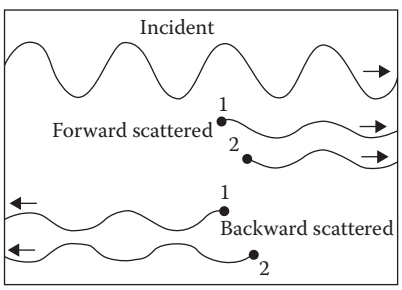


FIGURE 5.109 Incident waves, forward and backward scattering.

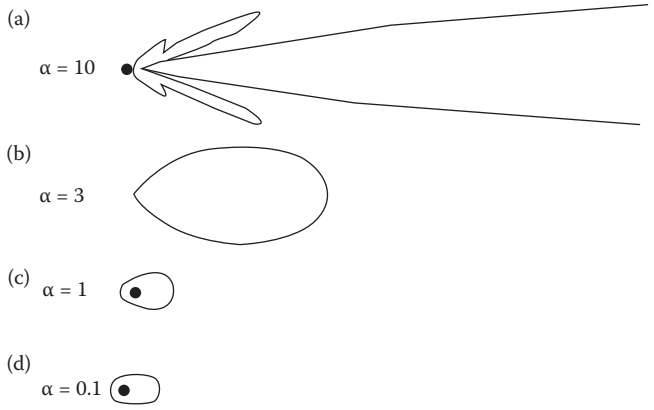


FIGURE 5.110 Polar plots of the Mie-derived scattering phase function $p(\Theta)$ for selected values of α. (a–c) Dominant forward scattering. (d) Equal forward and backward scattering ($\alpha = 0.1$ = Rayleigh scattering). (b and c) Mie scattering from raindrops in clouds with r = 1 μm, and Mie scattering from an aerosol particle with r = 0.1 μm. (a) With $\alpha > 10$, radiation can be described by geometric optics.

shown. At $\alpha = 0.1$, forward and backward scattering are the same (Rayleigh scattering), e.g., solar light on gas (air) with molecules with a radius $r = 10^{-4}$ μm. As α increases, the forward scattering predominates more and more over backward scattering, e.g., Mie scattering from an aerosol particle with r = 0.1 μm and Mie scattering from raindrops in clouds with r = 1 μm. With $\alpha > 10$, radiation can be described by geometric optics. When the size of the particle becomes larger than the wavelength, there is an extreme concentration of scattering in the forward direction with the development of maxima in different directions for different colors. The latter explain the optical phenomena of fogbows and glories (see Figure 5.113).

Typically Mie solutions are represented in terms of $K = \sigma/\pi r^2$, a dimensionless number that plays the role of the scattering area efficiency and depends on the size parameter $\alpha = 2\pi r/\lambda$ and the index of refraction n of the particles responsible for scattering. This coefficient measures the ratio of the effective cross-section of the particles to their geometric cross-section. In Figure 5.111, we plot K as a function of α.

For the lowest values of α (in the region with $\alpha \ll 1$ up to point A), the slope of the curve $(\Delta K/\Delta\alpha)$ in Figure 5.111 is positive, indicating that for a uniform set of spherical particles, scattering increases with decreasing wavelength (this typifies Rayleigh

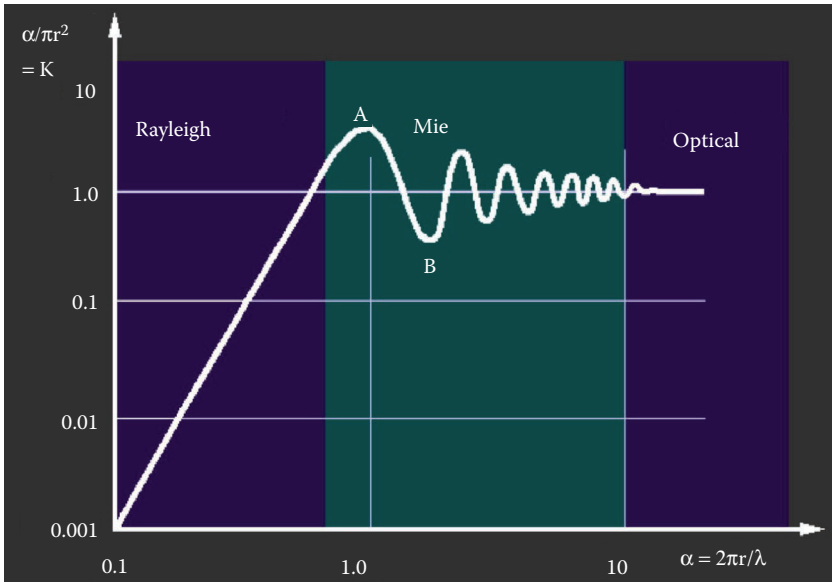


FIGURE 5.111 Scattering efficiency K as a function of size parameter α.

scattering). For a given refractive index, K is proportional to α^4. In this region, scattered radiation is evenly divided between the forward and backward hemispheres. Examples of Rayleigh scattering are scattering of solar radiation by air molecules (blue sky) and scattering of microwave radiation by raindrops. The latter is used in weather radar.

In the region between A and B the slope is negative, indicating that scattering decreases with decreasing wavelength. Examples include the scattering of sunlight by particles of haze, smoke, smog, and dust. In the region beyond point B, K exhibits oscillatory behavior. For $\alpha > 10$, K is more or less constant, and geometric optics applies.

In Figure 5.112 we show the log of the Mie phase function [log $p(\Theta)$] for Mie scattering by a cloud for red, green, and blue light. Scattering efficiencies in Mie scattering are similar for different colors, and that is why clouds are white. A large portion of the incident beam is scattered in the direction of the incident radiation (forward scattering), and a smaller portion of the incident beam is scattered in the opposite direction of the incident direction (backward scattering). Back scattering maxima occur in different directions for different colors, and this explains the beauty of fogbow and glories. A fogbow is a rare form of a rainbow. The water droplets in a fogbow are very small, so interference causes the colors to blend into pure white (Figure 5.113). The droplets then produce very weak colors. Glories are small luminous rings that surround an observer's shadow on a cloud deck or fog bank. A glory can

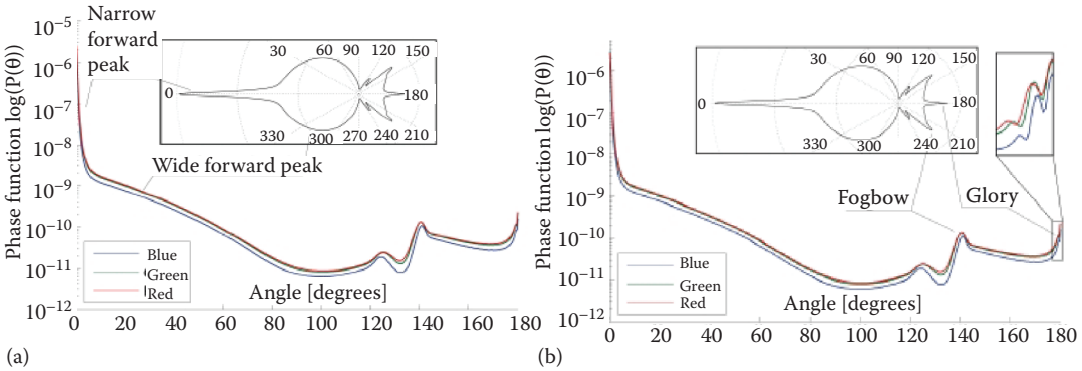


FIGURE 5.112 The log of the Mie phase function for Mie scattering of a cloud for red, green, and blue light as a function of scattering angle θ. Light is coming in from the right side. (a) Forward scattering. (b) Backward scattering. (See Figure 5.113 for an image of a fogbow and a glory.)

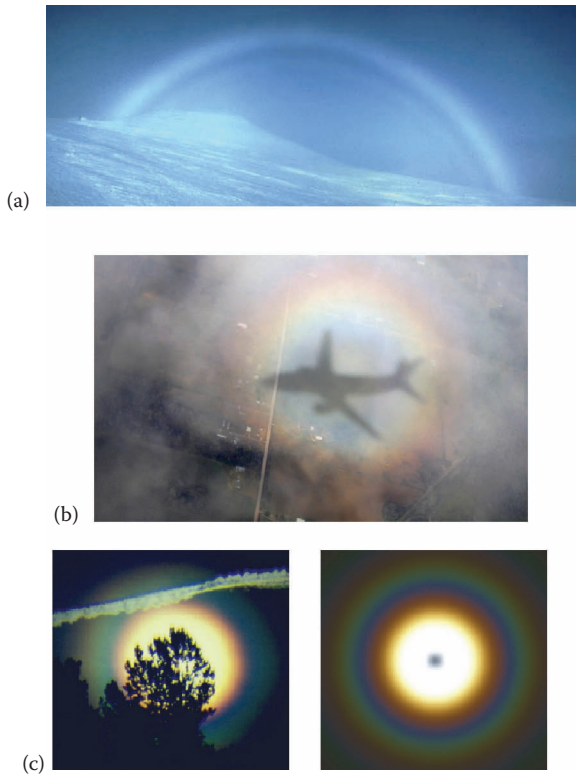


FIGURE 5.113 Fogbows, glories, and coronas. (a) White fogbow on Cairngorm, Scotland. (The fogbow was photographed by Simon Caldwell in February 1999.) (b) Jonathan Lansey saw this unusually bright glory while descending into Houston. (c) Corona simulation (right) matches Paul Neiman's photo of a corona with a jet contrail in mountain clouds above Nederland, Colorado (left).

often be seen surrounding an airplane's shadow. It is a pale, circular rainbow caused by the scattering and reflecting of sunlight from water droplets in clouds below the plane. A glory is not a true rainbow and is usually very small.

Finally, coronas occur when a thin cloud layer moves in front of the sun or moon. The light that shines through the cloud droplets curves (is diffracted) around the water droplets, and diffracted light casts a light larger than the original light source; several circles may appear because of the different wavelengths. The sharper they are, the more similar in size the droplets are.

Mie scattering software can be found at http://omlc.ogi.edu/software/mie.

Geometric Scattering ($\alpha > 10$)

If the scatterers are large compared with the wavelength of the incident light, then geometric optics provide a good approximation. We saw above that with $\alpha > 10$, K becomes constant (Figure 5.111). Scattering of visible radiation by cloud droplets, raindrops, and ice particles falls within this regime. Natural optical phenomena in this regime include rainbows and halos.

Plasmonic Applications 3

Introduction

We are returning to plasmonic applications for a final time in this chapter. This time we consider nanoscale confinement of EM radiation in near-field scanning optical microscopy (NSOM) for subwavelength resolution and detail NSOM applications with an emphasis on surface-enhanced Raman spectroscopy (SERS). We start by describing how NSOM manages to break through Abbe's diffraction limit. For more on NSOM see also Volume III, Chapter 6 on metrology.

Breaking the Diffraction Limit with NSOM

Earlier in this chapter, we saw how diffraction imposes a limit on the "resolution" achievable with conventional optics (see Abbe's diffraction limit, Equation 5.1). The best optical microscopes, for example, can only reach an ~200–250-nm resolution (confocal microscopy). For a long time, Abbe's law was regarded as being insurmountable: small features (or large wave numbers) of an object are lost because of the exponential evanescence of short-wavelength waves (see Figure 5.114). The evanescent near-field mode waves of the source decay exponentially away from the source, and they are never detected in conventional imaging; hence the diffraction limit on the image resolution, i.e., subwavelength features, is not resolved. One would need to amplify the near field to go beyond this diffraction limit.

For higher resolution, the textbook knowledge for a long time was that a complicated and costly electron microscope (0.1-nm resolution) or scanning tunneling microscope (STM) would be required. These techniques have a better resolution than Abbe's 200-nm limit but sacrifice many of the advantages associated with optical microscopes (i.e., nondestructiveness, low cost, high speed, reliability, versatility, accessibility, ease of use, informative contrast, spectroscopy, and real time). However, there

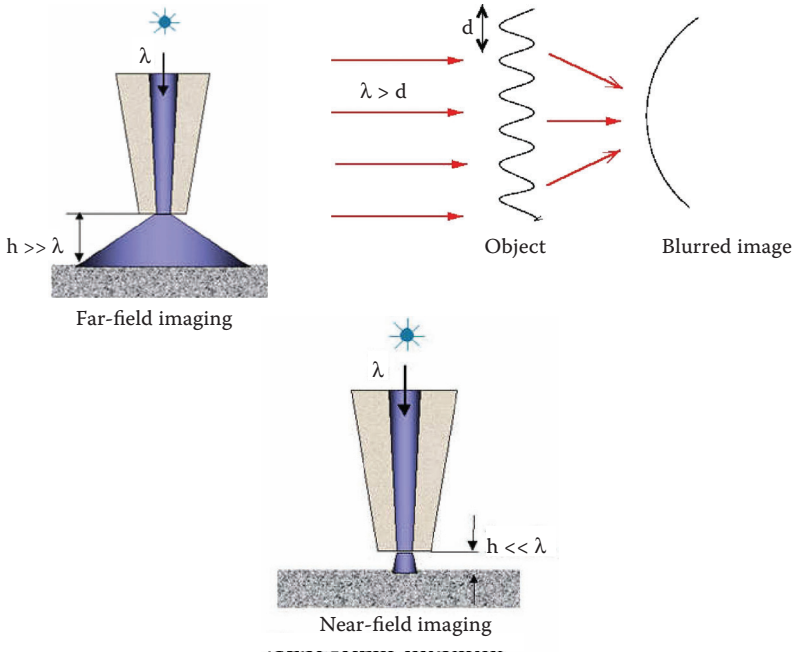


FIGURE 5.114 Near-field and far-field imaging. Small feature *d* is not resolved in far field.

is a solution, i.e., to combine a scanning proximal probe technique with optical microscopy in so-called *scanning near-field optical microscopy* (NSOM or SNOM). If we gather the light coming from the object as close as possible to the object, then the exponential evanescent waves are collected as well (see Figure 5.114). There are two fundamental differences then between near-field and far-field (conventional) optical microscopy: the size of the specimen area that is illuminated, and the separation distance between the source of radiation and the specimen. In NSOM the illuminated area is nanometric in dimensions, whereas conventional microscopes use macroscopic fields of view. In conventional far-field optical microscopy, the distance between the light source and the specimen is typically much greater than the wavelength of the incident light, whereas in NSOM, a necessary condition of the technique is that the illumination source is closer to the specimen than the wavelength of the illuminating radiation. To probe the "near field" requires one to be within 100 nm of the object. This is achieved in a variety of ways, for example, by using nanoscale apertures or by using apertureless techniques that enhance light interaction over nanoscale dimensions with the use of scattering nanoscale tips, nanospheres, and so on. Thus, NSOM is a form of lensless optics with subwavelength resolution, independent of the wavelength of the light being used. The light coming from an imaged object in the near field contains a considerable amount of nonpropagating, evanescent waves containing high lateral spatial resolution, information that decays exponentially within a distance comparable to a wavelength. It is this evanescent wave information, the information-rich light, we capture in NSOM, thus enhancing the resolution. Detecting the collected light in NSOM can be achieved with a wide variety of instruments: an Avalanche Photo Diode (APD), a Photomultiplier Tube (PMT), a CCD, or a spectrometer. The signals obtained by these detectors is then used to create an NSOM image of the surface.

For high-resolution imaging with NSOM, the tip must be positioned and held within nanometers of the sample surface during scanning. There are various feedback mechanisms that have been used for this purpose: electron tunneling, photon tunneling, impedance, and reflection measurements. The most widely adopted method is the shear-force technique in which the NSOM tip mounted on a tuning fork is dithered laterally at one of the mechanical resonances of the fork. To excite the mechanical resonance of the fork, it is rigidly mounted on a ceramic piezoelectric tube serving the purpose of a dither. The tuning fork and the tip are vibrated parallel to the sample surface. To avoid losing optical resolution,

the amplitude of the vibration is kept low, usually <10 nm. As the tip approaches a sample surface, shear forces acting between tip and sample dampen the amplitude of the tip vibration. That amplitude change can be monitored by several methods and is used to generate a feedback signal to control the tip-sample gap during imaging. NSOM sorts out in two major categories, i.e., aperture near-field optics and apertureless near-field optics also called *scattering-based NSOM* (SNSOM). The optical apertures often consist of a tapered, drawn optical fiber that is coated by a metal, typically aluminum, to confine the light and prevent it from leaking out of the sides of the tip, except for a tiny hole at its very apex. Aluminum is used because of ease of application and a minimum skin depth in the visible and IR. Metal film coatings have a finite skin depth (the minimum thickness required to produce opacity) that limits the minimum probe size that can be coated with sufficient thickness to prevent penetration of light through the walls of the tip. The light emanates through the tip opening, which is in the range of ~50–100 nm. The distance to the sample is in the same range; thus, the sample senses the near-field distribution of the light. Hence, the light is confined to a dimension smaller than its wavelength, and thus near-field optics overcomes the diffraction limit. There are other aperture types, such as micromachined cantilevered AFM/NSOM tips (Si_3N_4, SiO_2) with tetrahedral tips featuring a hole at the apex. Finally, one might also use the tip of a tapered pipette filled with a light-emitting compound, which can then be excited either by light or by applying a voltage. It is also possible to use chemical luminescence. Several implementations using an aperture are illustrated in Figure 5.115.

Illumination by the tip and measuring in transmission mode is probably the easiest to interpret and gives the most signal (Figure 5.115a). Transmission NSOM allows images of the intensity of the light transmitted through the transparent sample to be obtained. The sample is illuminated via a probe aperture during the scanning, and transmitted light may be collected by an inverted optical microscope and directed to a photomultiplier tube (PMT). However, it requires a transparent sample, and so it cannot always be used. Reflection modes give less light and are more dependent on the details of the probe tip but allow one to study opaque samples. Reflection NSOM allows images of the intensity of light reflected from the sample to be taken (Figure 5.115b). The sample is illuminated via a probe aperture, and reflected light is directed by a mirror system through the objective of the inverted microscope to the PMT. In collection mode the sample is illuminated with a macroscopic light source from the top or bottom, and the probe is used to collect the light from the sample surface (Figure 5.115c). The

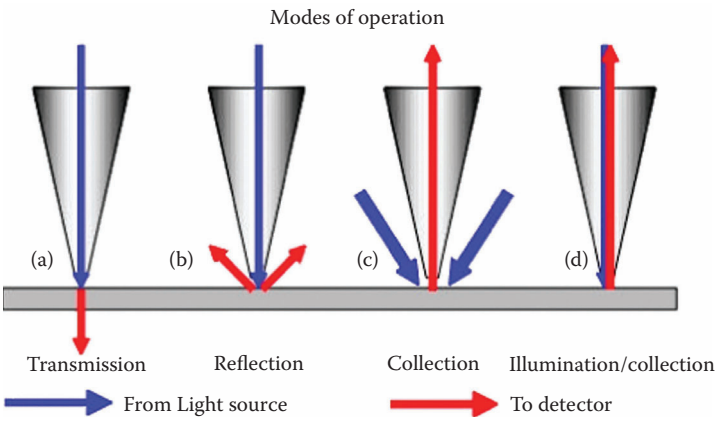


FIGURE 5.115 NSOM aperture-based proximal probes/tips. (a) Transmission mode imaging. The sample is illuminated through the probe, and the light passing through the sample is collected and detected. (b) Reflection mode imaging. The sample is illuminated through the probe, and the light reflected from the sample surface is collected and detected. (c) Collection mode imaging. The sample is illuminated with a macroscopic light source from the top or bottom, and the probe is used to collect the light from the sample surface. (d) Illumination/collection mode imaging. The probe is used for both the illumination of the sample and for the collection of the reflected signal.

illumination/collection mode provides a complement to the reflection modes, but the signal contains a large background from the light reflected without ever making it out of the probe tip (Figure 5.115d). In this hybrid mode the probe is used for both the illumination of the sample and for the collection of the reflected signal. Oscillating the tip vertically and measuring the signal at that oscillating frequency helps remove this background.

In Figure 5.116 we illustrate how an aperture-based (tapered optical fiber, coated with Al on it sides) near-field microscope can be used to probe luminescent semiconductor quantum dots (QDs). Luminescence NSOM allows images of local luminescence intensity to be gathered. The sample is illuminated via a probe aperture, and transmitted light is directed to the PMT through an inverted microscope objective and a notch filter.

Tapered coated fiber probes, as discussed above, have a resolving power of 30–50 nm at most because the amount of light that can be put through the small pinhole is limited. Apertureless near-field microscopy resolves the illumination intensity problem by shining light obliquely at a Rayleigh scatterer and periodically perturbing the interaction between the sample and its illumination. The perturbation can involve probe-induced phase shifts, or it can involve the periodic obscuration of the illuminating beam by a sharp probe. Measurement of the periodic perturbation involves lock-in detection. The perturbation is synchronous with the probe movement, and since the interaction of the sample with the probe is not the same at the low point than at the high point of the probe trajectory, lock-in detection allows the discrimination of the near-field information (which is modulated at the probe's oscillation frequency) from the far-field optical information (which varies more slowly). Apertureless SNSOM typically uses a sharp metallic subwavelength tip instead of an optical tip as described above. The sharp metal tip neither acts as excitation nor as detection light waveguide. Instead, it acts as a nanoscopic light scatterer (Rayleigh scatterer) when placed in near-field distance with respect to the sample. For "apertureless" NSOM the resolution can reach 10 nm; however, background from diffraction-limited illumination is a problem. Oblique incident outside light is scattered at the tip, and the resulting strongly enhanced local plasmonic field is used for excitation. The resulting optical response signal is collected via a lens or an objective with the collecting focus being adjusted to the tip. Exciting the surface plasmons propagating toward a tip of a tapered metal nanowire can make for a very intense local nanoscopic light source. As we learned in this chapter, plasma oscillations in a metal are vibrational modes of the electron gas density oscillating about the metallic ion cores (plasmons). Surface plasmons describe the special case in which the charges are confined to the surface of a metal. In this case, the electric field is strongest in the plane of the metallic surface. As we illustrated in Figure 5.44, surface plasmons confined to a plane do not radiate light (the surface plasmon mode always lies below the light line). However, when the local planar symmetry is disturbed, as illustrated with a couple of different configurations in Figure 5.46, plasmons can radiate. It is thus possible to design structures that take advantage of this emission of surface plasmon radiation, and this may be used to enhance radiative processes. This phenomenon leads to a giant concentration of energy on the nanoscale. The strength-of-field enhancement at the metal tip depends both on the tip material and its geometrical design. The enhancement at the tip is caused by a combination of the electrostatic lightning rod effect, which is due to the geometric configuration at the tip, and localized surface

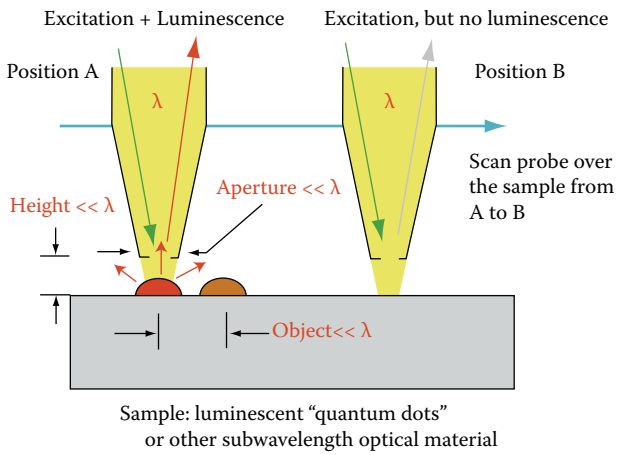


FIGURE 5.116 Use of NSOM to image a luminescent quantum dot.

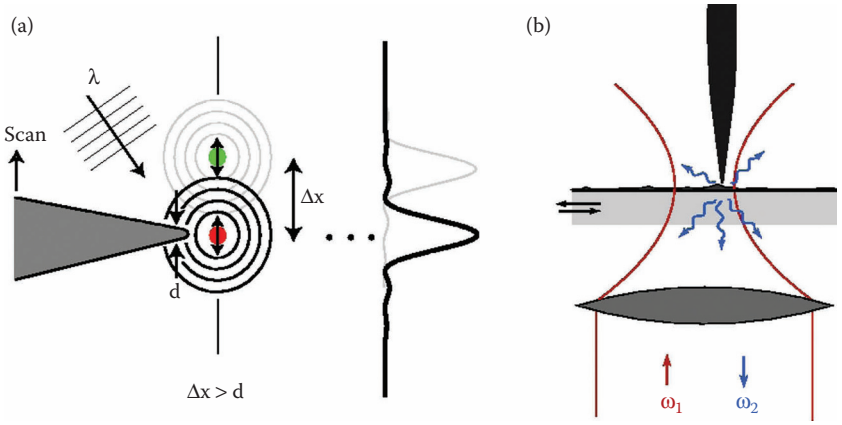


FIGURE 5.117 Local field enhancement technique. A laser-irradiated metal tip enhances the incident electric field near its apex, thereby creating a localized photon source. (a) Schematic of the method. (b) Practical implementation. A laser beam is focused on a sample surface, and a sharply pointed metal tip is positioned into the laser focus. The enhanced fields at the tip locally interact with the sample surface, thereby exciting a response that is collected by the same objective and directed on a detector. (From http://www.optics.rochester.edu/workgroups/novotny/muri03/progress04-01.htm.)

plasmon resonances that depend sensitively on the excitation wavelength (Figure 5.117).

Under "The μ-ε Quadrant and Metamaterials," we will learn that superlenses also prevent image degradation and also beat the diffraction limit established by Abbe.

Pros and Cons of NSOM

From the above, near-field scanning optical microscopy (NSOM) refers to the interaction of light with a sample close to a metal aperture or a solid metal tip, which constrains the lateral extent of the light. The light source is held in place in a manner similar to those used for other scanning proximal probe microscopes (e.g., AFM, STM; see Volume III, Chapter 6 on metrology). Advantages of the near-field interactions in NSOM include:

1. Improved spatial resolution (e.g., 10-nm resolution can be achieved with scattering-based NSOM (SNSOM)).
2. NSOM is a complementary optical tool to other scanning probe microscopes such as AFM and STM, which can only provide topographic and mechanical and electronic information, respectively.
3. A particular advantage of NSOM is to allow for a variety of powerful means of local surface analysis such as fluorescence near-field imaging and spectroscopy, which are highly sensitive and selective techniques to determine the physical and chemical properties of surfaces; specifically, surface-enhanced Raman spectroscopy (SERS) can be coupled with NSOM (see below).

Disadvantages:

1. Practically zero working distance.
2. Extremely small depth of field so that only features at the surface of specimens can be studied.
3. Extremely long scan times for high-resolution images or large specimen areas.
4. Very low transmissivity of apertures smaller than the incident light wavelength (does not apply to apertureless NSOM).
5. Fiberoptic probes are somewhat problematic for imaging soft materials because of their high spring constants.

NSOM Applications

Near-field optics technology offers an important window on the future of a wide range of vibrant new nanotechnologies. These new nanoapplications range from:

- Ultrahigh resolution optical imaging
  - Spectroscopy
  - Near-field surface-enhanced Raman spectroscopy (near-field SERS)

- Local spectroscopy of semiconductor devices
- Modification of surfaces
- Subwavelength photolithography (see Volume II, Chapter 2)
- Ultrahigh density data storage
- Laser ablation
- Deposition of various materials such as in the photodecomposition of chemical gases, i.e., near-field optics photochemical vapor deposition (NFO-PCVD)
- Near-field femtosecond studies, etc.

Several of these applications are dealt with in detail elsewhere in the book. Here we briefly describe some of them. As we saw above, measuring the Raman spectrum of scattered light provides a powerful "fingerprinting" technique for the identification of molecules (Raman spectroscopy). However, typical Raman signals are rather weak, and we saw how the EM fields carrying the fingerprinting information can be enhanced by many orders of magnitude through plasmons generated on assemblies of resonating metal nanoparticles. The combination of scanning near-field optics (NSOM), which enhances locally the Raman effect (SERS), enables localization and scanning of illumination or collection of light over a surface. This provides the means to detect trace amounts on surfaces that would otherwise be undetectable. Trace analysis down to a detection limit of about 100 molecules has been demonstrated. This is of use, for example, in environmental, biomedical, and pharmaceutical research.[23] Even single-molecule detection is possible with SERS. The important evidence of single-molecule SERS is blinking of the SERS signal and spectral diffusion of Raman shift with time. Single-molecule detection by Raman scattering is useful, especially for the detection and distinction of single DNA bases: the nucleotide bases show well-distinguished SERS spectra.[24]

In Volume II, Chapter 2, on next-generation lithographies, we will see that NSOM is used to write submicron patterns into a photoresist. Aperture NSOM in illumination mode (Figure 5.115) can also be used to write and read data on magneto-optical materials. When writing, the probe heats the medium under the probe. The domain under the write beam becomes magnetized opposite of the surrounding material. This varying domain of magnetization can then be read using the same probe but with a lower output power.

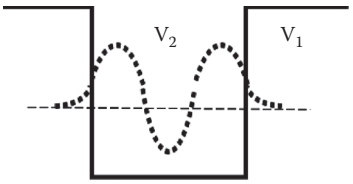
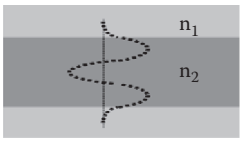
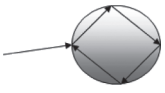
Near-field optical ultrafast laser ablation relies on extremely fast-energy deposition in the femto-/picosecond range, resulting in subsequent material ablation without thermal heat shocks or mechanical damaging (i.e., no melting, burr formation, or cracks). Near-field optics photochemical vapor deposition (NFO-PCVD) uses NFO in a vacuum as a patterning method to deposit material very locally via photodissociation.

Comparing Photons with Electrons—Photonic Crystals

Introduction

Photons and electrons exhibit wave and particle characteristics. When treated as waves, the wavelength associated with either particle is given by the de Broglie equation, i.e., $\lambda = h/mv$ (Equation 3.96), where $mv = p$ is the particle momentum and h the Planck constant. From Schrödinger's equation we established that the space propagation of electrons (Chapter 3), and, from Maxwell's equations that of photons (current chapter), can be described by plane waves as summarized in Table 5.12. Photons, we saw, satisfy Bose-Einstein statistics, whereas electrons are fermions. In Chapter 3 we also calculated the wave equations for EM waves from Schrödinger's equation because, according to Bohr's correspondence principle, the results from the older theory (Maxwell) should also follow from the newer theory (Schrödinger). The probability of finding matter at position x in quantum mechanics scales with $|\Psi_0(x,t)|^2$, and this may be compared with the probability of detecting light in Maxwell's theory scaling as $|\mathbf{E}_m(x,t)|^2$. Light waves are described by an electric ($\mathbf{E}$) and magnetic ($\mathbf{B}$) vector field, whereas electron wave functions are scalars. Comparing the photon and electron momentum [$\mathbf{p} = (h/2\pi)\mathbf{k}$ or $\hbar\,\mathbf{k}$], we find that the electron momentum is generally bigger than the photon momentum because the electron wave vector ($\mathbf{k} = m_e v/\hbar$) is bigger than the photon wave vector ($\mathbf{k} = 2\pi/\lambda$). Because the photon wave vector is

Table 5.12 Comparison of Electrons and Photons

		Electrons	Photons
Equation		Schrödinger	Maxwell's equations for light
Applicable particles distribution		Electrons are fermions and follow Fermi-Dirac statistics	Photons are bosons and follow Bose-Einstein statistics
Solution	Generic	Describes allowed energies of electrons	Describes the allowed frequencies of light
	Free space	$\psi = \psi_0\left(e^{i\mathbf{k}\cdot\mathbf{r}-\omega t} + e^{-i\mathbf{k}\cdot\mathbf{r}+\omega t}\right)$: a scalar wave function	$\mathbf{E} = \mathbf{E}_m\left(e^{i\mathbf{k}\cdot\mathbf{r}-\omega t} + e^{-i\mathbf{k}\cdot\mathbf{r}+\omega t}\right)$: a vector field described by **E** and **B**
	In a medium	Propagation of electrons affected by Coulomb potential	Propagation of light affected by the dielectric medium (refractive index)
Wavelength	Equation	$\lambda = \frac{h}{\mathbf{p}} = \frac{c}{mv}$	$\lambda = \frac{h}{\mathbf{p}} = \frac{c}{v}$
	Remark	Electron momentum **p** is larger so the wavelength of electrons is generally <1 nm	Photon momentum **p** of light is smaller so the wavelength of light is generally <1 μm
Momentum	Equation	$\mathbf{p} = (h/2\pi)\mathbf{k}$ or $\hbar\mathbf{k}$	$\mathbf{p} = (h/2\pi)\mathbf{k}$ or $\hbar\mathbf{k}$
	Remark	Same equation	Same equation
Energy	Equation	$E = \frac{(\hbar\mathbf{k})^2}{2m_e}$	$E = \mathbf{p}c = (h/2\pi)\mathbf{k}c = \hbar\mathbf{k}c$
	Remark	Parabolic	Linear
Speed		The velocity of an electron, *v*, can be any value smaller than *c*.	Fixed = c for electromagnetic radiation is always 3 10^8 m/s
Mass		$m = 9.1 \times 10^{-31}$ kg	m = 0
Propagation through classically forbidden zones		 $V_1 > V_2$ Wave function for a particle in a potentail well Electron wave functions decay exponentially in forbidden zones	 Planner waveguide, $n_1 > n_2$ Photons tunnel through classically forbidden zones. **E** and **B** fields decay exponentially and **k** vector is imaginary.
	0D (no size confinement)	Single crystal: bulk material	Photonic crystals: bulk material
Confinements	1D	Thin film-quantum well: 1D confinement	Planar waveguide: 1D confinement
	2D	Quantum wire: 2D confinement	Optical fiber: 2D confinement
	3D	Quantum dot: 3D confinement	 Microsphere: 3D confinement
Bandgap, forbidden energies		Electronic bandgap	Photonic bandgap
Possible states		Spin up and spin down	Polarization
Charge		$q = -1.6 \times 10^{-19}$ C	q = 0

small compared with the electron wave vector, the wavelength λ for an electron is of the order of the atomic spacing (<1 nm), whereas the wavelength of light is generally much larger (<1 μm). For photons, the energy $E = pc = \hbar \mathbf{k}c$, whereas for electrons, the energy $E = (h/2\pi)^2\mathbf{k}^2/2m$. Thus, whereas the energy dispersion, i.e., the variation of energy with wave vector $\mathbf{k}$ ($=2\pi n/\lambda$), is linear for photons, for electrons it is a parabola (Figure 5.118). The energy of an electron with a 1-Å wavelength is 12,000 eV (= hard x-rays), and the energy of a photon with a 1-Å wavelength is 150 eV. Simply comparing the far superior resolution of electron microscopy with that of an optical microscopy confirms that in most instances the wavelength of an electron is far shorter than that of a photon. Moreover, electrons feature much more punch with a momentum given by $\mathbf{p} = m_e v$, where m_e is the electron's rest mass, far larger than the relativistic mass of photons of similar energies, with $m = h\nu/c^2$ (Equation 5.173). Remember also that for photons the speed is fixed, i.e., c for EM radiation is always 3×10^8 m/s, whereas the velocity of an electron, v, can be any value smaller than c.

The propagation of light is affected by the dielectric medium (refractive index), whereas the Coulomb potential affects the propagation of electrons. Solution of Schrödinger's equation in a 3D periodic Coulomb potential for an electron crystal forbids propagation of free electrons with energies within the energy bandgap. Imposing boundaries on photons by making them move in a material with a periodic dielectric constant in one, two, or three dimensions leads to photonic crystals as illustrated in Figure 5.119. A typical repeat unit for a photonic crystal is in the 100-nm range, whereas in a crystal the repeating units are the atoms (±1 nm) (see Figure 5.120). Photonic crystals generally possess photonic bandgaps (PBGs), ranges of frequency in which light cannot propagate through the structure. The forbidden range of frequencies depends on the direction of the light with respect to the photonic crystal lattice. For a sufficiently large refractive index

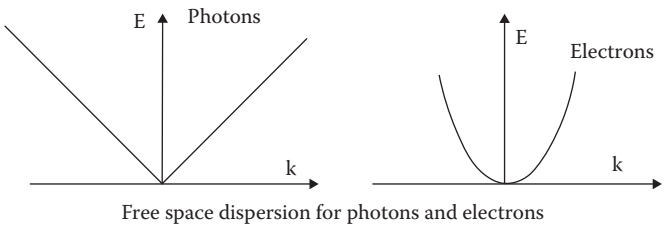


FIGURE 5.118 Dispersion relationships of photons and electrons in free space.

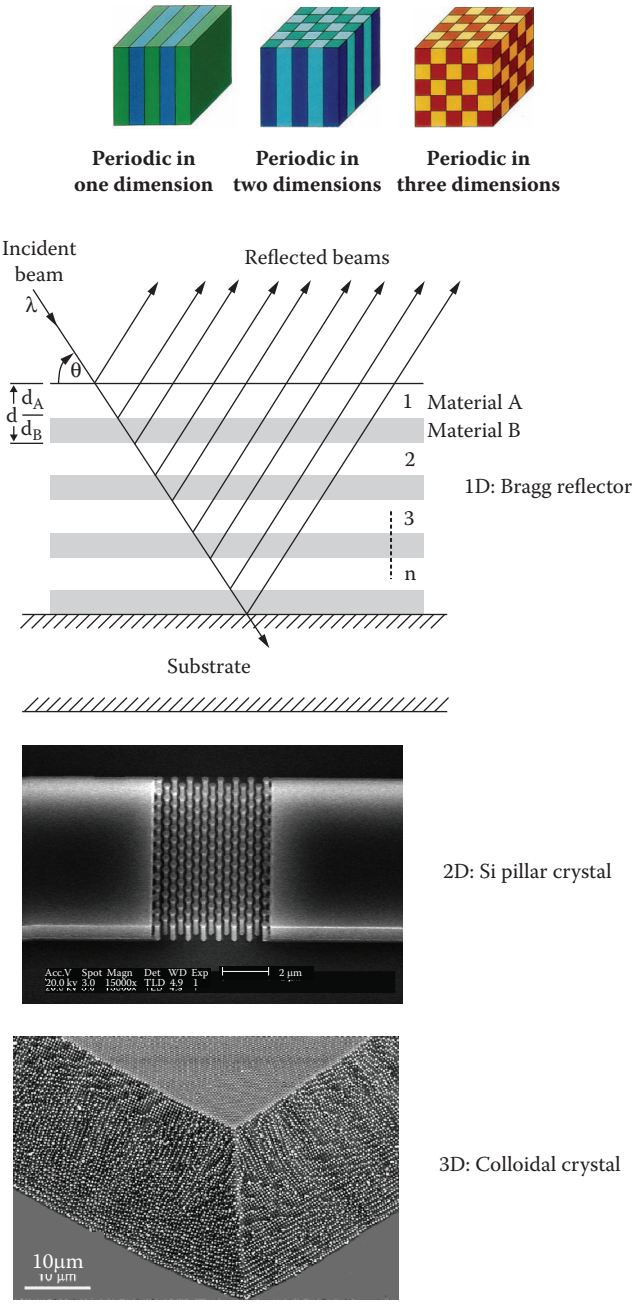


FIGURE 5.119 Photonic crystals are materials with a periodic dielectric constant in one, two, or three dimensions; typical repeat distance is 100 nm. Different colors represent different refractive indices. 1D photonic crystal: stack of alternating refractive index material in one direction; 2D photonic crystal: periodic variation of refractive index in two directions; 3D photonic crystal: periodic variation of refractive index in three directions.

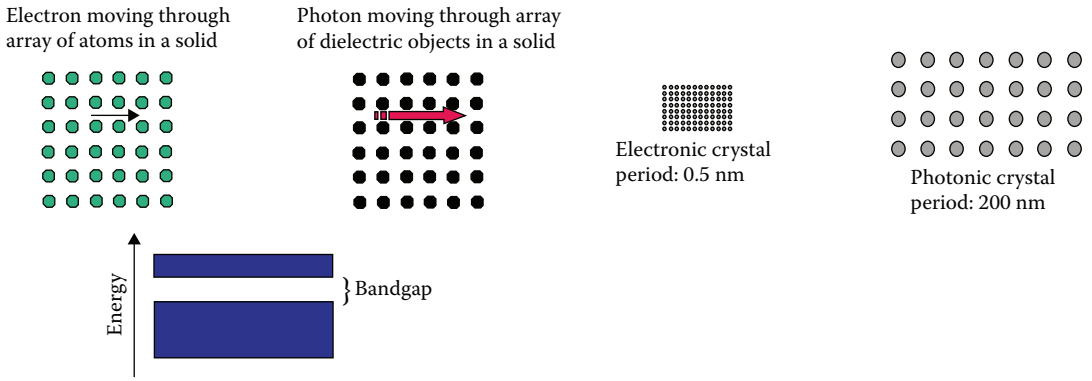


FIGURE 5.120 Similar to the periodic electron crystal lattice, one can fabricate a photonic crystal lattice. The refractive index varies with a much larger period of around 200 nm.

contrast (ratio n_1/n_2), there exists a bandgap that is omnidirectional. Each constitutive component in these periodic structures is perfectly transparent, but their periodic arraying blocks EM waves of certain frequencies. Photonic crystals may be regarded as semiconductors/insulators of light. Rather than a periodic array of atoms that scatter and modify the energy momentum of electrons, photonic crystals consist of a spatially periodic dielectric constant, with a periodicity of the order of the wavelength of light. They are mostly artificial structures that control the flow of light and are expected to lead to a new photonic era, with an impact similar to that of the electronic revolution brought on by semiconductors.

Inside a semiconductor, electrons experience a periodic potential as a result of the regularly spaced atomic cores in the crystal lattice, leading to multiple electron wave reflections and electron wave interference. The multiple reflections result in eigenmodes that are affected by the exact shape of the periodic potential $V(x)$. Felix Bloch pioneered the study of wave propagation in 3D periodic media in 1928, unknowingly extending an 1883 theorem in 1D by G. Floquet. Bloch proved that waves in such a medium could propagate without scattering, their behavior governed by a periodic envelope function multiplied by a plane wave (see Equations 3.198 and 3.199). Bloch studied the problem from the viewpoint of quantum mechanics. He postulated that electrons in a conductor scatter only from imperfections and not from the periodic ions. The same techniques can be applied to electromagnetism by casting Maxwell's equations as an eigenproblem in analogy with Schrödinger's equation.

The tunneling of electrons through a forbidden zone has its equivalent in light leaking out of a waveguide to create an evanescent field. Table 5.12 summarizes the above comparison of photons with electrons.

The Photonic Bandgap

During the past 17 years, John Maddox's lament (see quote at the beginning of this chapter) was heard loud and clear, and tremendous progress was made toward realizing photonic crystals. Stacks of dielectrics (with repetition in one direction, 1D) have been known since 1887 (Bragg reflectors). Also known and used were optically confined structures, with 1D confinement (planar waveguides), 2D confinement (fiberoptic cables), and 3D confinement (microspheres). The breakthrough in photonics in 1987 was the proposal of photonic crystals with periodicity in 2D and 3D (see Figure 5.119). This concept was introduced independently and simultaneously that year by two researchers, each following different paths. Photonic crystals were originally proposed as a means to realize two fundamental optical principles: the localization and trapping of light in a bulk material and the complete inhibition of spontaneous emission over a broad frequency range. Sajeev John, from the University of Toronto, was formulating an answer to the first question—whether localization of electrons in a disordered solid can be extended to photons in a strongly scattering medium—and he predicted that localized states of the EM field can be created in a periodic dielectric medium in three dimensions.[25] At the same time, Eli Yablonovitch, then at Bell

Communications Research in New Jersey, was trying to address the possibility of suppressing the unwanted spontaneous emission affecting semiconductor lasers, and he predicted that a periodic dielectric in thee dimensions could produce a forbidden gap in the EM spectrum.[26] Spontaneous emission of light is a natural phenomenon, which is of great practical and commercial importance. For example, in semiconductor lasers, spontaneous emission is the major sink for threshold current I_{th}, which must be surmounted to initiate lasing. In solar cells, spontaneous emission determines the maximum voltage available. Yablonovitch reasoned that if there are no optical modes available for photons to go to there cannot be spontaneous emission. From Fermi's golden rule (Equation 3.285), we know indeed that the density of final states is the density of optical modes available to the photon and an optical bandgap would indeed cut the optical modes that are available.

The origin of the photonic bandgap is best understood by looking back at the dispersion curve of a 1D photonic crystal [ω(**k**)] (i.e., a Bragg reflector). Similar dispersion curves in electronic crystals reveal the electronic bandgap. A 1D photonic crystal is simply a Bragg stack of dielectric planes, with a period of refractive index variation labeled a in Figure 5.121a. Light incident in the *z*-direction with a wave vector (**k**) is diffracted by this 1D photonic crystal into many possible directions, and the wave vectors of incoming and diffracted waves are related as:

$$\mathbf{k}_z - \mathbf{k}'_z = \pm N(2\pi/a) \qquad (5.265)$$

where *N* is the multilayer number (see also Equation 2.20). Thus, the incident and diffracted *z*-components of a wave vector are shifted with respect to each other by integral multiple of $(2\pi/a)$. A special case of diffraction is the Bragg diffraction, which occurs when incident light is totally reflected by the photonic crystal (it behaves like a metal!). This occurs whenever $\mathbf{k}_z$ is a multiple of (π/a):

$$\mathbf{k}_z = \pm N(\pi/a) \qquad (5.266)$$

as seen in Figure 5.120b. This wave cannot propagate, and its group velocity $(d\omega/d\mathbf{k})$ is zero. Figure

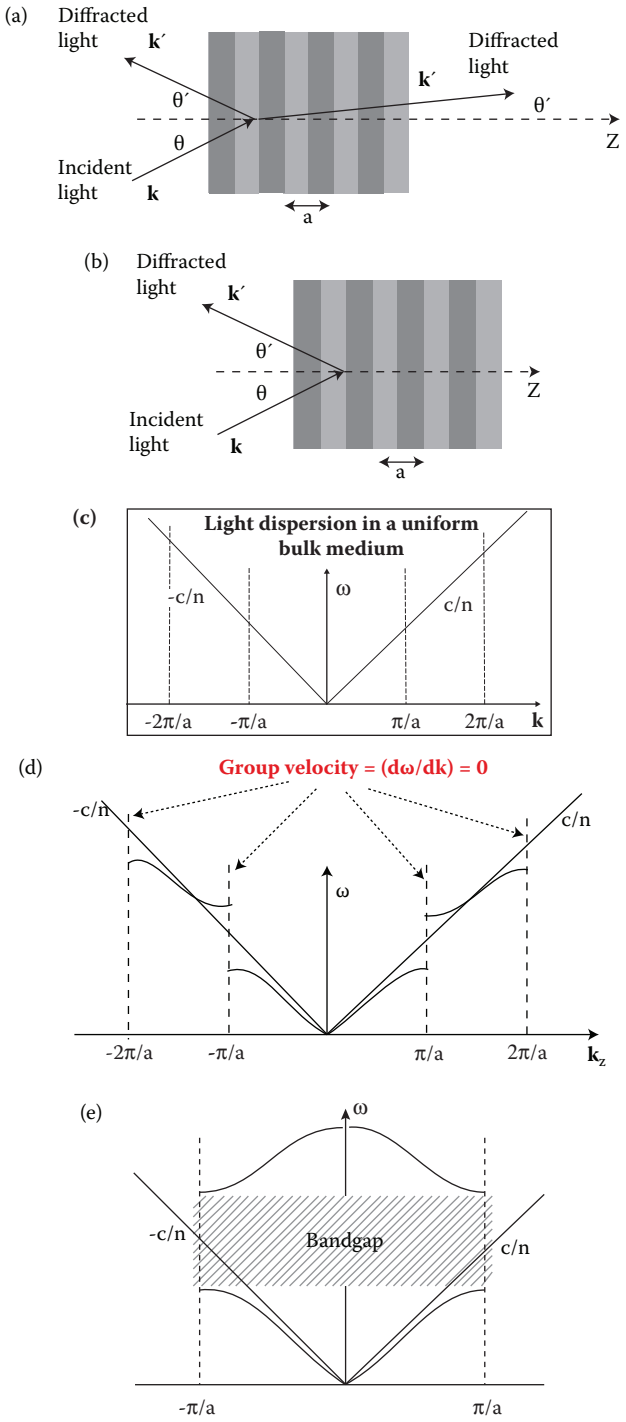


FIGURE 5.121 1D photonic crystal: a stack of alternating refractive index materials in the *z*-direction. (a) Light incident in the *z*-direction with a wave vector (**k**) is diffracted by the 1D photonic crystal into many possible directions. (b) A special case of diffraction is the Bragg diffraction, which occurs when incident light is totally reflected by the photonic crystal. (c) The dispersion character of light in a bulk medium with a uniform refractive index, *n*. (d) The dispersion curve of a 1D photonic crystal deviates from the straight-line dispersion curve of a uniform bulk medium. (e) The region of $\mathbf{k}_z$ between $-\pi/a$ and π/a is called the *first Brillouin zone*. The shaded area marks the photonic bandgap.

5.120c shows the dispersion character of light in a bulk medium with a uniform refractive index, *n*. In contrast, in Figure 5.120d, we notice that for the 1D photonic crystal, the dispersion curve deviates from the straight-line dispersion curve of such a uniform bulk medium. To ensure Bragg reflection for $\mathbf{k}_z = \pm N(\pi/a)$, the curve becomes horizontal when approaching π/a ($d\omega/d\mathbf{k} = 0$). It is necessary to plot the dispersion curve in Figure 5.120d only for values of $\mathbf{k}_z$ between $-\pi/a$ and π/a as illustrated in Figure 5.120e. All other values of $\mathbf{k}_z$ can be obtained by diffraction of waves with $\mathbf{k}_z$ between $-\pi/a$ and π/a. The region of $\mathbf{k}_z$ between $-\pi/a$ and π/a is called the *first Brillouin zone*. There is a band of frequencies that are forbidden for all possible values of $\mathbf{k}_z$. This is the bandgap for the 1D photonic crystal.

Compared with a 1D photonic crystal, it is relatively more difficult to deduce the dispersion curves for 2D and 3D photonic crystals. The software for doing this is publicly available at http://www.elec.gla.ac.uk/groups/opto/photoniccrystal/Software/SoftwareMain.htm.

Complete photonic bandgaps can only be obtained under favorable circumstances, i.e., with the right structures and sufficient (threshold) refractive index contrast. Ho et al.[27] showed theoretically that photonic "atoms" arranged in a diamond lattice structure may have a complete bandgap. Simple face-centered cubic photonic crystals, on the other hand, even with high dielectric contrast, do not yield a complete photonic bandgap.[28] The threshold refractive index contrast for complete bandgaps in 3D photonic crystals obviously depends on the exact periodic structure; e.g., for diamond it is 1.87, and for an inverse opal structure it is 2.80.

Manufacturing of 2D and 3D Photonic Crystals

In 1991, the first photonic diamond structure was fabricated by the group of Yablonovitch. This first photonic crystal was made by mechanically drilling holes, 1 mm in diameter, into a block of dielectric with a refractive index n = 3.6. The material, which became known as *Yablonovite*, prevented microwaves from propagating in any direction; in other words, it exhibited a 3D photonic bandgap.[28] The Yablonovite was the first "crystal" that showed experimentally a full 3D bandgap in the microwave range, and at these long wavelengths a mechanical fabrication method (drilling holes!) could be used. A schematic drawing of the Yablonovite is illustrated in Figure 5.122. Materials like this, in theory, guide and confine light without losses. A 3D photonic crystal provides the ultimate tool for the manipulation of light.

Several ingenious ways to produce these so-called *photonic crystals* for use in the microwave regime have been devised in addition to Yablonovitch's method. For example, in 1994 Ekmel Özbay, then at the Ames Laboratory in the United States, and his coworkers fabricated a photonic crystal with a bandgap at microwave wavelengths by stacking micromachined silicon wafers in a "woodpile" or "picket fence" structure.[29] The key to their success was the accurate alignment of successive layers—inaccurate alignment destroys the bandgap. Other structures that have bandgaps at microwave and radio frequencies are currently being used to make antennae that direct radiation away from the heads of mobile phone users.

Despite these successes in the microwave range, it took more than a decade to fabricate photonic crystals that work in the near-infrared (780–3000 nm) and visible (450–750 nm) regions of the spectrum. To block wavelengths in the infrared and visual spectrums, one needs to get a bit more creative. The main challenges are to find suitable materials and processing techniques to fabricate structures that

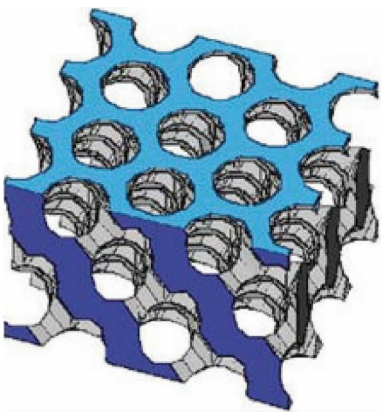


FIGURE 5.122 Yablonovite was the first crystal that showed experimentally a full 3D bandgap in the microwave range.

are about one thousandth the size of the microwave crystals. It becomes increasingly difficult to achieve the required accuracy as the dimensions of the structures are reduced and as the number of layers is increased in an attempt to make a device that operates at infrared or optical frequencies. As so often seems the case, nature is way ahead of the scientists: the iridescence of various colorful living creatures, from beetles to peacock feathers to butterflies [see chapter-opening image (b) before the Introduction of this chapter], is the result of photonic crystals. Unlike pigments, which absorb or reflect certain frequencies of light as a result of their chemical composition, the way that photonic crystals reflect light is a function of their physical structure. That is, a material containing a periodic array of holes or bumps of a certain size may reflect blue light, for example, and absorb other colors even though the crystal material itself is entirely colorless. Because a crystal array looks slightly different from different angles (unlike pigments, which are the same from any angle), photonic crystals can lead to shifting shades of iridescent color that may help some animals attract mates or establish territories. Another naturally occurring example of a photonic crystal is opal [see chapter-opening image (a) before the Introduction of this chapter]. Opals are naturally occurring 3D photonic crystals: their microstructure consists of silica spheres about 150–300 nm in diameter, which are tightly packed into repeating hexagonal or cubic arrangements. Synthetic opals use this same pattern, although they can be made from different materials. A type of photonic crystal made using the gaps between the spheres is called an *inverse opal*. Self-assembly methods enable the fabrication of large polymer opals and inverse opals at low cost. However, the contrast in the refractive index in opal is rather small, which results in a rather small bandgap. Yablonovitch pointed out that the larger the difference in refractive indices between the two repeating materials of the photonic crystal, the more extremely light is influenced.

Two-dimensional (2D) photonic crystals are relatively easy to fabricate using Si micromachining and UV and electron beam lithography. Periodic arrays of macropores in single-crystal Si were first shown by Lehmann and Föll in 1990; some examples of this work are presented in Volume II, Figure 4.75.[30] Grüning et al. realized a bandgap in the near IR using such macropores in n-type silicon etched under back side illumination.[31] If one retains a periodicity of an alternating index of refraction in the *x*-*y* plane while the *z*-direction remains unmodulated, a 2D photonic crystal results as shown in Figure 5.123. Grüning et al. obtained for the first time a complete 2D bandgap in the near infrared (at a wavelength of about

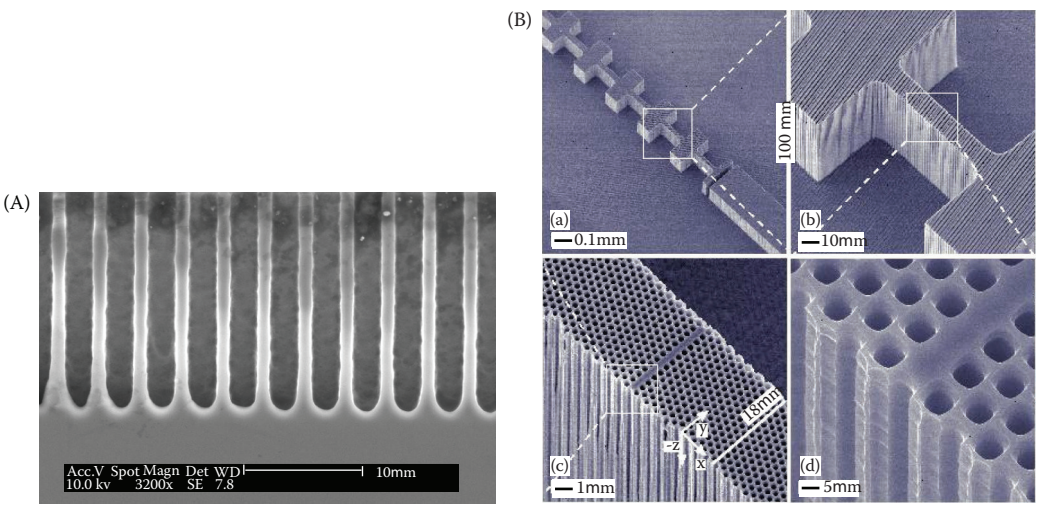


FIGURE 5.123 (A) SEM image of a slab of macroporous silicon, representing a 2D photonic crystal (a). The pore walls are about 100 μm tall and about 25 μm wide in the direction of transmission (b). Omitting some pores yielded a waveguide structure (c). The extremely smooth finish of the structure is clearly visible (d). (Courtesy of Dr. R. Wehrspohn.) (B) Macropores in n-type silicon etched under back side illumination form a 2D photonic crystal.[31]

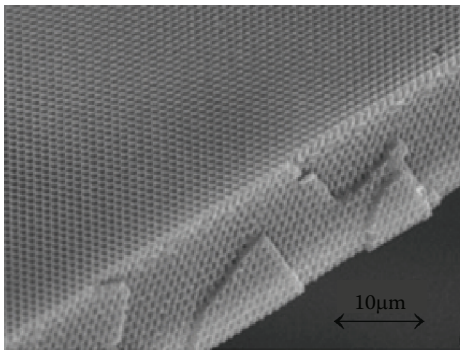


FIGURE 5.124 A 3D photonic crystal fabricated using holographic lithography.[32]

4.9 μm). A complete bandgap in this case means that a bandgap always exists independent of the polarization and propagation of the light wave.

As pointed out earlier, fabrication of 3D photonic crystals is more difficult than 2D crystals. In addition to the methods listed above, it has been achieved by sedimentation of monodisperse colloidal nanospheres of polystyrene or silica and by using holographic two-photon lithography. Two-photon lithography (see Volume II, Chapter 2) is based on the concept that electronic transitions that would normally require absorption of one UV photon for excitation can under some circumstances be accomplished by simultaneous absorption of two IR photons (two-photon overlap in time and space for exposure). When this high-intensity IR light enters a photoactive material, the excitation reaction is confined to a focal point at a precise depth in a substrate. Because two-photon absorption takes place only close to the focal point of a laser, one can fabricate structures deep within the volume of the polymer matrix. Atomic layer deposition (Volume II, Chapter 7) has also been used to build photonic crystals for coatings inside porous alumina and to fabricate inverted opals. The current experimental efforts are directed at decreasing the disorder present in the structures and increasing their scalability, and in the meantime, there is a continuous search for designing new structures with more robust bandgaps.

A 3D photonic crystal fabricated using holographic lithography is shown in Figure 5.124. In Figure 5.125 we show a self-assembled silica opal grown on a silicon substrate, and an inverse opal is made using LPCVD to fill the opal template with silicon, and wet etching yields the inverse opal silicon structure. By 1998, artificial 3D photonic crystals in the near infrared and visible were fabricated based on inverted opals, and in 1999 the first 3D photonic crystal-based optical devices (lasers, waveguides) were made. Another way of making photon crystals is by using block copolymers as described in Volume II, Chapter 2.

Defects in Photonic Crystals

As light is forbidden within the photonic bandgap, we can anticipate that we can manipulate the behavior of light using photonic crystals just as we can manipulate electrons using a semiconductor. Most proposals for devices that make use of photonic crystals do not use the properties of the crystal directly but make use of defects. Defects are made when the lattice is changed locally, and as a result, light with a frequency inside the bandgap can now propagate locally in the crystal, i.e., at the position of the defects. It remains impossible, however, to

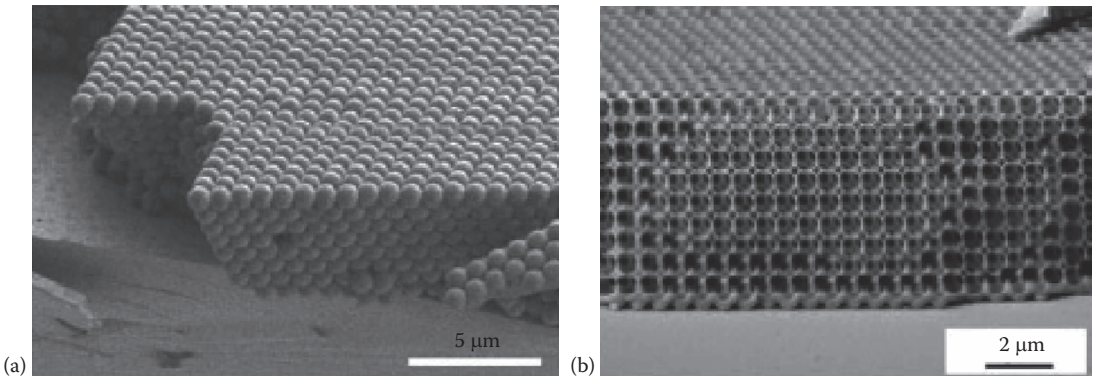


FIGURE 5.125 (a) Self-assembled silica opals grown on silicon substrate. (b) LPCVD is used to fill opal template with silicon; wet etching yields inverse opal silicon structure.[33]

propagate in the surrounding photonic crystal material. For example, if we consider the introduction of a light-emitting layer, with the same wavelength as the photonic bandgap in the PC, light emission is impossible as light is forbidden to be present in the photonic bandgap. This leads to the control of spontaneous emission as first recognized by Yablonovitch. However, if we disrupt the periodic characteristics of the photonic crystal by introducing defects, energy levels corresponding to the defects are formed in the photonic bandgap. Thus, light is now permitted at the energy levels corresponding to those defects. This in turn enables the realization of single-mode light-emitting diodes (LEDs) and zero-threshold lasers. In addition, if we introduce a line of defects, the propagation of light through the defects is possible, and as light propagation is forbidden outside the defects, a no-loss microoptical circuit can be formed. Therefore, to provide functionality to photonic crystals, the introduction of controlled defects is necessary; the importance of defects in photonic crystals is comparable to that of dopants in semiconductors.

Most of the work on introducing controlled defects is carried out in 2D photonic lattices, so-called *PC slabs*. These slabs exhibit some of the useful properties of a truly 3D photonic crystal and are far simpler to make with current IC manufacturing type of equipment. These structures can block certain wavelengths of light at any angle in the plane of the device and can even prevent light entering from certain angles in the third dimension (i.e., perpendicular to the surface). Thus, 2D photonic crystals are a good compromise for many applications and are easily incorporated within planar waveguides.

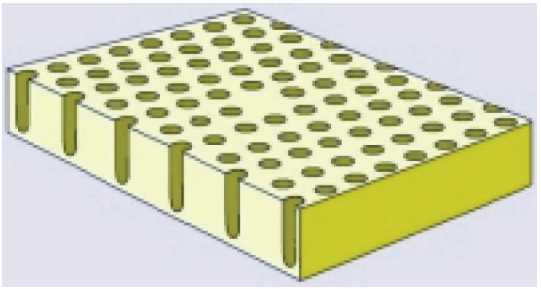


FIGURE 5.126 Point defect as high-*Q* resonator cavity in photonic crystals.

Point defects in a photonic crystal lattice (Figure 5.126) remove a single dielectric unit and produce microcavities that have size-dependent radiation modes. They are analogous to a single F-center (atom vacancy) defect in an electronic material. A microcavity resonant for wavelength λ satisfies $D = N(\lambda/2)$, where N is an integer and D is the characteristic size of the cavity. These "microcavities," just like in electronic crystals, produce defect states in the photonic bandgap. They come with a very high resonance Q (10^4–10^8).

Line defects in a photonic crystal remove lines of a dielectric material and act as waveguides confining photons (Figure 5.127). This is analogous to a line of F-centers (atom vacancies) for electronic

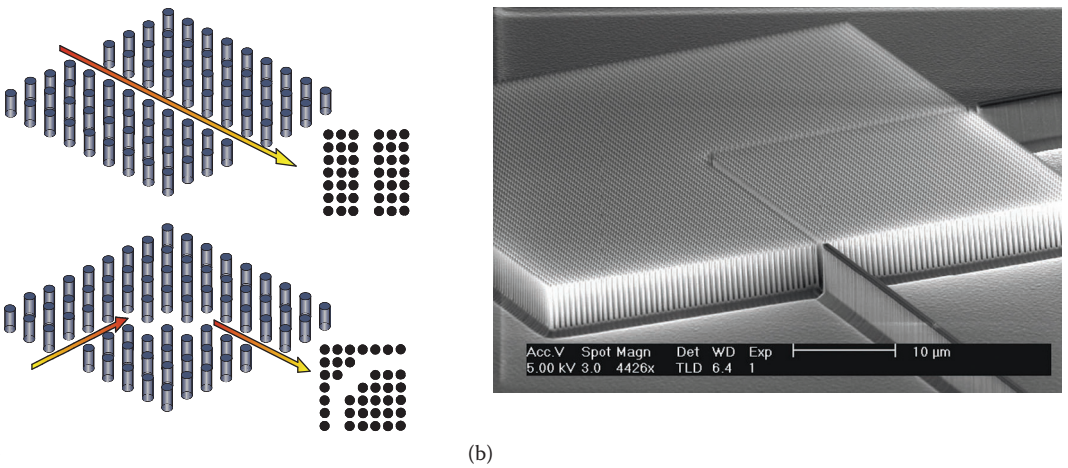


FIGURE 5.127 (a) Line defects in PC slabs. (b) 2D Silicon photonic crystal waveguide bend. Zijlstra, van der Drift, De Dood, and Polman (DIMES, FOM) [Delft Institute of Microelectronics and Submicron Technology (DIMES), Foundation for Fundamental Research of Matter (FOM); http://www.dedood.demon.nl/PhotonicCrystals.html].

defects. The *E* field is confined to the defects and cannot propagate in the rest of the material. Unlike the conventional refractive waveguides, these photonic waveguides work by diffraction. Modes that cannot propagate through the photonic crystal lattice (i.e., those within the bandgap) can propagate easily within the waveguides, including those waveguides that feature bends at right angles. The defect mode cannot penetrate the crystal in the *x*-*y* plane because of the bandgap but extends in the *z*-direction. Whereas photonic crystal waveguiding is based on diffraction, conventional optical fibers and waveguides operate based on total internal reflection. Fiberoptic technology has revolutionized telecommunications owing to the speed of data transmission, which is equivalent to >3 h of television per second or 24,000 simultaneous telephone calls (0.1 kg of fiber carries the same information as 30,000 kg of copper cable). But this benefit of speed comes with a penalty: fiber optic cables are large.

For an optical fiber, the contrast in refractive index between the glass core of the fiber and the surrounding cladding material determines the maximum radius through which light can be bent without losses. For conventional glass fibers, this bending radius is a few millimeters. However, the interconnects between optical components on a dense integrated optical circuit require bending radii of 10 μm or less! This is where photonic crystal-based waveguides offer a solution. It is possible to form narrow-channel waveguides within photonic crystals by creating a row of holes in an otherwise regular pattern (see Figure 5.127). Light will now be confined within that line of defects for wavelengths that lie within the bandgap of the surrounding photonic crystal. Because the surrounding material has no available modes at this wavelength, an optical quantum well forms in the waveguide region and traps the light. Under these conditions we can introduce a pattern of sharp bends that will either cause the light to be reflected backward or directed around the sharp bends.

Thus, we can create point defects corresponding to resonant cavities, or line defects behaving like waveguides.

A great effort in studying defects in photonic crystals, as building blocks for photonic integrated circuits, has been carried out by Professor John D. Joannopoulos' group at MIT, and the interested reader can refer to their publications (e.g., *Photonic Crystals: Molding the Flow of Light*[34]).

Super Refraction or Super Prism Effect

The prism effect (see Figure 5.128) refers to separation of colors by refraction through a prism. This is related to the dispersion or variation of refractive index with wavelength, i.e., the derivative $(dn/d\lambda)$ of the refractive index with wavelength. In a normal bulk medium like glass, dispersion is small, but $dn/d\lambda$ can be made unusually large in photonic crystals. This can be understood from the dispersion curves for photonic crystals mentioned earlier (see Figure 5.121).

Photonic Crystal Optical Fibers

Photonic crystal optic fibers are a special class of 2D photonic crystals where the dimension of the medium perpendicular to the crystal plane can be hundreds of meters long. These types of fibers are known by several names: photonic crystal fiber, photonic bandgap fiber, holey fiber (photonic crystal fiber using air holes in their cross-section), microstructured fiber, and Bragg fiber (Figure 5.129).

Besides the bending losses, another major drawback associated with conventional optical fibers is that different wavelengths of light travel through the material at different speeds. Over long distances, time delays then occur between signals that are encoded at different wavelengths. This dispersion is worse if the core is large, as the light can follow more different paths or "modes" through the fiber. A pulse of light traveling through such a fiber will broaden out, thereby limiting the amount of data that can be sent through it. Adding new signals at other optical wavelengths can also increase the fiber data transfer capacity of an optical fiber, a method known as *dense wavelength division multiplexing* (DWDM). Unfortunately, the optical fiber is only transparent over a small range of wavelengths, so the number of separate conversations that can be transmitted depends on the line width of neighboring optical channels (which is currently at the subnanometer level). These problems can be solved effectively using the "holey fiber" developed by Philip Russell

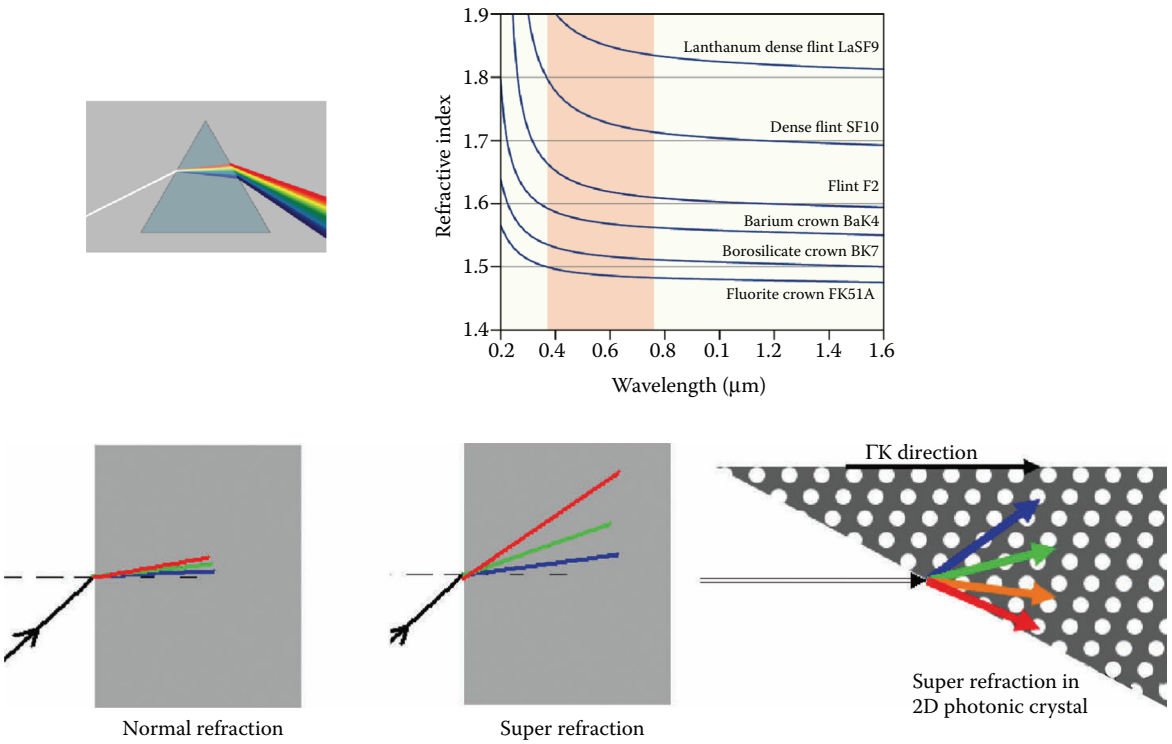


FIGURE 5.128 Super refraction or super prism effect compared with normal refraction.

and Jonathan Knight and coworkers at the University of Bath.[35] This fiber has a regular lattice of air cores running along its length and transmits a wide range of wavelengths without suffering any dispersion. One version of the photonic crystal fiber is made by packing a bundle of glass capillary tubes around a solid glass core that runs through the center of the bundle. This assembly is then heated and stretched out to create a long monolithic fiber that is only a few micrometers in diameter. The resulting fiber has the unusual property that it transmits a single mode of light, even if the diameter of the core is very large. Frequencies in the bandgap propagate within the fiber core, which is like a defect in a 2D photonic crystal. The same Bath team created an even more unusual photonic crystal fiber by removing the

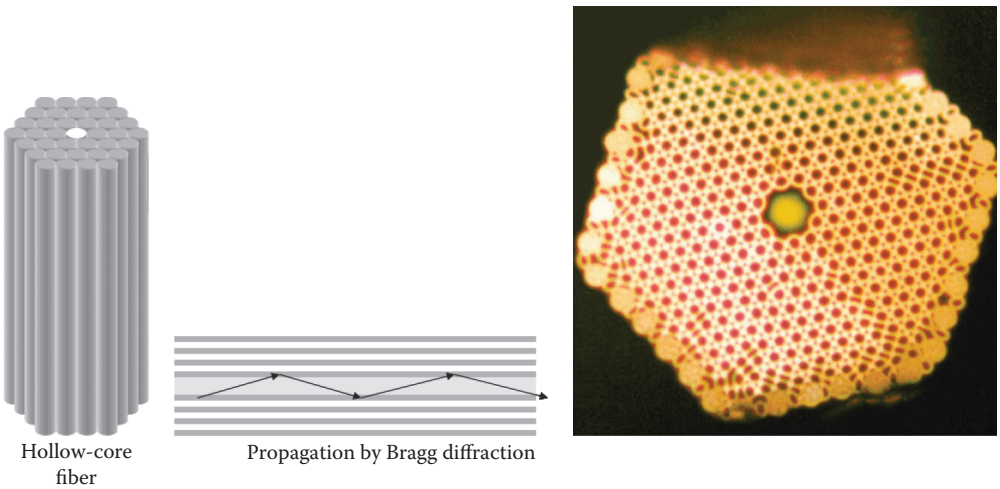


FIGURE 5.129 Photonic crystal fiber, photonic bandgap fiber, holey fiber, microstructured fiber, or Bragg fiber are all used as synonyms in this field.

central solid glass core to form a long air cavity. In this case, the light is guided along the low-refractive-index air core by the photonic bandgap confinement effect. Because the light is not guided by the glass material, very high-power laser signals could potentially be transmitted along such a fiber without damaging it. Because light travels in air, group velocity dispersion is zero for all wavelengths.

In a hollow-core fiber, even those wavelengths that experience high losses in conventional fibers can propagate. Propagation over ~1000 m has been demonstrated for wavelengths like 1.5 μm and 10 μm with an attenuation of only around ~1 dB/km (can be made lower than index guiding fiber). By filling the hollow core with gas, the fiber can be also used as a very sensitive gas sensor.

Applications

It is easy to see why photonic crystals are generating so much interest. Photonic bandgap crystals hold the promise for several fantastic integrated optics applications, including thresholdless microlasers, better signal filters, photonic computers, lossless mirrors and lossless waveguides (even in bends)/waveguide couplers, photonic bandgap optical fibers, and so on. The impact this will make can be gauged by looking at today's information technology (IT). IT derives its power from our ability to manipulate the flow of electrons and holes in semiconductors in the most intricate ways; photonic crystals promise to give us even greater flexibility because we have far more control over the properties of the larger photonic crystals than we do over the electronic properties of semiconductors. Structure possibilities are limited only by our imagination. The benefits of photonic crystals include the ultrahigh refractive indices, the ability to control spontaneous emission, to modify the state density and group velocity, and to locate light by introducing structural defects. This way, photonic crystals could eventually address many of the problems that currently limit the speed and capacity of communications networks. For example, these structures could be used to create novel light-emitting diodes (LEDs) and lasers that emit light in a very narrow wavelength range, together with highly selective optical filters that could be integrated on one and the same chip.

A short sampling of photonic crystal applications follows:

1. Low-threshold microlasers. Within a few years a number of basic photonic crystal applications are expected to start making an appearance in the marketplace. Among these will be highly efficient photonic crystal lasers and extremely bright LEDs. Point defects in photonic crystals act as high-Q resonator cavities and are used in low-threshold photonic crystal lasers. In photonic lasers, only the wavelength that matches the wavelength of the defect mode is amplified because it can propagate freely through the material. The intensity of the propagating light increases as it undergoes successive reflections and travels back and forth through the photonic crystal. Meanwhile, light at other wavelengths is trapped within the photonic crystal and cannot build up. This means that the laser light is emitted in a narrow wavelength range that is directly related to the diameter of the microcavity.
2. Highly efficient light-emitting diodes (LEDs). LEDs, like lasers, play a key role in optical communication systems. Encoding and sorting thousands of multiplexed wavelength channels poses quite a problem: at the transmitting end of an IT system, each wavelength channel requires a very stable light source that only emits in a very narrow range of wavelengths. Although LEDs offer high switching speeds, they emit light over a wide range of wavelengths, which makes them less suitable than lasers for dense wavelength division multiplexing (DWDM) systems. LEDs are made from photoemissive materials that emit photons once they have been excited electrically or optically. These photons are typically emitted in many different directions and also have a range of wavelengths, which is not ideal for communications applications. Photonic crystals can be used to design a mirror that reflects a selected wavelength of light from any angle with high efficiency (Figure 5.130). Moreover, they can be integrated within the photoemissive layer to create an LED that emits light at a specific wavelength and direction.

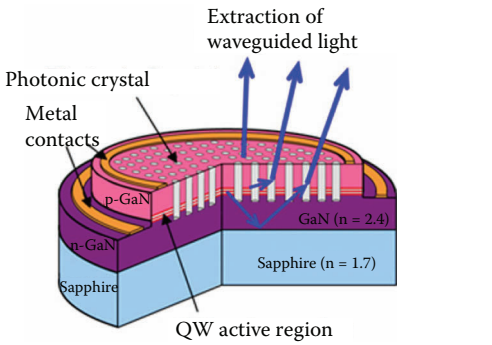


FIGURE 5.130 GaN-based diode with photonic crystal slab.

3. Filters with a very narrow range of wavelengths. Photonic crystal microcavities that are fabricated from passive materials, such as silicon dioxide and silicon nitride, are used to create filters that only transmit a very narrow range of wavelengths.
4. Micrometer-sized optical benches. To develop a micrometer-sized optical bench one first needs to develop small-scale optical "interconnects"—tiny planar waveguides that can steer light around tight corners. Although it is straightforward to route electrons around sharp bends in microchips, it is impossible to direct light in the same way using conventional glass waveguides because the losses at the bends are so large. Photonic crystal waveguides can be used to route light around micrometer-sized optical benches. Together with the very narrow range photonic crystal filters, this could lead to a viable "spectrometer on a chip." These developments are expected within the next 5 years.
5. On a 5–10-year timescale it is expected that the first photonic crystal "diodes" and "transistors" will be demonstrated. A demonstration of the first photonic crystal logic circuit could even take place in the next 10–15 years, whereas a prototype optical computer driven by photonic crystals could be available within the next 25 years.
6. Synthetic opals might find a niche in the valuable jewelry and artwork markets, and thin photonic crystal films could also be used as anticounterfeit devices on credit cards.
7. Negative refractive index. Photonic crystals can exhibit an effective negative refractive index as explained below under "The μ-ε Quadrant and Metamaterials." Whereas photonic crystals occur in nature, metamaterials do not. Metamaterials are engineered composites tailored for specific EM properties that are not found in nature and not observed in the constituent materials. Cubukcu et al.[36] have experimentally demonstrated negative refraction by a photonic crystal.

Although complex photonic crystal LED and laser structures are being developed, fundamental questions about the formation, operation, and efficiency of photonic bandgap materials remain unanswered. Real photonic crystal waveguides are far from ideal because they suffer losses as a result of diffraction and scattering and because of an optical mismatch at the surface joining the device to input and output waveguides. These loss mechanisms need to be fully understood and the devices need to be optimized for particular wavelengths before they can become commercially viable.

The μ-ε Quadrant and Metamaterials

Introduction

As we saw in Chapter 2, material properties are determined by the properties of the subunits (atoms/molecules) plus their spatial distribution. In Figure 5.131 we show a diagram of the EM properties as a function of the ratio of a, the "lattice constant" of a material/structure, and λ, the wavelength of the incoming light.

- Natural crystals (in Figure 5.131 we show the stunning violet tanzanite as an example) have lattice constants much smaller than the light wavelengths: $a \ll \lambda$. In this case, the lattice constant a is the distance between constituent atoms. Such materials are treated as homogeneous media with parameters ε and μ, which are tensors in anisotropic crystals. They have a positive refractive index, $n > 1$, and show no magnetic response at optical wavelengths ($\mu = 1$).
- In the above case, the EM wave is much longer than the typical lattice distance a ($\lambda \ggg a$), and one can apply effective medium theory; however, when a is in the same range of the

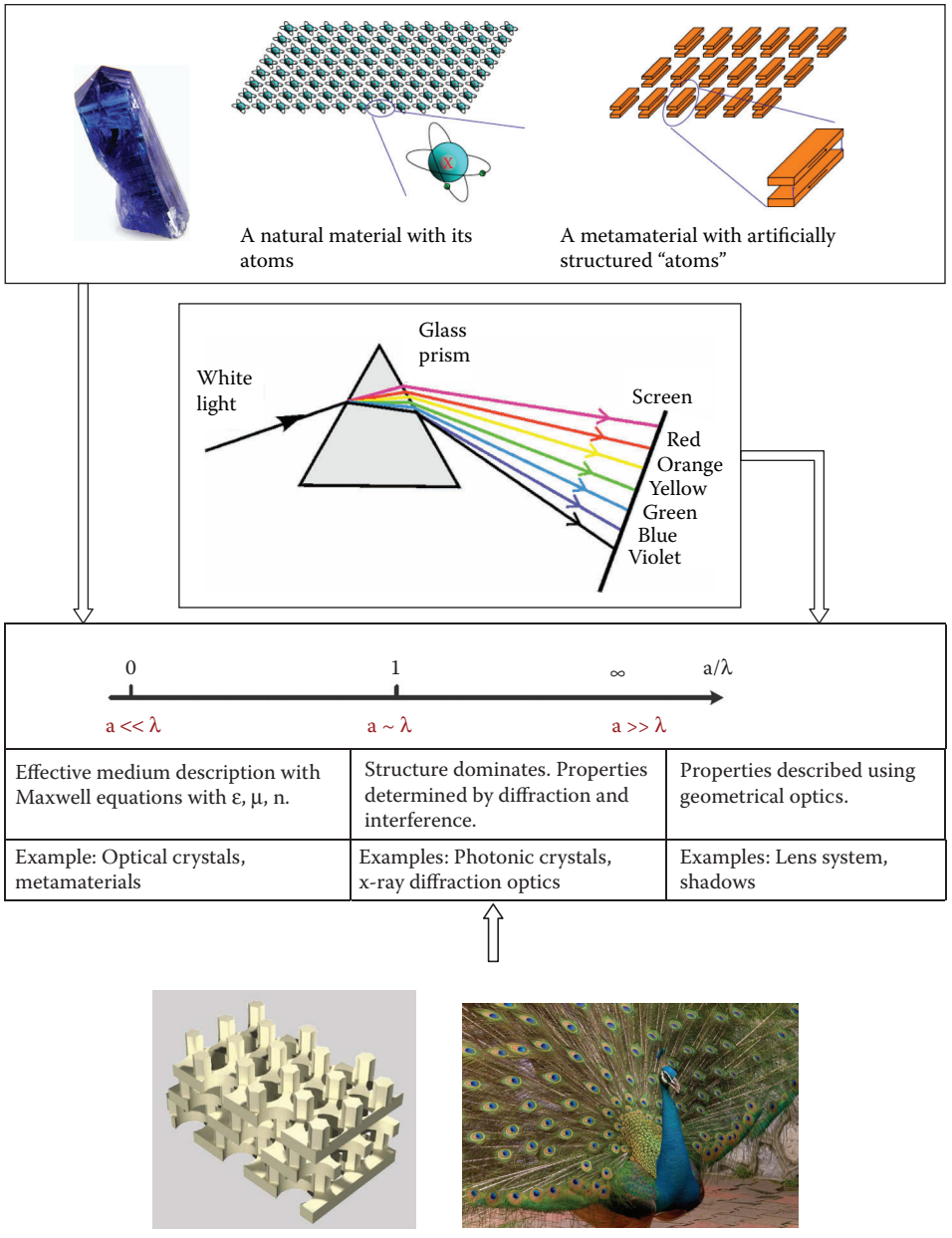


FIGURE 5.131 Electromagnetic properties as a function of the ratio a of the "lattice constant" of a material/structure and λ the wavelength of the incoming light.

wavelength of the incoming light, one defines a photonic crystal, a mesoscale material with subunits bigger than atoms but smaller than the EM wavelength. In photonic crystals *a* is the distance between repeat units with a different dielectric constant, so we can still call it a "lattice constant." Photonic crystals have properties governed by the diffraction of the periodic structures and may exhibit a bandgap for photons. They typically are not described well using effective parameters ε and μ and may be artificial or natural (in the figure we show an artificial 3D photonic crystal and peacock feathers as an example of natural photonic crystals).

- When considering macroscopic optic components with a critical dimension $a >>> \lambda$, geometric optics do apply. In this case *a* might be the repeat unit of a grating, the aperture of a lens, or the length of a side of an optical triangular prism.

Metamaterials are artificially engineered materials possessing properties (e.g., mechanical, optical,

electrical) that are not encountered in naturally occurring materials. *Meta* originates from the Greek word μετα: "after/beyond." The dimensions of the unit cell are less than the wavelength of the excitation EM wave; thus, the effective media theorem can be applied. We can design and control the properties of metamaterials. Novel properties, such as negative electric permittivity, negative magnetic permeability, negative refractive index, and so on, have been explored. Photonic crystals often behave like—but are not—true metamaterials, and they do occur in nature. *Metamaterials* is a general EM term: an object that gains its (EM) material properties from its structure rather than inheriting them directly from the materials it is composed of.

The μ-ε Quadrant

We can classify all materials according to their specific conductivity σ, their magnetic permeability μ, and their dielectric constant (or permittivity) ε. Materials with $\sigma > 0$ are conductors, and those with $\sigma \approx 0$ are insulators or dielectrics. All materials have two fundamental EM parameters: permeability μ and permittivity ε, which respectively measure the capacities of a medium to form magnetic and electrical fields. The values of those parameters produce the characteristic bending of a light beam when it travels from one medium into another. In addition, because both parameters are usually positive in nature, the electric and magnetic vector field components are directed according to the "right-hand rule," which can be represented by pointing the index finger of the right hand in the direction of propagation. The thumb and middle finger are then oriented at right angles to the index finger, showing the field vector directions. Mathematically, from the boundary conditions in Maxwell's equations, we concluded earlier that for a positive μ and ε, the vectors **E** and **H** and **k** constitute a right set of vectors and define right-handed materials (RHMs) as illustrated in Figure 5.30a. Most materials do come with $\varepsilon > 1$ (e.g., $\varepsilon = 12$ for Si and 2.25 for glass), but this is not without exceptions; we know that materials with a negative ε at optical frequencies (visible, IR) include metals such as Au, Ag, and Al (see, e.g., Figure 5.68). Most materials have $\mu \approx 1$ at optical frequencies and are nonmagnetic, but if $\mu \neq 1$, one deals with a magnetic material. With $\mu > 1$ we have a paramagnetic material at hand; with $\mu < 1$ it is diamagnetic. (Diamagnetic materials have a relative magnetic permeability that is less than 1, thus a magnetic susceptibility that is less than 0, and are therefore repelled by magnetic fields.) And finally with $\mu \gg 1$ we define a ferromagnetic material, which is associated with hysteresis effects because of fields. Materials with a negative μ are again the exception; they include resonant ferromagnetic or antiferromagnetic systems at microwave frequencies.

In 1968, Victor Veselago, a Russian physicist, showed that if ε and μ are both negative, we are forced by the Maxwell's equations to choose a negative square root for the refractive index:

$$n = -\sqrt{\varepsilon\mu}, \varepsilon < 0, \mu < 0 \qquad (5.267)$$

For such a material with μ and ε negative, the vectors **E** and **H** and **k** form a left set of vectors, and one refers to them as left-handed materials (Figure 5.30b). For right-handed materials, the phase velocity ($v_p = \omega/\mathbf{k}$ along **k**) and group velocity ($v_g = d\omega/d\mathbf{k}$ along the Poynting vector **S**) are in the same direction, but in left-handed materials group and phase velocity are in opposite directions (see Figure 5.32).

The dispersion equation:

$$\mathbf{k}(\omega) = \frac{\omega}{v} = \frac{n(\omega)\omega}{c} = \frac{\omega}{c}\sqrt{\varepsilon\mu} \qquad (5.81)$$

which enables one now to construct a μ-ε plane (see Figure 5.132 for a few different representations of such a diagram). For most natural materials, ε and μ are positive, $n > 0$ (top right), and waves propagate through the medium. If either ε or μ is negative, n is imaginary, and there is no wave propagation through the material. A substance with simultaneously negative values of permittivity ε and permeability μ (bottom left quadrant) does have a negative refraction index, as $n = \sqrt{\varepsilon}\sqrt{\mu} = -\sqrt{|\varepsilon\mu|}$.

Left-Handed Materials

Left-handed materials (LHMs) have EM properties that are distinct from any known material and

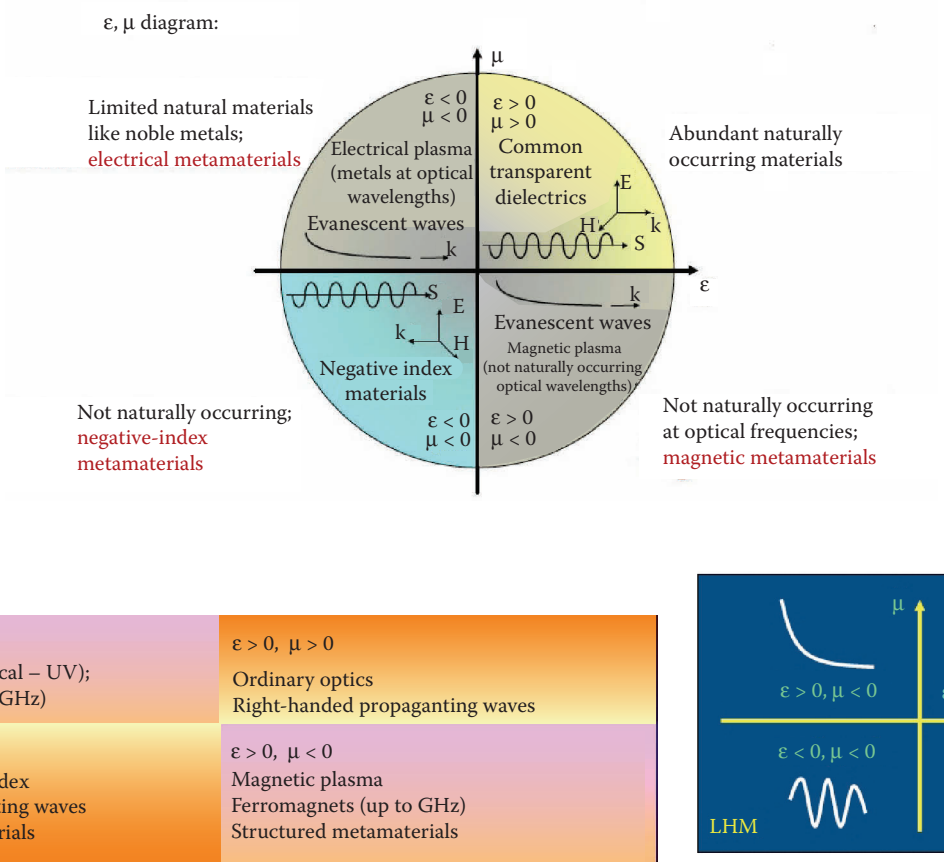


FIGURE 5.132 Various representations of ε-μ plane and typical materials.

hence are examples of metamaterials. Veselago expected that such a left-handed metamaterial would, besides a negative refractive index n, also have a reverse Doppler effect, backward Cerenkov radiation,* and make it possible to design perfect lenses and have a phase velocity directed against the flow of energy. In addition, the direction of the refracted light bends away from the normal to the interface between two media, rather than toward the normal, as in Snell's law (see metawater in Figures 5.17 and 5.133).

There are no known naturally occurring left-handed or negative-index (*n*) materials, but today experiments have shown that artificially designed structures/materials can act as negative refractive index materials in the microwave region and even in the visible.[37]

Veselago's fascinating prediction that perfect lenses could be made from metamaterials lay dormant until about 1996, when John Pendry, a physicist at Imperial College in London, showed that certain metal structures could be engineered to respond to electric fields as though the field parameters were both negative. Most electric resonances for natural materials are in the terahertz region or

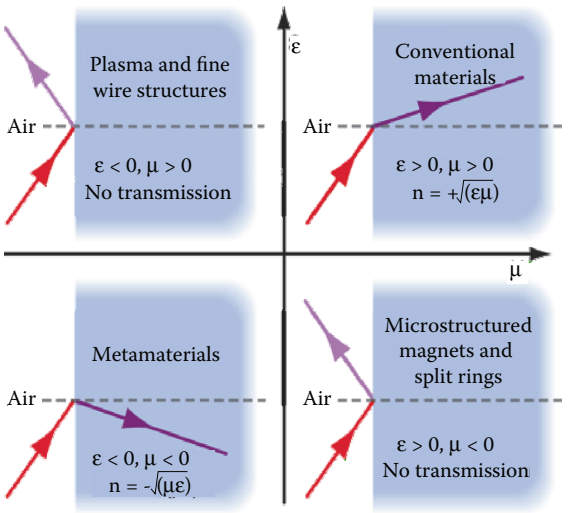
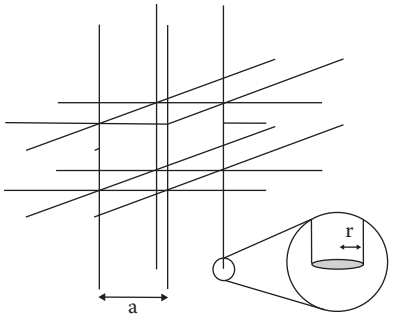


FIGURE 5.133 Refraction and the ε-μ plane.

* Cerenkov radiation (also spelled *Cherenkov*) is electromagnetic radiation emitted when a charged particle passes through an insulator at a speed greater than that of light in the medium.

higher. Magnetic natural systems typically have resonances through the gigahertz (a few magnetic systems have resonances up to terahertz frequencies, e.g., MnF_2, FeF_2). Negative ε, with relatively low loss in the visible region, we know from the Drude-Lorentz model, can be obtained for metals such as Au and Ag. Natural materials with a negative μ also exist; they include resonant ferromagnetic or antiferromagnetic systems at microwave frequencies. Unfortunately, electric and magnetic resonances do not overlap in existing materials, but, as we will see, this restriction does not exist for artificial materials!

The Drude model, we saw, describes the negative dielectric constant of conducting materials and is a basic type of material response with $\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega(\omega + i\gamma)}$. The important point is that ε is essentially negative below the plasma frequency, at least down to frequencies comparable to γ. The plasma frequency ω_p is given by $\omega_p^2 = \frac{ne^2}{\varepsilon_0 m_e}$ (with n the electron density and m_e the electron mass), and for many metals falls in the UV region. At lower frequencies, the permittivity of most metals is imaginary. The damping factor γ is given by $\gamma = \frac{\varepsilon_0}{\sigma}$ with σ the electric conductivity. For aluminium $\omega_p = 15$ eV and $\gamma = 0.1$ eV. On a macroscale, Pendry et al.[38,39] found a metamaterial with $\varepsilon < 0$ can be fashioned by a simple periodic structure composed of thin, infinite wires arranged in a simple cubic lattice (Figure 5.134). A periodic array (lattice constant a) of thin metal wires (radius r) with $r \ll a \ll \lambda$ acts as a low-frequency plasma. The effective ε is described with a modified ω_p, and the plasma frequency depends on geometry rather than on material properties. Although the concept of using wires to mimic a Drude material had been around in a practical sense since the 1940s (or earlier), Pendry reintroduced the idea in 1996 and noted the important connection with optical plasmon phenomena. With wire spacing $a = 5 \times 10^{-3}$ m and a radius of the wires r of 1.0×10^{-6} m, one calculates an ω_p of about 8.3 GHz (see Figure 5.134). Note that the wire model brings down the plasma frequency substantially, even to the microwave regime, but the same Drude equations apply.



$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2 - \omega_0^2} \qquad \omega_p^2 = \frac{2\pi c_0^2}{a^2 \ln\left(\frac{a}{r}\right)} \approx [8.2\ \text{GHz}]^2$$

FIGURE 5.134 On a macroscale, a metamaterial with $\varepsilon < 0$ can be fashioned by a simple periodic structure composed of thin, infinite wires arranged in a simple lattice, which mimics the response of a plasma. The plasma frequency is in the gigahertz band instead of the ultraviolet band.[39]

As mentioned above, both negative ε materials and negative μ materials exist in nature but in different frequency domains. Natural magnetic responses (ferromagnetic/ferrimagnetic/antiferromagnetic resonances) typically occur at lower than a few gigahertz, and because there are no free magnetic monopoles, there is no such thing as a magnetic plasma and conventional optics (lenses, mirrors, crystals, etc.) essentially have nothing to do with magnetic fields. Pendry et al.[40] also discussed how to obtain negative permeability in metamaterials. Magnetism in materials originates from the orbital motion of electrons and unpaired electron spins. The magnetic response of most natural materials fades away in the gigahertz region. But Pendry realized that conducting, nonmagnetic split-ring resonators (SRRs) can realize artificial magnetism even in nonmagnetic Cu. The magnetic flux-induced current loops form magnetic dipoles, and the intrinsic conductance and inductance of the split rings cause strong paramagnetic or diamagnetic activity around the resonance frequency. The magnetic response can be extended to the terahertz range or even higher frequencies with large positive or negative permeability (see Figure 5.135).

By 2001, researchers at Imperial College and Marconi Caswell (London) announced a magnetic resonance medical imaging system using a magnetic metamaterial based on Pendry's design (Wiltshire

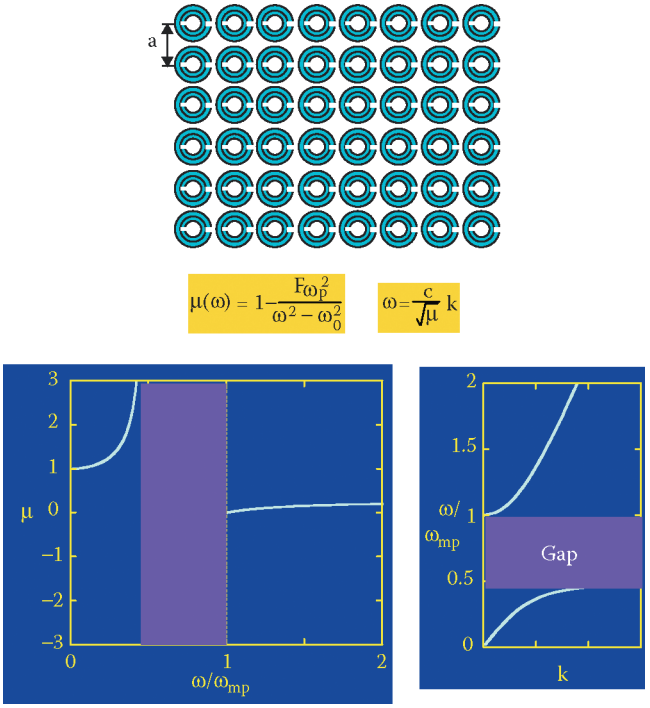


FIGURE 5.135 Pendry's split rings result in a negative μ.

et al.[41]). This group stated that "exploiting this class of materials could fundamentally change existing approaches to magnetic resonance imaging and spectroscopy."

Also in 2001, 3 years after Pendry's split ring proposal and close to 40 years after Veselago's first paper on negative refraction, physicist Richard Shelby's group at the University of California, San Diego demonstrated a left-handed composite metamaterial that exhibited a negative index of refraction for microwaves.[42] In Figure 5.136a we show two embodiments of these composite microwave metamaterial structures. The device pictured in the left photo shows sets of SRRs on common substrates alternating with separate rows of wires. The San Diego group reported that they had followed Pendry's prescriptions and succeeded in constructing a material with both a negative μ and a negative ε at microwave frequencies. The raw materials used, copper wires and copper rings, do not have unusual properties of their own and are nonmagnetic. But, when incoming microwaves fall on these alternating rows of rings and wires, a resonant reaction between the light and the whole of the ring-and-wire array sets up tiny induced currents, which contribute fields of their own. The net result is a set of fields moving to the left, even as EM energy is moving to the right. Therefore, this effective medium is a left-handed material. In Figure 5.136b a transmission measurement is shown for this metamaterial. The results show that there is a frequency region where both ε and μ are negative.

In the photo on the right in Figure 5.136a, we show a more integrated microwave metamaterial with wire and SRR on the same substrate. The latter arrangement consists of a planar pattern of copper SRRs and wires on a thin fiberglass circuit board. The SRRs and wires are arranged into a 2D structure with a repeated 5-mm lattice, with the wires located on the opposite side of the circuit board from the SRRs. This material exhibits negative refractive index at around 10.5 GHz. The San Diego team also demonstrated that microwaves were refracted to the opposite side (negative angles) of the normal as expected for a left-handed sample. From the measured angle, a refractive index of –2.7 ± 0.1 was derived.

FIGURE 5.136 (a) Two embodiments of composite metamaterial structures. The structure on the left shows an SRR on a circuit board substrate and separate wires, whereas the structure on right is more integrated with wire and SRR on the same substrate.[43,44] (b) Transmission spectrum for the metamaterial on the left in (a).[44]

These metamaterials might have applications in microwave communications: a medium that focuses waves when other materials would disperse them (and vice versa) ought to be useful in improving existing delay lines, antennas, and filters.

Development of metamaterials that refract light at ever-shorter wavelengths has been fast. A negative refractive index material for the optical range, specifically for wavelengths close to 1.5 μm (200-THz frequency), was accomplished by Shalaev et al. in 2005[45] with a metal dielectric composite consisting of a 2 × 2-mm array of nanorods on a glass substrate fabricated using electron beam lithography. A "fishnet" design developed by Costas Soukoulis' group at the US Department of Energy's Ames Laboratory and produced by researchers Stefan Linden and Martin Wegener of the University of Karlsruhe was made by etching an array of holes into layers of silver and magnesium fluoride on a glass substrate.[46] This constituted the first negative refractive index metamaterial with a refractive index of −0.6 at the far red end of the visible spectrum (wavelength 780 nm). This discovery, detailed in the January 5, 2007 issue of *Science* and the January 1 issue of *Optic Letters* and noted in the journal *Nature*, marked a significant step forward from existing metamaterials that operate in the microwave or far infrared—but still invisible—regions of the spectrum.[46]

Thus, negative refraction is clearly demonstrated in Pendry-type quasihomogeneous metamaterials with unit cell dimensions less than the wavelength of the incoming radiation, consisting of interleaved arrays of wires ($\varepsilon < 0$) and SRRs ($\mu < 0$). In 2000, Masaya Notomi at NTT Basic Research Laboratories in Japan presented strong theoretical evidence that negative refraction does not imply large losses and that negative refraction can also be achieved in photonic crystals, which are inhomogeneous periodic media with a lattice constant comparable to the wavelength of the incoming radiation (see Figure 5.131).[47] The theoretical structures that Notomi imagined could in reality be composed of insulating dielectric materials, which have negligible losses even at optical wavelengths. In 2003, Cubukcu et al.[36] experimentally demonstrated negative refraction by a photonic crystal (a square array of alumina rods in air).

The physical principles that allow negative refraction in a photonic crystal arise from the dispersion characteristics of wave propagation in a periodic medium and are very different from that of the Pendry metamaterials we just studied. The dispersion surface in a photonic crystal becomes rounded near the photonic bandgap (similar to the behavior of a Fermi surface in metals). Such a dispersion surface resembles that of homogeneous dielectric materials; thus, Snell's Law can be applied, and an effective refractive index can be defined. However, the propagation direction can be negative because $d\omega/dk$ can be negative in photonic crystals. Effective refractive index states in photonic crystals have their analogy in the effective-mass states in semiconductors as illustrated in Figure 5.137. For many applications such as imaging, one requires index matching between a negative refractive index material and its surroundings so as to incur negligible losses. These requirements are more easily met with photonic crystals than with composite negative index metamaterials. Furthermore, a microwave photonic crystals can be more easily scaled to three dimensions and to optical frequencies, which is more difficult with composite negative index metamaterials.

The Ideal Lens

The performance of conventional lenses is limited by diffraction: they need a wide aperture NA for good resolution, but even so their resolution is

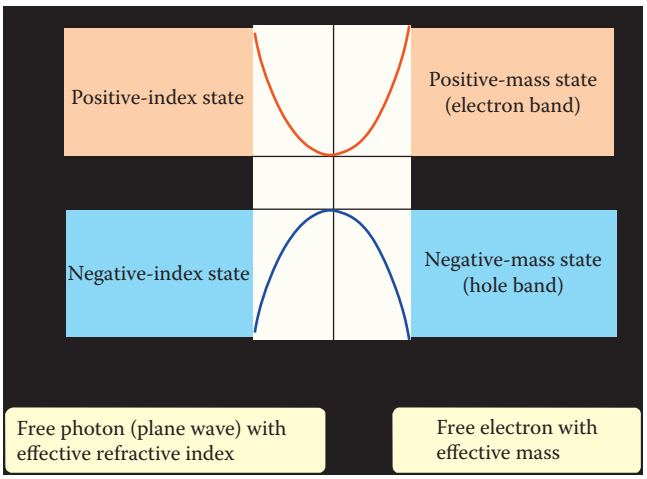


FIGURE 5.137 Effective refractive index states in photonic crystals have their analogy in the effective-mass states in semiconductors.

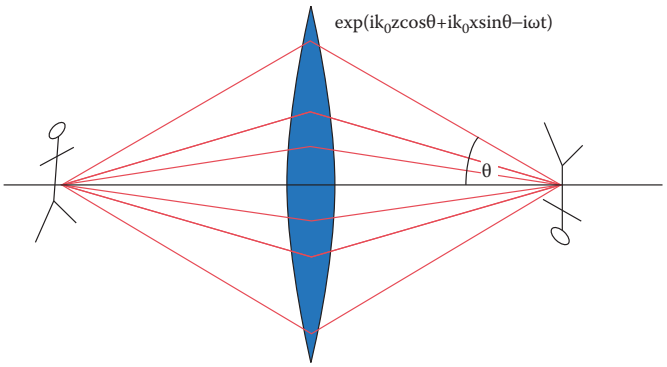


FIGURE 5.138 The resolution limit of a traditional lens.

limited by the wavelength used (Figure 5.138). The contribution to the resolution of an image from the far field is limited by the free space wavelength λ_0. Considering an object and a lens placed along the z-axis with rays from the object travelling in the $+z$-direction, we know that with $\theta = 90°$ (see Abbe's limit in Equation 5.1), we get a maximum value of the wave vector $\mathbf{k}_0 = \omega/c_0 = 2\pi/\lambda_0$—the shortest wavelength component of the image. Hence the resolution R cannot be better than:

$$R \approx \frac{2\pi}{\mathbf{k}_0} = \frac{2\pi c_0}{\omega} = \lambda_0 \qquad (5.268)$$

Contributions from the near field to the image come from large values of $\mathbf{k}_x$ (or $\mathbf{k}_y$) responsible for the finest details in the source. The light emanating from the object is made up of a superposition of plane waves, and the resulting $\mathbf{k}_z$ wave vector may be written as of function of $\mathbf{k}_x$ and $\mathbf{k}_y$ as follows:

$$\mathbf{k}_z = \sqrt{\frac{\omega^2}{c_0^2} - (\mathbf{k}_x^2 + \mathbf{k}_y^2)} \qquad (5.269)$$

In the far field, with energy going in the $+z$-direction, only the positive square root is taken, and all the components of the angular spectrum of the image for which $\mathbf{k}_z$ is real are transmitted through a conventional lens and refocused. However, if the second term under the square root in Equation 5.269 is larger than the first, $\mathbf{k}_z$ becomes imaginary, and the wave turns evanescent with an amplitude decaying as it propagates along the z-axis. What this really means is that the high-angular frequency components of the image are lost, resulting in an inability to resolve the smallest features of the object. The near-field, high-information content contributions to the image are not transmitted by the lens. For that, a lens with an NA of 1 would be required.

It is here that Pendry again had an amazing insight*: a superlens can be made from a flat slab of negative refractive index material that not only brings rays to a focus but also has the capacity to amplify the near field so that it can contribute to the image, thus removing the wavelength limitation. With the Pendry superlens with a refractive index $n = -1$ ($\varepsilon = -1$ and $\mu = -1$), transport of light in the $+z$-direction requires now that $\mathbf{k}_z$ in Equation 5.269 comes with the opposite sign (–). The evanescent wave now grows instead of decaying, and with the proper lens thickness all components of the angular spectrum are transmitted through such an ideal lens undistorted. A superlens prevents image degradation and beats the diffraction limit established by Abbe! However, the resonant nature of the amplification places severe demands on materials: they must be very low loss! The latter explains why there is so much interest in using photonic crystals rather than metamaterials for superlensing.

In Figure 5.139 we compare imaging with a traditional and a "Veselago" lens. RH and LH stand for right-handed and left-handed, respectively. A negative refractive index medium (LH) bends light to a negative angle relative to the surface normal. Light formerly diverging from a point source is set in reverse and converges back to a point. Released from the medium, the light reaches a focus for a second time. Note that the parallel component of the wave vector is always preserved in transmission but that the energy flow is opposite to the wave vector. The new Pendry lens based on negative refraction has unlimited resolution, provided that the condition $n = -1$ is met exactly. This can happen only at one frequency. Therefore, the secret of the new lens is that it can focus the near field, and to do this it must amplify the highly localized near field to reproduce the correct amplitude at the image. This

* "All this was pointed out by Veselago some time ago. The new message in this Letter is that, remarkably, the medium can also cancel the decay of evanescent waves. The challenge here is that such waves decay in amplitude, not in phase, as they propagate away from the object plane. Therefore to focus them we need to amplify them rather than to correct their phase. We shall show that evanescent waves emerge from the far side of the medium enhanced in amplitude by the transmission process." J.B. Pendry.[54]

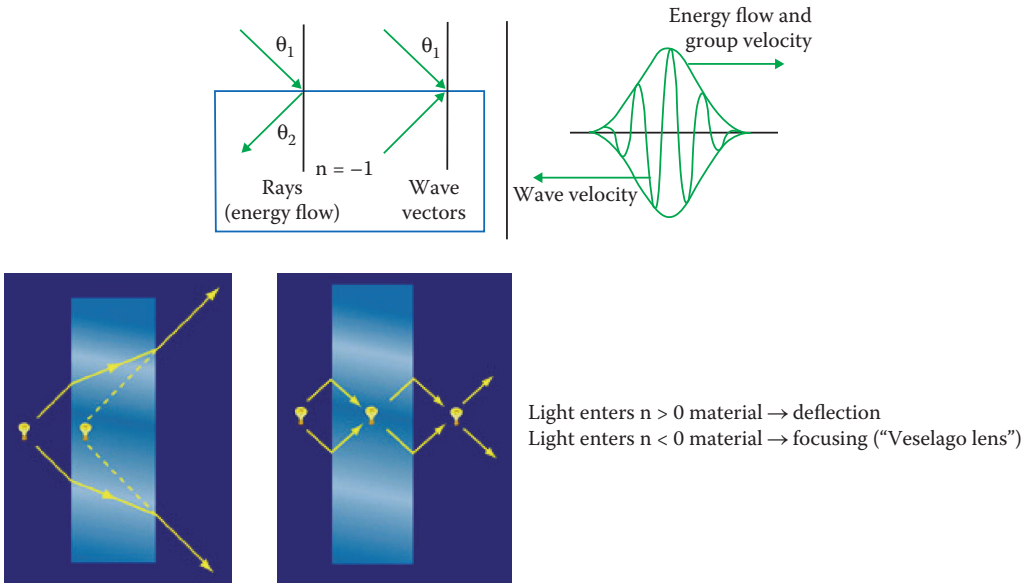


FIGURE 5.139 Comparing imaging with a traditional and a "Veselago" lens. RH and LH stand for right-handed and left-handed, respectively.

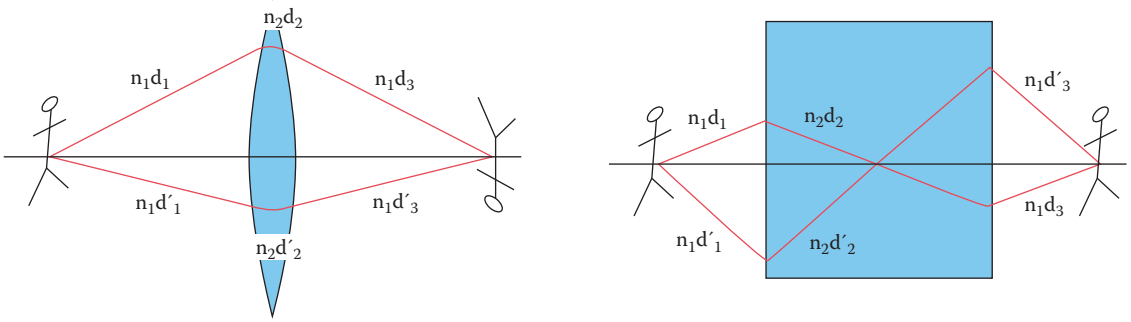


FIGURE 5.140 Fermat's principle applied to right-handed and left-handed lenses.

can be understood from Fermat's principle that light takes the shortest optical path between two points, as illustrated in Figure 5.140. For a traditional lens, the shortest optical distance between object and image is:

$$n_1d_1 + n_2d_2 + n_1d_3 = n_1d'_1 + n_2d'_2 + n_1d'_3 \quad (5.270)$$

Both paths converge at the same point because both correspond to a minimum. In the Pendry lens n_2 is negative, and the ray transverses negative optical space.

For a perfect lens, the shortest optical distance between object and image is zero:

$$n_1d_1 + n_2d_2 + n_1d_3 = 0 = n_1d'_1 + n_2d'_2 + n_1d'_3 \quad (5.271)$$

In other words, for a perfect lens the image is the object. The evanescent waves are "regrown" in a negative refractive index slab and fully recovered at the image plane as illustrated in Figure 5.141.

A poor man's near-field superlens ($\varepsilon < 0$ but $\mu = 1$) has already been demonstrated. In 2003, Zhang's group at the University of California, Berkeley

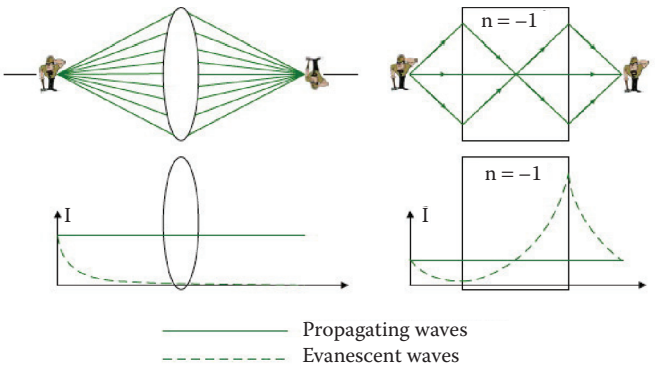


FIGURE 5.141 For a perfect lens, the image is the object. The evanescent waves are "regrown" in a negative refractive index slab and fully recovered at the image plane.

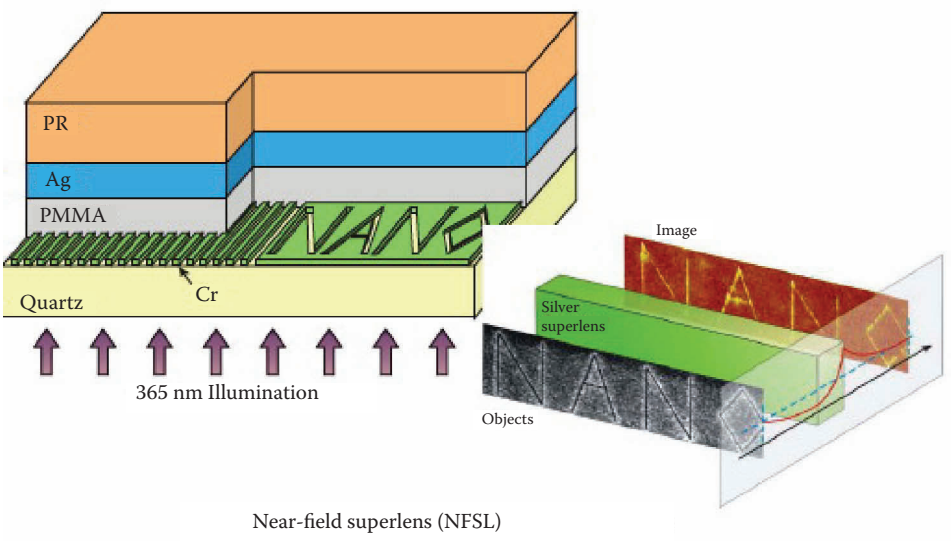


FIGURE 5.142 This poor man's superlens images 10-nm features with 365-nm light.[49]

showed that optical evanescent waves could be enhanced as they passed through a silver superlens.[48] The same group took this work one step further and imaged objects as small as 40 nm across with blue light at 365 nm with a superlens that was just 35-nm thick (see Figure 5.142).[49] In contrast, current optical microscopes can only resolve objects down to around 400 nm. With the Zhang's poor man's superlens at 365 nm, illumination features of a few tens of 10 nanometers were imaged using 365-nm light, clearly breaking Abbe's diffraction limit.

Lasers

Introduction: Masers and Lasers

LASER stands for Light Amplification by the Stimulated Emission of Radiation. The key word here is "stimulated." Most of the light emission we have mentioned so far is spontaneous; it happened as a result of randomly occurring "natural" effects. Stimulated emission refers to electron transitions that are "encouraged" by the presence of other photons. Albert Einstein put a mathematical analysis of stimulated emission forward as far back as 1916 that showed that an incident photon with $E \geq E_g$ (with E_g the semiconductor bandgap) was equally likely to cause stimulated emission of light as to be absorbed (Figure 5.143).

The emitted light has the same energy and phase as the incident light; in other words, it is coherent. Under normal circumstances, there are few excited electrons, and most are in the ground state, so we get predominantly absorption of light. If we could arrange for more excited than nonexcited electrons, then we would get mostly stimulated emission. Such a condition is called a *population inversion*. Because we get more photons out than we put in, this constitutes optical amplification, and the chain reaction of stimulated emissions is referred to as *amplification by stimulated emission of radiation*, or ASER for short. It took about 40 years after Einstein's initial theory was put forward until a technique was invented to sustain a chain reaction of stimulated emissions. The stimulated emissions system was first used in 1954 to amplify microwaves for communications by James Gordon, H.J. Zeiger, and Charles Townes in a so-called *Maser*, the microwave counterpart of a laser. It was not until 1960 when the first laser was successfully tested. The first laser used a ruby rod, giving off a beam of red laser light, the same kind one sees at a grocery checkout counter. Applications of laser technology went into full swing only in the early 1980s. Clearly, random spontaneous emission "wastes" electron transitions by giving an incoherent

FIGURE 5.143 An incident photon with $E \geq E_g$ is equally likely to cause stimulated emission of light to be absorbed.

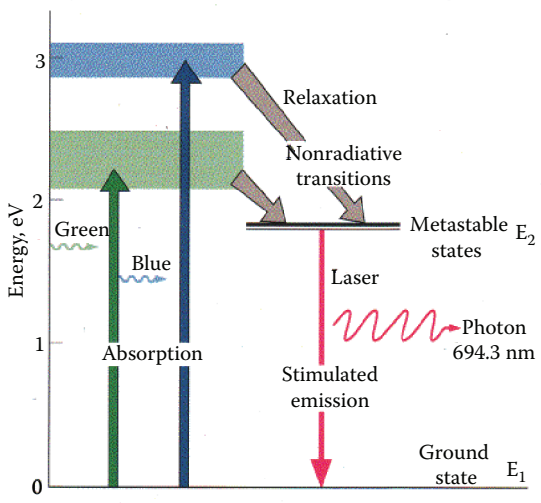


FIGURE 5.144 The energy levels of a typical laser material with metastable states at E_2.

output, so one must minimize those random transitions by using transitions for which the spontaneous emissions are of low probability, i.e., so-called *metastable states*. Therefore, the energy levels of a typical laser material look as we sketched in Figure 5.144, with E_2 representing the metastable states.

A ruby rod is a common laser material; it consists of Al_2O_3 (sapphire) with Cr^{3+} impurities. All one needs to make this rod into a laser is to 1) achieve population inversion, and 2) provide enough photons to stimulate emission. The first objective is met by filling (or populating) the metastable states with electrons generated by light from a xenon flash lamp (pumping = energy input), and the second condition is achieved by confining the photons such that they are forced to travel back and forth along the ruby rod of ruby by equipping the rod with mirrored ends. To keep the coherent emission, we must ensure that the light that completes a round trip between the mirrors returns in phase with itself. Hence the distance between the mirrors should obey $2L = N\lambda$, where N is an integer, λ is the laser wavelength, and L is the cavity length. Semiconductor lasers work in just the same way except that they achieve the population inversion electrically by using a carefully designed band structure. The ruby laser has an output at 694.3 nm. The basic lasing process is summarized in Figure 5.145.

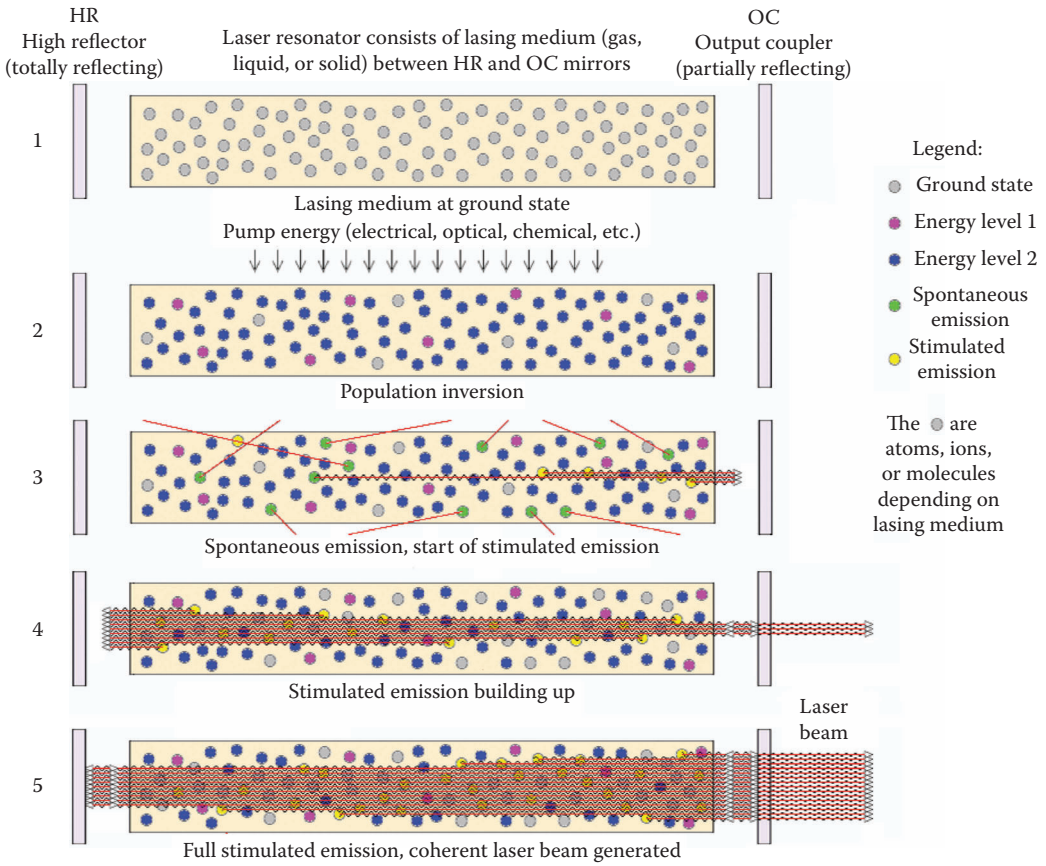


FIGURE 5.145 The basic lasing process.

From this figure, a laser comprises:

1. An active medium: the active medium may be made up by solid crystals such as ruby or neodymium-doped/yttrium aluminum garnet (Nd:YAG), liquid dyes, gases like CO_2 or helium/neon, or semiconductors such as GaAs. Active mediums contain atoms whose electrons may be excited to a metastable energy level by an energy source.
2. Excitation mechanism: excitation mechanisms pump energy into the active medium by one or more of three basic methods: optical, electrical, or chemical.
3. High-reflectance mirror: a mirror that reflects essentially 100% of the laser light.
4. Partially transmissive mirror: a mirror that reflects less than 100% of the laser light and transmits the remainder.

A laser diode is a laser where the active medium is a semiconductor, and it is similar to the light-emitting diode (LED) described in Chapter 4. Compared with other laser types, semiconductor lasers are compact, reliable, and last a very long time. The most common laser diode is formed from a p-n junction, and the excitation mechanism is the injection of an electrical current (Figure 5.146). When electron and holes find themselves present in the same region of the solid, they may recombine, resulting in spontaneous emission. This spontaneous emission is necessary to initiate the lasing action, but it is also one among several sources of inefficiency once the laser is working. It is also this spontaneous emission that makes a laser diode, in the regime below the lasing threshold, behave like an LED.

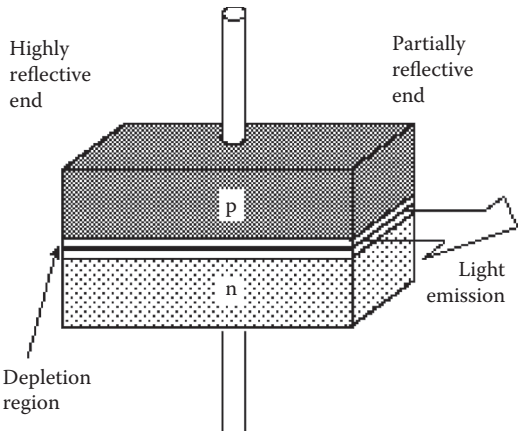


FIGURE 5.146 Simple laser diode.

In a conventional semiconductor junction diode, the energy released from the recombination of electrons and holes is carried away as phonons, i.e., lattice vibrations, rather than as the emerging photons. The difference between the photon-emitting semiconductor lasers and LEDs (on one hand) and conventional phonon-emitting (nonlight-emitting) semiconductor junction diodes (on the other hand) lies mostly in the type of semiconductor that is used in their construction. For lasers and LEDs we need a "direct bandgap" semiconductor such as gallium arsenide. Diode lasers then consist of these two basic components: an optical amplifier and a resonator. The resonator continuously recirculates light through the amplifier region and helps to focus it. The amplifier consists of the direct bandgap semiconductor junction. The two ends of the resonator structure need to be optically flat and parallel with one end mirrored and one partially reflective (see Figure 5.146). The length of the resonator is precisely related to the wavelength of the light to be emitted. The junction is forward biased, and the recombination process produces light as in an LED (incoherent). Above a certain current threshold (I_{th}) a photon moving parallel to the junction with energy equal to the recombination energy causes recombination by stimulated emission. This generates another photon of the same frequency, traveling in the same direction, with the same polarization and phase as the first photon. This means that stimulated emission causes gain in an optical wave (of the correct wavelength) in the injection region, and the gain increases as the number of electrons and holes injected across the junction increases. Finally, if there is more amplification than loss, the diode begins to "lase." Photons emitted into a mode of the waveguide will travel along the waveguide and be reflected several times from each end face before they are emitted. However, light is also lost because of absorption and by incomplete reflection from the end facets.

The spontaneous and stimulated emission processes are vastly more efficient in direct bandgap semiconductors than in indirect bandgap

semiconductors; thus, silicon is not a common material for laser diodes.

Some important properties of laser diodes are determined by the geometry of the optical cavity. Generally, in the vertical direction, the light is contained in a very thin layer, and the structure supports only a single optical mode in the direction perpendicular to the layers. In the lateral direction, if the waveguide is wide compared with the wavelength of light, then the waveguide can support multiple lateral optical modes, and the laser is known as *multimode*.

Vertical cavity surface-emitting lasers (VCSELs) have the optical cavity axis along the direction of the current flow rather than perpendicular to the current flow as in conventional laser diodes (Figures 5.146 and 5.147). The active region in this case is very short compared with the lateral dimensions, so that the radiation emerges from the surface of the cavity rather than from its edge. The reflectors at the ends of the cavity are dielectric mirrors made from alternating high- and low-refractive index quarter-wave thick multilayer. Such dielectric mirrors provide a high degree of wavelength-selective reflectance at the required free surface wavelength λ if the thicknesses of alternating layers d_1 and d_2 with refractive indices n_1 and n_2 are such that $n_1d_1 + n_2d_2 = \frac{1}{2}\lambda$, which then leads to the constructive interference of all partially reflected waves at the interfaces. However, there is a disadvantage because of the high mirror reflectivities; VCSELs have lower output powers when compared with edge-emitting lasers.

The laser gain g is the amplification factor by which the radiation intensity is multiplied after traversing a certain distance L of active laser medium. If the gain g is positive, the intensity increases exponentially as a function of distance:

$$I = I_0 e^{(g \times L)} \tag{5.272}$$

with g in units of inverse distance. In other words, when the absorption coefficient $\alpha > 0$, we are dealing with absorption, and with $\alpha < 0$, we have a gain g (so $g = -\alpha$ is used for negative absorption). The increase of intensity in a laser is much greater than can be obtained by spontaneous emission mechanisms. The threshold of a laser is the state where the gain just equals the cavity losses. For electrically pumped lasers this corresponds to a certain threshold current I_{th}.

Quantum Mechanics and Lasers

Quantum Confinement

Lasers work because of the quantum properties of photons—one photon tends to cause another to be emitted—and one photon cannot be distinguished from another (they are bosons; see Chapter 3). If there are many excited atoms, the photons can "cascade," and very intense, collimated light is emitted, forming a beam of precisely the same color of light. Because photons cannot be distinguished, which atom emitted a given photon is completely uncertain, but that also means that the direction and energy of the photons can be very certain.

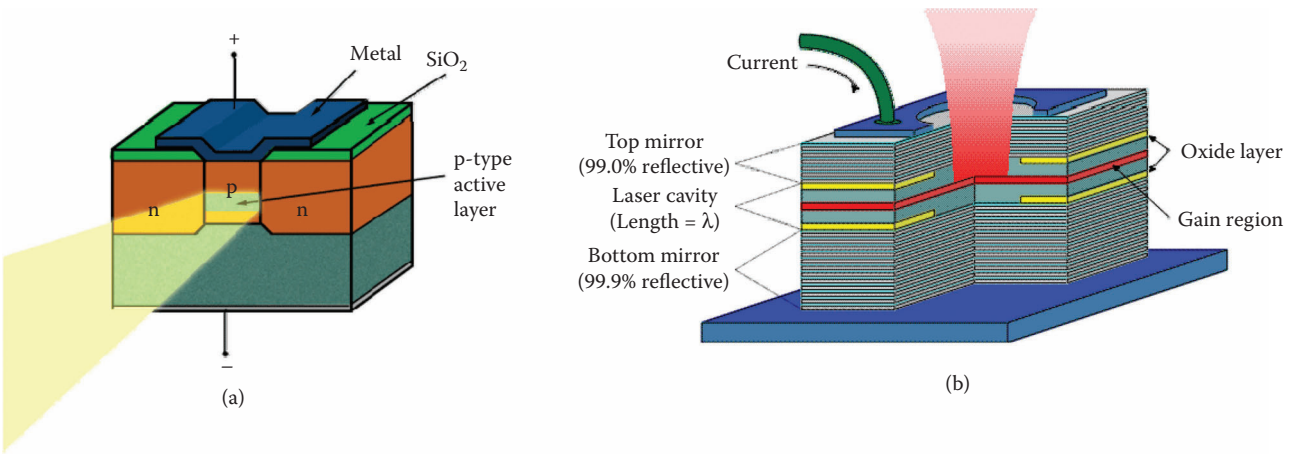


FIGURE 5.147 Comparison of an edge-emitting laser diode (a) and a vertical cavity surface-emitting laser (VCSEL) (b).

In a quantum well, we saw in Chapter 3, there are no allowed electron states at the very lowest energies (zero-point energy). What that means is that the lowest allowed conduction state in a quantum well is broad so that many electrons can be accommodated. Similarly, the top of the valence band has plenty of states available for holes. This means that it is possible for many holes and electrons to combine and produce photons with identical energy for enhanced probability of stimulated emission (lasing), making the quantum well laser much more efficient than a diode laser. The two important reasons that better lasers are enabled with lower-dimensionality semiconductors is the quantum confinement of carriers, which leads to a concentrated density of states, and a sharper energy distribution. The carrier distribution is given by n(E) = G(E)f(E) (see Equation 3.230), and bulk semiconductors have a broad carrier distribution compared with lower-dimensional semiconductor structures. Arakawa and Sakaki predicted significant increases in the gain and decreases in the threshold current I_{th} of semiconductor lasers using QDs or boxes in the active layer of the laser.[50] The gain g predicted by Asada et al.[51] for 3D, 2D, 1D, and 0D semiconductor structures is shown in Figure 5.148.

Quantum Cascade Lasers

Esaki was not only the originator of the Esaki tunnel diode, covered in Chapter 4, but also the originator

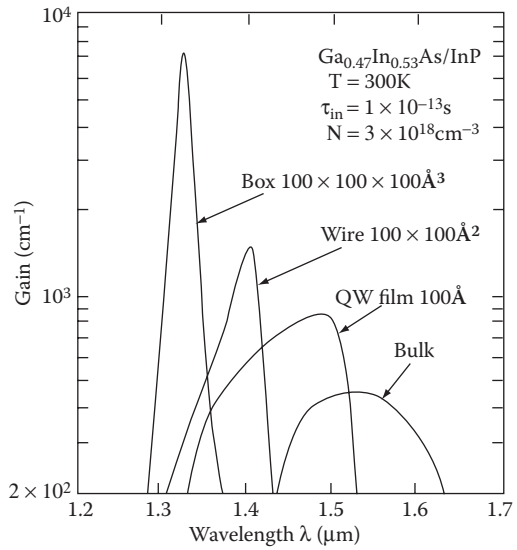


FIGURE 5.148 Gain for 3D, 2D, 1D, and 0D semiconductor structures proposed by Asada et al.[51]

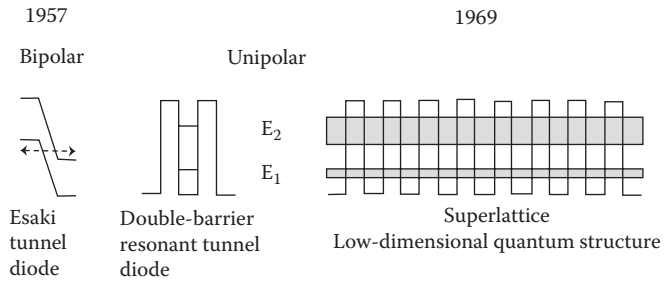


FIGURE 5.149 Esaki inventions from the 1957 Esaki tunnel diode to the double-barrier resonant diode to the superlattice. In bipolar devices, electrons and holes are involved; in unipolar devices, only one type of charged particle is involved.

of superlattices (1969) and double-barrier resonant tunneling diodes (1974). The resonant tunnel diode has even more spectacular characteristics than the Esaki tunnel diode and constitutes one of the possible contenders to replace CMOS technology. In Figure 5.149 we summarize the intellectual path followed by Esaki and colleagues in going from the first tunnel diode to the double-barrier resonant tunnel diode and finally to low-dimensional quantum structures.

Before arriving at the superlattice concept, Esaki and Tsu examined the feasibility of structural formation of potential barriers and wells that were thin enough to exhibit resonant tunneling (see Figure 5.149).

Resonant tunneling in quantum mechanics is an effect observed in quantum wells that are coupled to electron reservoirs by thin tunnel barriers and corresponds to a resonant enhancement of transmission that is seen whenever the energy in the reservoirs matches one of the quantized values in the well. With barriers that are thin enough, an electron with energy lower than that of the height of the barrier can tunnel through the barrier, as long as there is an empty state of the same energy level waiting on the other side. The well is then "in resonance." Resonant tunneling can be compared to the transmission of an EM wave through a Fabry-Pérot resonator.

The double-barrier resonant tunneling diode shown in Figure 5.150 is a variant of the tunnel diode of Chapter 4, with thin barrier layers of GaAlAs and a GaAs quantum well with sharp interfaces and layers with a width comparable with the Schrödinger wavelengths of the electrons

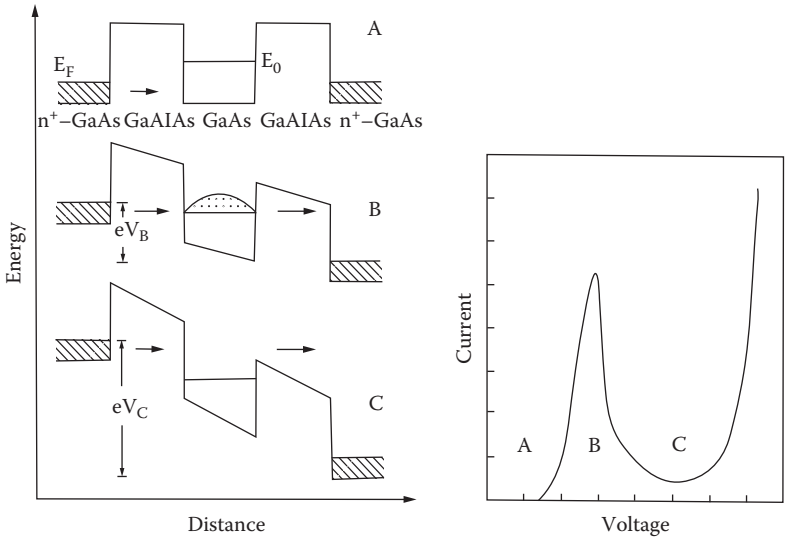


FIGURE 5.150 Semiconductor resonant tunneling diode: device schematic and current voltage plot.

permitting resonant behavior. In the device schematic we see that the contacts/leads are made of highly doped n-type GaAs (n^+ GaAs) and that two thin GaAlAs barrier layers (e.g., 20-Å thick) enclose a thin GaAs well region (e.g., 50-Å thick). The narrower bandgap material GaAs is enclosed by the considerably larger bandgap of GaAlAs to establish potential barriers at the surface of the GaAs well. Because the GaAs well is thin, the motion of electrons and holes is restricted in one dimension and is forced to occupy discrete states of energy instead of staying arbitrarily within an energy continuum. With a low bias applied, the quantized energy level in the well E_o is situated above E_F, and no current flows (A in the schematic and the current voltage plot). At a medium bias, the level in the well has lined up with the conduction band, and current flows (B in the schematic and the current voltage plot). At yet higher bias, the energy level in the well is now situated below the bottom of the conduction band, and the current decreases (C in the schematic and the current voltage plot). The current voltage plot is typical for a tunnel diode (see Figure 5.150), but whereas the Esaki diode is a bipolar device (two types of charge carriers are involved: electrons and holes) and features relatively large dimensions, the resonant tunnel diode is a unipolar device (only one type of charge carrier) with layers that have a thickness comparable with the Schrödinger wavelengths of the electrons.

A superlattice is a material with periodically alternating layers of different substances. Such structures possess periodicity both on the scale of each layer's crystal lattice and on the scale of the alternating layers. Quantum superlattices are a type of superlattice featuring arrays of periodic quantum structures: quantum wells, quantum dots, or quantum wires. Esaki and Tsu first proposed research on such artificially structured materials in 1960 to the Army Research Office (ARO). They envisioned engineered semiconductor superlattices based on a periodic structure of alternating layers of semiconductor materials with wide and narrow bandgaps using high-precision heteroepitaxial deposition methods such as molecular beam epitaxy (MBE) and metalorganic chemical vapor deposition (MOCVD) and involving III–V semiconductors. Although the proposal was favorably received by ARO, *Physical Review* rejected their first paper on the basis that it was "too speculative" and involved "no new physics." In some circles, the ARO proposal was criticized as impossible to implement in practice. The main objection being that a human-made structure with compositional variations on the order of a few nanometers would be thermodynamically unstable because of interdiffusional effects. Fortunately, it turned out that interdiffusion is negligible at the temperatures involved, and innovations and improvements in epitaxial techniques such as MBE and MOCVD made it possible to prepare high-quality heterostructures

with predesigned potential profiles and impurity distributions with dimensional control close to the interatomic spacing. This great precision opened up the mesoscopic quantum regime to practical experimentation. Because a 1D potential could now be introduced along with the growth direction in the laboratory, famous examples in the history of 1D mathematical physics, including the above-mentioned resonant tunneling (1930s), Krönig-Penney bands (1931), Zener band-to-band tunneling (1934), and so on, all of which had remained textbook exercises, could for the first time be practiced in the laboratory.

The superlattice example shown in Figure 5.151A is formed by a periodic array of quantum wells produced by the growth of alternate layers of a wide-bandgap (AlGaAs) and narrower bandgap (GaAs) semiconductors in the growth direction. At 300 K, the bandgap of GaAs is 1.43 eV, whereas it is 1.79 eV for $Al_xGa_{1-x}As$ (x = 0.3). Thus, the electrons and holes in GaAs are confined in a 1D potential well of length L in the *z*-direction. Two or more quantum wells side by side give rise to a multiple quantum well (MQW) structure. This is called a *compositional superlattice*. An artificial periodic potential in a semiconductor crystal can also be fabricated through periodic n-type and p-type doped layers as shown in Figure 5.151B. Such a superlattice is called a *doping superlattice*. When the individual quantum wells are widely separated, the wave functions of the electrons and holes remain confined within the individual wells, and the array can be treated as a set of isolated quantum wells. Noninteracting MQWs like this are used to enhance/amplify absorption or emission of photons compared with what is possible with a single quantum well. In lasing, for example,

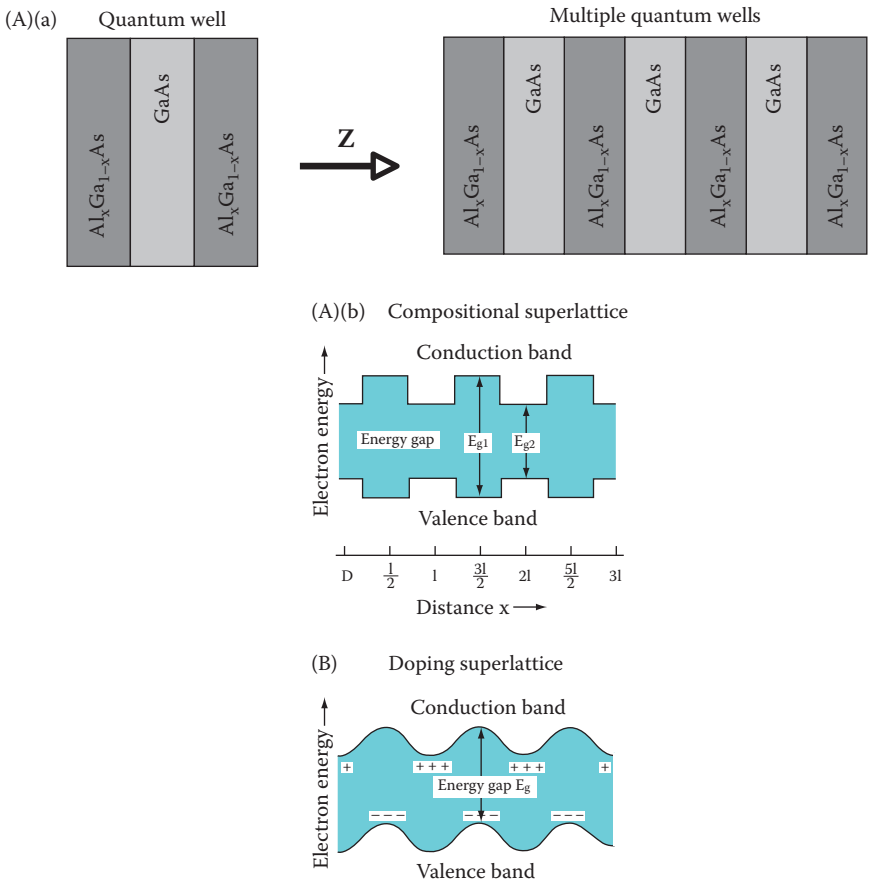


FIGURE 5.151 (A) Superlattice formed by a periodic array of quantum wells produced by the growth of alternate layers of wider bandgap (AlGaAs) and narrower bandgap (GaAs) semiconductors in the growth direction. Schematics of the arrangement (a) and the energy bands of multiple quantum wells (b). This is a compositional superlattice. (B) Artificial periodic potential in a semiconductor crystal made through periodic n-type and p-type doped layers. Such a superlattice is called a *doping superlattice*.

the stimulated emission is amplified by photons transversing through MQWs, with each well acting as an independent lasing medium.

One of the most interesting applications of superlattices is the quantum cascade laser. In the April 1994 issue of *Science*, Faist, Capasso, Sivco, Sirtori, Hutchnson, and Cho of AT&T Bell Labs reported a new type of far-infrared laser called the *quantum cascade laser*.[52] The idea of this superlattice device was first proposed by Rudolf Kazarinov (who is now with AT&T Bell Labs) and Robert Suris of A. F. Ioffe Institute in 1971 (http://www.mtmi.vu.lt/legacy/pfk/funkc_dariniai/nanostructures/superlattice.htm). The Bell Labs' work thus represents a real breakthrough that had been anticipated for almost a quarter of a century. Unlike quantum-confined lasers, which involve interband transitions between the conduction band and valence band, these quantum cascade lasers involve transitions of electrons between the various sub-bands (minibands) corresponding to different quantized levels of the conduction band. Because only electron transitions are involved, one refers to these types of lasers as unipolar lasers. There could be ~50 quantum wells in an MQW laser geometry. The barrier layers are very thin (1–3 nm), and an excited electron may emit 25–75 photons as it cascades down the ladder of subbands in the conduction band. Quantum cascade lasers have been demonstrated for wavelengths between 3 and 20 μm, which is useful for sensing atmospheric pollution. Quantum cascade lasers, operating on intraband transitions and emitting in the mid-infrared region, are used, e.g., for trace gas analysis.

In what follows we clarify how these bands within the conduction band come about. If the thickness of the wide-bandgap barrier layer in Figure 5.151 is small enough so that electrons may tunnel through, then the situation becomes similar to what happens when individual atoms are brought together in a crystal. To understand the interaction between these very closely spaced quantum wells better, we follow the thinking of Esaki and Tsu. Let us bring a set of N quantum wells together as shown in Figure 5.152. These wells consist again of alternate layers of GaAs (well) and AlGaAs (barriers). First, consider two wells in this array only, each with their own set of quantized levels E_n and labeled with their own quantum numbers n = 1, 2, 3, along the confinement direction (the growth direction). Reducing the separation between these two wells allows the wave functions to overlap, and the same (degenerate) energy levels E_n of the two wells now form two new states: $E_n(1)$ and $E_n(2)$ (just as we saw in Chapter 3 when bringing atoms together), where $E_n(1) = E_n - \Delta_n$ and $E_n(2) = E_n + \Delta_n$. The magnitude of the splitting $2\Delta_n$ depends on the level E_n. The higher the energy E_n, the more the wave function extends into the energy barrier between the wells, allowing for stronger interaction between the wave functions in those wells and leading to a larger splitting Δ_n. Bringing N wells together, instead of only two, increases the energy degeneracy for all N wells, producing N levels closely spaced in a so-called *miniband*. In the

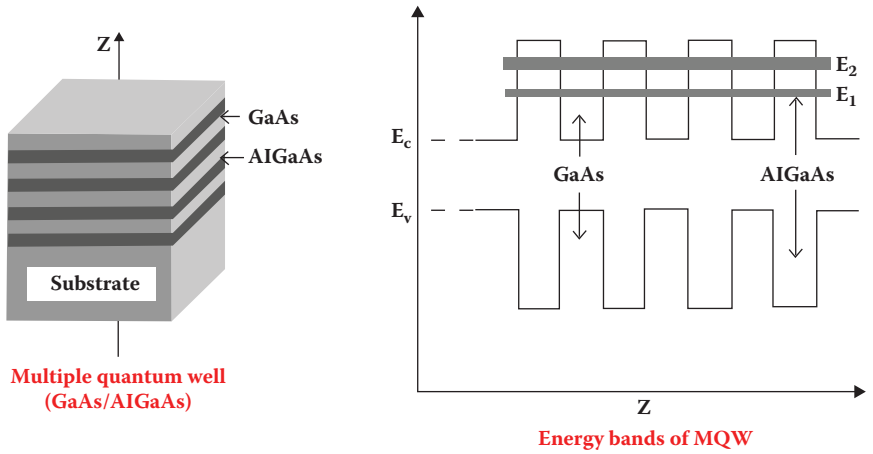


FIGURE 5.152 Schematic of formation of minibands in a superlattice consisting of alternate layers of GaAs (well) and AlGaAs (barriers).

case illustrated in Figure 5.152, individual levels E_1 and E_2 in the quantum wells are split into two minibands; the higher-energy miniband has a greater bandwidth than the lower-energy miniband for the reason explained above (larger splitting for higher-energy levels). Just as in a crystal where the periodic atomic potential leads to band formation in a superlattice, an artificial, human-made periodical potential causes the formation of minibands. Thus, the introduction of the 1D superlattice potential perturbs the band structure of the host materials, yielding a series of narrow sub-bands and forbidden gaps that arise from the subdivision of the Brillouin zone into a series of minizones. Most semiconductor devices are operated based on the interband transitions, namely, between the conduction and valence bands. These devices are usually bipolar involving a p-n junction. A new class of devices is based on the transitions between the sub-bands or minibands in the conduction or valence bands as described here. These intraband devices are unipolar and are faster than the intraband devices. Today a wide variety of engineered structures have been developed that exhibit extraordinary electronic transport and optical properties that do not exist in any natural crystals. This new degree of freedom offered in semiconductor research through advanced materials engineering has inspired many ingenious experiments, resulting in observations of not only predicted effects but also totally unknown results, and it appears that nearly half of all semiconductor physicists in the world are working now in this area. In addition to being used in quantum cascade layers, superlattice structures have been used in field effect transistors, where several quantum wells provide parallel conducting channels, increasing the device current-carrying capabilities and hence the output power. Superlattices are also used for photodetectors and for novel LEDs.

Photonic Crystal and Lasers

Earlier in this chapter we listed as an important benefit of photonic crystals the ability to control spontaneous emission and to modify the state density and group velocity of light by creating structure defects. We pointed out that these structures could be used to create novel, highly efficient lasers that emit light in a very narrow wavelength range. Point defects in photonic crystals act as high-Q resonator cavities and have been used in low-threshold photonic crystal lasers. In photonic lasers, only the wavelength that matches the wavelength of the defect mode is amplified because it can propagate freely through the material. The intensity of the propagating light increases as it undergoes successive reflections and travels back and forth through the photonic crystal. Meanwhile, light at other wavelengths is trapped within the photonic crystal and cannot build up. This means that the laser light is emitted in a narrow wavelength range that is directly related to the geometry of the microcavity.

Questions

Questions by Chuan Zhang, UC Irvine

5.1: In a traditional lithography system using i-line (365 nm) wavelength, the numerical aperture is NA = 0.8. Can we use this system to fabricate a pattern with a minimum line width of 65 nm? List several ways of increasing the resolution of the lithography system.

5.2: A beam of light with a wavelength of 500 nm goes through an aperture that has a diameter of 0.1 mm and generates an Airy disk on the image plane very far from the aperture. If the diameter of the Airy disk is 50 mm, what is the distance between the aperture and the image plane? If the distance becomes 3 times longer, then what would the size of the Airy disk be?

5.3: A beam of monochromatic red light with a frequency of 500 terahertz and a wavelength of 600 nm goes from vacuum into a piece of glass with an incidence angle of 30°. After getting into the glass, the wavelength of the light becomes 450 nm. What is the angle of refraction?

5.4: A beam of monochromatic light with a frequency of 500 terahertz and a wavelength of 600 nm transmits in water, which has a refractive index of 1.333, and then the light is totally internally reflected at the interface

of water and an unknown material with an incident angle 60°. What is the minimum possible wavelength of that light when it is transmitted in the unknown media?

5.5: Make a comparison between right-handed and left-handed materials. List the two most important applications for left-handed materials.

5.6: A stationary proton with an electric charge of $1.60217653 \cdot 10^{-19}$ C and a mass of $1.6726 \cdot 10^{-27}$ kg is accelerated to 0.8c under an electric field intensity of $E_0 = 1 \cdot 10^6$ V/m; what is the weight of the particle at this speed and how long did it take to reach the final speed?

5.7: In an optical microscope, if we replace the light source (λ = 550 nm) of visible light with ultraviolet (λ = 275 nm), what is the percentage increase of resolution? If the numeral aperture in air is 0.8, what is the smallest distance between two parallel lines that can be resolved with the ultraviolet light source? And what if we immerse the lens system into an oil medium with a refractive factor of 1.5?

5.8: A beam of light has a wavelength of 500 nm in air. After crossing through another medium, its wavelength becomes 200 nm. Assuming the difference of permeability between this medium and air is very small, what is the relative dielectric constant for this medium?

5.9: The refractive index of a crystal is 1.7, $dn/d\lambda = -0.15\ um^{-1}$; what is the ratio of phase velocity to group velocity for light with a wavelength of 400 nm?

5.10: What is the optical phenomenon called *light polarization*? A beam of light becomes completely polarized after being reflected from the interface between water and an unknown medium. We know the incident angle is 60°. What is the refractive index of the unknown medium?

5.11: How does an optical fiber work? Explain why fiber losses increase if the fiber is bent? How can plasmonics avoid the bending problem?

5.12: How can surface plasmons be excited? What are some applications?

5.13: What is the difference between a direct bandgap and indirect bandgap semiconductor? Which one has a higher photon absorption and electron transition possibility? For optoelectronic applications, which kind of semiconductor is preferred?

5.14: Compare the similarities and differences between photonic crystals and metamaterials.

5.15: Why is there a resolution limit for traditional lenses? How does a superlens made from negative refractive index material exceed that limit?

5.16: What is Raman scattering? Compare Raman spectroscopy with infrared (IR) spectroscopy.

5.17: As the size of a metal particle becomes smaller and smaller, will there be any changes in its color? Explain why.

5.18: Why is the sky blue, while clouds are white?

5.19: Show that in the Maxwell equation $\Delta \times H = J_c + \frac{\partial D}{\partial t}$ all three terms have the same dimensions.

5.20: If an eye has a pupil opening (lens diameter) of 4 mm, can it resolve two point sources of 600 nm light (direct or reflected) that are 5 mm apart and 10 m away? Does the eye's resolving ability improve if the pupil enlarges to 5 mm?

References

1. Cheng, D. K. 1989. *Field and wave electromagnetics.* 2nd ed. Boston: Addison-Wesley.
2. Taya, M. 2005. *Electronic composites.* New York: Cambridge University Press.
3. Veselago, V. G. 1968. The electrodynamics of substances with simultaneously negative values of ε and μ. *Sov Phys Usp* 10:509–14.
4. Ashkin, A. 1970. Acceleration and trapping of particles by radiation pressure. *Phys Rev Lett* 24:156–59.
5. Berns, M. W., and D. E. Rounds. 1970. Cell surgery by laser. *Scientific American* 222: 98–103.
6. Atwater, H. A. 2007. The promise of plasmonics. *Scientific American* 296:56–63.
7. Greene, B. 1999. *The elegant universe.* New York: W.W. Norton.
8. Parker, B. 1987. *Search for a super-theory: from atoms to superstrings.* New York: Plenum Press.
9. Faraday, M. 1858. The Bakerian lecture: experimental relations of gold (and other metals) to light. *Philos Trans Royal Soc London* 147:145–81.
10. Padovani, S., C. Sada, P. Mazzoldi, B. Brunetti, I. Borgia, A. Sgamellotti, and A. Giulivi. Copper in glazes of renaissance luster pottery: nanoparticles, ions, and local environment. *J Appl Phys* 93:10058–63.

11. Bohren, C. F., and D. R. Huffman. 1983. *Absorption and scattering of light by small particles.* New York: John Wiley and Sons.
12. Gans, R. 1915. Fortpflanzung des Lichts durch ein inhomogenes Medium. *Ann Physik* 352:709–36.
13. Link, S., and M. El-Sayed. 1999. Size and temperature dependence of the plasmon absorption of colloidal gold nanoparticles. *J Phys Chem B* 103:4212–17.
14. Westcott, S. L., S. J. Oldenburg, T. R. Lee, and N. J. Halas. 1998. Formation and adsorption of clusters of gold nanoparticles onto functionalized silica nanoparticle surfaces. *Langmuir* 14:5396–401.
15. Chen, J., B. Wiley, Z.-Y. Li, D. Campbell, F. Saeki, H. Cang, L. Au, J. Lee, X. Li, and Y. Xia. 2005. Gold nanocages: engineering their structure for biomedical applications. *Adv Mater* 17:2255–61.
16. Maier, S. A., M. L. Brongersma, and H. A. Atwater. 2001. Electromagnetic energy transport along arrays of closely spaced metal rods as an analogue to plasmonic devices. *Appl Phys Lett* 78:16–18.
17. Frenkel, J. 1931. On the transformation of light into heat in solids. I. *Phys Rev Lett* 37:17–44.
18. Frenkel, J. 1931. On the transformation of light into heat in solids. II. *Phys Rev* 37:1276–94.
19. Gross, E. F. 1956. Optical spectrum of excitons in the crystal lattice. *Il Nuovo Cimento* 3:672–701.
20. Gustav, M. 1908. Beiträge zur Optik Trüber Medien, Speziell Kolloidaler Metallösungen. *Ann Physik* 330:377–445.
21. Jeanmaire, D. L., and R. P. V. Duyne. 1977. Surface Raman spectroelectrochemistry. Part I. Heterocyclic, aromatic, and aliphatic amines adsorbed. J Elecroanal Chem 84:1–20.
22. Shalaev, V. M. 1999. *Nonlinear optics of random media: fractal composites and metal-dielectric films*. Berlin: Springer.
23. Laserna, J. J. 1993. Combining fingerprinting capability with trace analytical detection: surface-enhanced raman spectometry. *Anal Chim Acta* 283:607–22.
24. Kneipp, K., H. Kneipp, V. B. Kartha, R. Manoharan, G. Deinum, I. Itzkan, R. R. Dasari, and M. S. Feld. 1998. Detection and identification of a single DNA base molecule using surface-enhanced raman scattering (SERS). *Phys Rev* E57:R6281–84.
25. John, S. 1987. Strong localization of photons in certain disordered dielectric superlattices. *Phys Rev Lett* 58:2486–89.
26. Yablonovitch, E. 1987. Inhibited spontaneous emission in solid-state physics and electronics. *Phys Rev Lett* 58:2059–64.
27. Ho, K. M., C. T. Chan, and C. M. Soukoulis. 1990. Existence of a photonic gap in periodic dielectric structures. *Phys Rev Lett* 65:3152–55.
28. Yablonovitch, E., and T. J. Gmitter. 1989. Photonic band structure: the face centered cubic case. *Phys Rev Lett* 63:1950.
29. Ozbay, E. 1994. Micromachined millimeter-wave photonic band-gap crystals. *Appl Phys Lett* 64:2059–61.
30. Lehmann, V., and H. Foll. 1990. Formation mechanism and properties of electrochemically etched trenches in n-type silicon. *J Electrochem Soc* 137:653–60.
31. Grüning, U., V. Lehmann, S. Ottow, and K. Busch. 1996. Macroporous silicon with a complete two-dimensional photonic band gap centered at 4.9 μm. *Appl Phys Lett* 68: 747–49.
32. Campbell, M., D. N. Sharp, M. T. Harrison, R. G. Denning, and A. J. Turberfield. 2000. Fabrication of photonic crystals for the visible spectrum by holographic lithography. *Nature* 404:53–56.
33. Vlasov, Y. A., X. Z. Bo, J. C. Sturm, and D. J. Norris. 2001. On-chip natural assembly of silicon photonic bandgap crystals. *Nature* 414:289–93.
34. Joannopoulos, J. D. 2008. *Photonic crystals: molding the flow of light*. Princeton, NJ: Princeton University Press.
35. Coen, S., A. H. Chau, R. Leonhardt, J. D. Harvey, J. C. Knight, W. J. Wadsworth, and P. S. Russell. 2001. White-light supercontinuum generation with 60-ps pump pulses in a photonic crystal fiber. *Opt Lett* 26:1356–58.
36. Cubukcu, E., K. Aydin, E. Ozbay, S. Foteinopoulou, and C. M. Soukoulis. 2003. Electromagnetic waves: negative refraction by photonic crystals. *Nature* 423:604–05.
37. Research Highlights. 2007. Invisibility cloak in sight. *Nature* 444:4–5.
38. Pendry, J. B., A. J. Holden, D. J. Robbins, and W. J. Stewart. 1998. Low frequency plasmons in thin-wire structures. *J Phys Condens Matter* 10:4785–809.
39. Pendry, J. B., A. J. Holden, W. J. Stewart, and I. Youngs. 1996. Extremely low frequency plasmons in metallic mesostructures. *Phys Rev Lett* 76:4773–76.
40. Pendry, J. B., A. J. Holden, D. J. Robbins, and W. J. Stewart. 1999. Magnetism from conductors and enhanced nonlinear phenomena. *IEEE Trans Microwave Theory Techn* 47:2075–84.
41. Wiltshire, M. C. K., J. B. Pendry, I. R. Young, D. J. Larkman, D. J. Gilderdale, J. V. Hajnal. 2001. Microstructured magnetic materials for RF flux guides in magnetic resonance imaging. *Science* 291:849–51.
42. Shelby, R. A., D. R. Smith. and S. Schultz. 2001. Experimental verification of a negative index of refraction. *Science* 292:77–79.
43. Smith, D.R., and N. Kroll. 2000. Negative refractive index in left-handed materials. *Phys Rev Lett* 85:2933.
44. Smith, D. R., W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser, and S. Schultz. 2000. Composite medium with simultaneously negative permeability and permittivity. *Phys Rev Lett* 84:4184–87.
45. Shalaev, V. M., W. Cai, U. K. Chettiar, H. K. Yuan, A. K. Sarychev, V. P. Drachev, and A. V. Kildishev. 2005. Negative index of refraction in optical metamaterials. *Opt Lett* 30:3356.
46. Soukoulis, C. M., S. Linden, and M. Wegener. 2007. Negative refractive index at optical wavelengths. *Science* 315:47–49.
47. Notomi, M. 2000. Theory of light propagation in strongly modulated photonic crystals: refraction like behavior in the vicinity of the photonic band gap. *Phys Rev B* 62:10696–705.
48. Fang, N., and X. Zhang. 2003. Imaging properties of a metamaterial superlens. *Appl Phys Lett* 82:161.
49. Fang, N., H. Lee, C. Sun, and X. Zhang. 2005. Sub-diffraction-limited optical imaging with a silver superlens. *Science* 308:524–37.
50. Arakawa, Y., and H. Sakaki. 1982. Multidimensional quantum well laser and temperature dependence of its threshold current. *Appl Phys Lett* 40:939–41.
51. Asada, M., Y. Miyamoto, and Y. Suematsu. 1986. Gain and the threshold of three-dimensional quantum-box lasers. *IEEE J Quant Electron* QE-22:1915–21.
52. Faist, J., F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho. 1994. Quantum cascade laser. *Science* 264:553–56.
53. Meinhart, C. D., and S. T. Wereley. 2003. The theory of diffraction-limited resolution in microparticle image velocimetry. *Meas Sci Technol* 14:1047.
54. Pendry, J. B. 2000. Negative refraction makes a perfect lens. *Phys Rev Lett* 85:3966.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

6

Fluidics

Outline

Introduction

Macroscale Laws for Fluid Flow

Breakdown of Continuum Theory in Fluidics

Forces at Interfaces

Mixing, Stirring, and Diffusion in Low Reynolds Number Fluids

Chemical Reactions in Microchambers—Microreactors

Fluid Propulsion

Electrowetting

Centrifugal Fluidic Platform—CD Fluidics

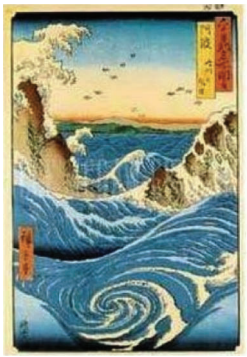
Scaling in Analytical Separation Equipment

Acknowledgments

Questions

Further Reading

References



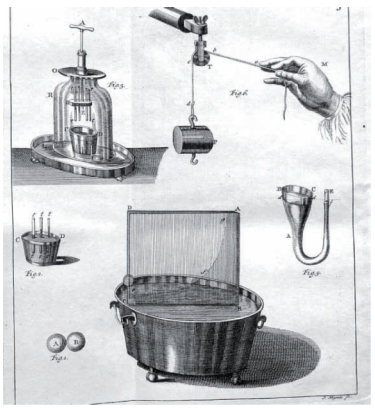
Angry Sea at Naruto
Ando Hiroshige,
1830

Observe the motion of the surface of the water, which resembles that of hair, which has two motions, of which one is caused by the weight of the hair, the other by the direction of the curls; thus the water has eddying motions, one part of which is due to the principal current, the other to random and reverse motion.

Leonardo da Vinci, 1510

The current nano hype makes us forget that already Sir Isaac Newton described nanoscale phenomena, the so-called Newton rings (left; no, that is not spaghetti he is looking at). Newton also very clearly formulated experiments of microfluidics by describing experiments with capillary flow between parallel glass plates (below). So, neither nano nor micro are [sic] really very novel it seems!

Based on an editorial by Andreas Manz and Harpal Minhas for *Lab on a Chip*



These [natural] laws may have originally been decreed by God, but it appears that he has since left the universe to evolve according to them and does not now intervene in it.

Stephen Hawking
A Brief History of Time

Introduction

Fluid mechanics (fluidics) is the study of motion of fluids (liquids and gases) and the forces associated with that motion. In Chapter 6 we introduce fluidics, compare various fluidic propulsion mechanisms, and discuss the influence of miniaturization on fluid behavior. Our treatment deals with macroscale, microscale, and nanoscale fluidic phenomena. Given the current academic and industrial interest, fluidics in miniaturized analytical equipment is detailed in this chapter as well. Example applications we discuss include electro-osmotic pumping systems, diffusive separation systems, micromixers, DNA amplifiers (PCR), compact disc diagnostic platforms, and cytometers.

Newton and da Vinci aside (see insets above), fundamental work on the physical behavior of fluids in micro- and nanomachined structures only started in the early 1980s. In the micro- and nanodomain, many factors seem to upset the old fluidics continuum order: Navier-Stokes equations (NSEs), with no-slip boundary conditions at the fluid interface, cannot always be applied, and slip flow, creep flow, intermolecular forces, etc., start suggesting molecular-based flow models instead. For gases, breakdown of macroscale physics is readily recognized, and some molecular-based modeling progress has been made. For liquids, with molecules much more closely packed, modeling has proven to be much more elusive.

Microfluidics and nanofluidics present researchers with exciting future applications. These include the characterization of individual molecules (e.g., reading the base sequence in DNA by feeding the molecule through a nanopore), studying the kinetics of very fast reactions, isolating a molecule in a nanometric space for detection purposes, further increasing the role of surfaces in catalytic reactions, manipulation of single molecules, labeling and studying of conformational changes of single molecules, applying intense electrical or magnetic fields to control chemical reactions, constructing hybrid systems by assembling biomolecules in NEMS structures, etc.

Macroscale Laws for Fluid Flow

Definition of Fluids and Types of Flow

A fluid, in our common experience, is a substance that has volume but no shape; that is, a fluid is a material that cannot resist a shear force or shear stress without moving.[1] The two most important parameters characterizing a fluid are density and viscosity. Density or specific mass, ρ, of a fluid is the mass per unit volume (in kgm^{-3}), and dynamic viscosity, η, is the fluid property that causes shear stresses when the fluid is moving; without viscosity there would be no fluid resistance. Dynamic viscosity is measured as kg/ms (SI system), that is, Pa·s or Poiseuille (Pl)*, and is given as the ratio of shear stress, τ_s, to shear rate, Γ_s:

$$\eta = \frac{\tau_s}{\Gamma_s} \tag{6.1}$$

The dynamic viscosity of air is 1.85×10^{-5} kg/ms, and for water it is 10^{-3} kg/ms. Equation 6.1 can be better understood from an inspection of Figure 6.1, in which we illustrate two large plates, of area *A*, separated by a fluid layer of uniform thickness, *d*. When one of the two plates is moved in a straight line relative to the other at an average velocity *V*, the shear force required for obtaining *V* is given by τ_{zx} (=F/A = force/area in N m^{-2}, i.e., Pa):

$$\frac{F}{A} = \tau_{zx} = \eta \frac{V}{d} \tag{6.2}$$

and V/d corresponds to the shear rate Γ_s (velocity gradient in s^{-1}). The first subscript in τ_{zx} is in the direction normal to the shearing force (*z*), and the second

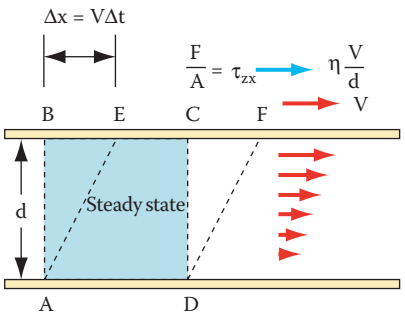


FIGURE 6.1 Schematic of shearing experiment.

* In the CGS system it is measured in poise (1 poise or P = 10^{-1} N·sm^{-2}) and 1 Pl = 10 P.

subscript is in the direction of the force (x). Equation 6.2 can be generalized as:

$$\tau_{zx} = \eta \frac{dV}{dz} \tag{6.3}$$

This relationship is known as Newton's Law of viscosity, and it relates the velocity gradient in the fluid to the shear stress. Thus, viscosity refers to friction between layers. The normal component of stress (F_n) for a fluid at rest is called the pressure, and the tangential component (F_t) is the shear stress (Figure 6.2). A solid can resist applied shear by deforming, and the stress is proportional to the strain, but a fluid deforms continuously under applied shear, and the stress is proportional to strain rate.

The kinematic viscosity, ν in m^2/s, or also myria (=10,000) Stokes (ma St, where the Stokes is the CGS unit, 1 $m^2/s = 10^4$ St), is the ratio of the absolute viscosity to density given by:

$$\nu = \frac{\eta}{\rho} \tag{6.4}$$

Its value is 1.43×10^{-5} m^2/s for air and 10^{-6} m^2/s for water. For Newtonian fluids such as water and air, the shearing stress and the velocity are linearly related so that the viscosity is only a function of the nature of the fluid.[2] For non-Newtonian fluids such as milk and blood,* the viscosity is also a function of the velocity gradient (see Figure 6.3). Most often, the viscosity in these cases diminishes when the velocity gradient increases, and these types of materials are called *pseudoplastics*. A Bingham plastic

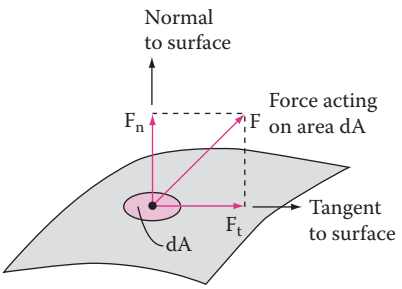


FIGURE 6.2 Pressure and shear stress in a fluid.

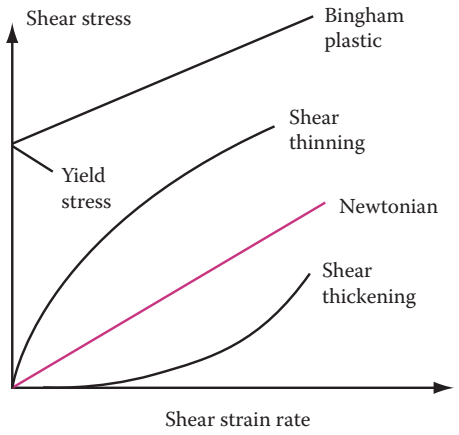


FIGURE 6.3 Newtonian and non-Newtonian fluids compared. The viscosity usually diminishes when the velocity gradient increases, and these types of materials are called *pseudoplastics*.

is a viscoelastic material that behaves as a rigid body at low stresses but flows as a viscous fluid at high stress. Once a critical shear stress (or "yield stress") is exceeded, the material flows as a Newtonian fluid.

Non-Newtonian flow behavior of fluids can offer real advantages. Paint, for example, should be easy to spread, so it must have a low viscosity at the high shear caused by the paintbrush. Many cleaning fluids and furniture waxes have similar properties. The causes of non-Newtonian flow depend on the colloid chemistry of the particular product.

The fundamental laws governing fluid motion are the so-called Navier-Stokes equations (NSEs) (see Navier and Stokes in Figure 6.4). The NSEs are 3D vector equations and represent the differential forms of three equations: 1) the conservation of mass, 2) the conservation of linear momentum, and 3) the conservation of energy (this third equation

(a) (b)

FIGURE 6.4 (a) French mathematician L. M. H. Navier (1758–1836) and (b) English mathematician Sir G. G. Stokes (1819–1903) formulated the Navier-Stokes equations by including viscous effects in the equations of motion.

* Although serum is a Newtonian fluid, the presence of red blood cells makes whole blood non-Newtonian. The same is true for mayonnaise; the oil in it is Newtonian, but the emulsion with yolk makes it non-Newtonian.

that makes up the complete set of NSEs will not be separately derived here). The Navier-Stokes equations are applicable to the description of the motion of a fluid volume at an arbitrary location in a flow field at any instant of time. One often classifies flows as a means to simplify the assumptions governing the partial differential Navier-Stokes equations. Lower dimensional flows reduce the complexity of analytical and computational solutions. We start by introducing the NSEs and show how they simplify depending on the type of flow. Next we derive two of the actual NSEs.

1. The Conservation of Mass Equation. The continuity equation in differential form is given as:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{V}) = 0 \tag{6.5}$$

Mass is neither created nor destroyed; there are no sources or sinks, or the net inflow rate, i.e., (inflow – outflow), must equal the accumulation rate. For one direction (e.g., the x-direction with velocity component u) this yields:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho u)}{\partial x} = 0 \tag{6.6}$$

A flow is classified as incompressible if the density remains constant. For an incompressible fluid $\left(\rho = \text{constant or } \frac{\partial \rho}{\partial t} = 0\right)$, this results in:

$$\nabla \cdot u = 0 \text{ (for 1D) and } \nabla \cdot \mathbf{V} = 0 \text{ (for 3D)} \tag{6.7}$$

Note that the above is true even for unsteady flows. Liquid flows typically are incompressible, but gas flows are compressible, especially at high speeds. The Mach number, Ma = V/c, is the ratio of flow velocity to the speed of sound (c) and may be regarded as the ratio of inertial forces to elastic ones. The Mach number is a good indicator of whether compressibility effects are important (Figure 6.5).

2. The Conservation of Momentum. In vector notation, the conservation of momentum is given as:

$$\underset{a}{\rho \frac{\partial \mathbf{V}}{\partial t}} + \underset{b}{\rho (\mathbf{V} \cdot \nabla)\mathbf{V}} = \underset{c}{-\nabla p} + \underset{d}{\eta \nabla^2 \mathbf{V}} + \underset{e}{\rho \mathbf{g}} \tag{6.8}$$

FIGURE 6.5 Ma = V/c. With Ma < 0.3 (~100 m/s), incompressible; with Ma < 1, subsonic; with Ma = 1, sonic; with Ma > 1, supersonic; and with Ma >> 1, hypersonic.

where $\boldsymbol{g}$ is the gravitational acceleration, p is the driving pressure, and t is time. The gravitational acceleration is related to the gravitational potential Ω as $\mathbf{g} = -\nabla\Omega$. Each term in Equation 6.8 has the dimension of $\{m/L^2t^2\}$.

The term a is the local acceleration of the fluid element. For a steady flow this term is zero. Steady flow implies no change at any point with time; transient terms in the NSE are then also zero:

$$\frac{\partial \mathbf{V}}{\partial t} = \frac{\partial \rho}{\partial t} = 0 \tag{6.9}$$

Transients usually describe a starting or developing flow (Figure 6.6). A change in coordinate system (e.g., Cartesian, cylindrical, spherical) may facilitate the mathematical treatment. For example, from Figure 6.6, for a fully developed pipe flow, velocity $\mathbf{V}(r)$ is a function of radius r, and pressure $p(z)$ is a function of distance z along the axis.

The next term in Equation 6.8 is the term b, which represents the convective acceleration of the fluid particle, and it predicts how the flow differs from one location to the next at the same instant of time. Uniform flow has no convective acceleration. The c term is the pressure acceleration caused by the "pumping" action. The d term is the viscous deceleration term resulting from the fluid's frictional resistance. The e term is the acceleration caused by gravity (weight). Regions where frictional effects are dominant are called viscous regions. They usually are close to solid surfaces (Figure 6.7). Regions where frictional forces are small compared with inertial or pressure forces are called inviscid. Such fluids are

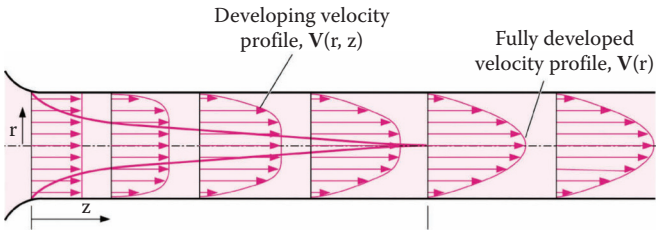


FIGURE 6.6 From a developing velocity profile to a fully developed velocity profile.

considered to be ideal, and in that case the *d* term in Equation 6.8 can be neglected:

$$\rho\frac{\partial \mathbf{V}}{\partial t}+\rho\left(\mathbf{V}\cdot\nabla\right)\mathbf{V}=-\nabla\mathbf{p}+\rho\mathbf{g} \qquad (6.10)$$

This expression is known as the Euler equation of motion. Even though this relation is simpler than the Navier-Stokes, an analytical solution is still not generally possible because the advective accelerations make the equations nonlinear.

Internal flows (e.g., inside a pipe), as illustrated in Figure 6.6, are dominated by the influence of viscosity throughout the flow field, whereas in external flows, viscous effects are limited to the boundary layer and wake as illustrated in Figure 6.8.

In the case of very slow fluid motion, the total acceleration term (a + b) as well as the acceleration caused by gravity (*e*) are zero. Equation 6.8 then reduces to the equation for Stokes flow or creep flow:

$$\nabla p=\eta\nabla^2\mathbf{V} \qquad (6.11)$$

This equation applies when dealing with very viscous fluids, capillary flow, and certain molten metals. Flows are also distinguished as being laminar, turbulent, or transitional, as shown in Figure 6.9.

An important concept in fluidics concerns the idea of streamlines. A streamline is a path traced out by a

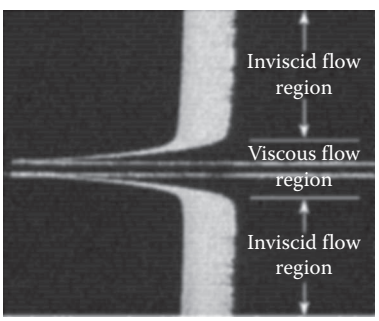


FIGURE 6.7 Inviscid and viscous regions in a liquid flow.

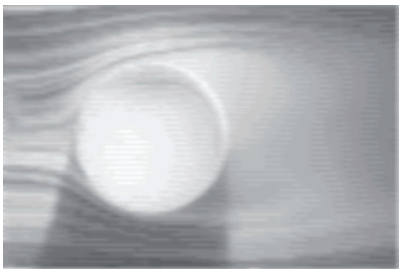


FIGURE 6.8 External flow. Viscous effects are limited to the boundary layer and wake around the object in the flow.

massless particle as it moves with the flow. Velocity is tangent to a streamline at every point.

In the laminar case, the fluid motion is highly ordered with smooth streamlines. In the case of turbulence, one has a highly disordered fluid motion characterized by velocity fluctuations and eddies. In the transitional case, one obtains a flow that contains both laminar and turbulent regions. We will see below that it is the Reynolds number, $Re = \rho VL/\eta$, that is the key parameter in determining whether a flow is laminar or turbulent. A rotational flow is one where streamlines close in on themselves while in an irrotational flow the streamlines do not close in on themselves.

Boundary Conditions

Continuum Hypothesis

In a continuum all quantities are defined everywhere and vary continuously from point to point. To understand this better, consider an "average" density, ρ, averaged over a fluid "particle" with volume L^3 as illustrated in Figure 6.10. For the "average density," it

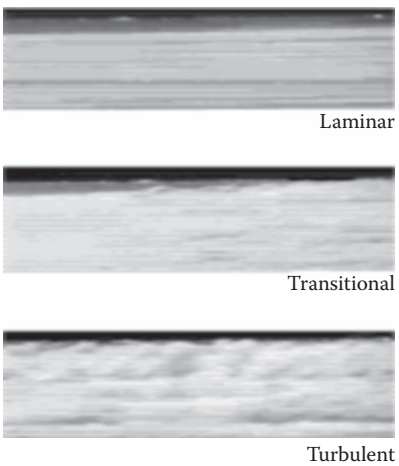


FIGURE 6.9 Laminar, transitional, and turbulent streamlines.

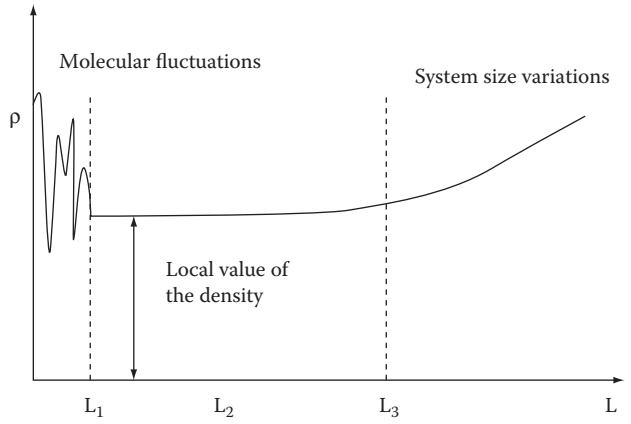


FIGURE 6.10 Schematic representation of the variation of instantaneous density of the fluid particle with respect to its volume.

holds that $\rho = N \cdot m/L^3$, with N the number of molecules and m the mass of each molecule. The continuum assumption fails when the sampling volume becomes too small. For the smallest particle size (L_1), molecular fluctuations may cause density variations. Also, for large volume elements, of the scale of the system itself (L_3), density can evolve because of spatial variations caused by changes in, for example, pressure (e.g., shocks) or temperature. It is only in the intermediate, "mesoscopic" size regime (L_2), the plateau in Figure 6.10, where it is possible to establish equations of fluid flow assuming a continuous medium. The continuum hypothesis thus assumes that in the L_2 regime (with $L_1 \ll L_2 \ll L_3$), the fluid can be treated as a continuum (i.e., not made up of individual molecules) with the properties of the system in the plateau region. Liquid particles in the plateau regime typically will have many, many molecules, even in a sample volume as small as a 1-μm cube. Liquid particles in a typical microchannel (e.g., more than a couple of micrometers wide) therefore still obey the Navier-Stokes equations. It is only when considering nanochannels that the situation needs to be re-evaluated. In the case of gases, it is easier to go beyond the NSE behavior, and even with sample volumes of 1 μm^3, deviations become apparent. For example, the continuum hypothesis is not valid anymore for:

1. Rarefied gases: $L_2 \sim \lambda$ (mean free path), as discussed in Volume III, Chapter 7, "Breakdown of Continuum Theory in Electrostatics" (see page 380).
2. Knudsen diffusion: diffusion within very small pores where d, the pore diameter, $< \lambda$ and $L_3 \sim d$ (i.e., no plateau), a situation we encountered when discussing evaporation from a Knudsen cell in Volume II, Chapter 7 (see Equation 7.5).

Summarizing, the continuum theory works well as long as local properties such as density and velocity can be defined as averages over elements large compared with the microscopic structure of the fluid but small enough compared with the scale of macroscopic phenomena. Importantly, we will learn below that the flow also should not be too far from thermodynamic equilibrium.

No-Slip Condition

A fluid in direct contact with a solid "sticks" to the surface as a result of viscous effects as illustrated in Figure 6.11. This sticking is responsible for the generation of wall shear stress τ_w, surface drag $D = \int \tau_w dA$, and the development of the boundary layer. The fluid property responsible for this no-slip condition is the viscosity, η. According to the no-slip condition, for a fluid in contact with a solid wall, the velocity of the fluid must equal that of the wall, as illustrated $v_{fluid} = v_{wall}$ in Figure 6.11.

This no-slip condition at the fluid-solid interface in momentum space has an equivalent in energy space with the analogous requirement for a no-temperature jump at the fluid-solid interface. Together these restrictions ensure that there are no finite discontinuities that would involve infinite velocity/temperature gradients. The latter, in continuum modeling, would lead to infinite viscous stress/heat flux. In other words, the fluid velocity must be zero relative to the solid, and the fluid temperature must equal that of the surface. This ultimately requires thermodynamic equilibrium between fluid and solid.[3] In the section on microscale laws for fluid flow, where we cannot

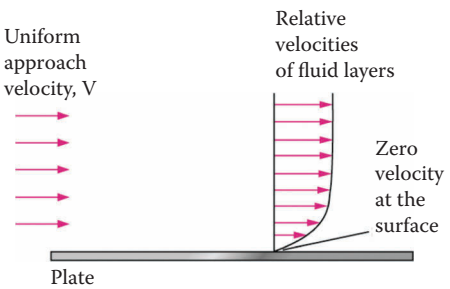


FIGURE 6.11 No-slip condition: zero velocity at the wall.

necessarily assume that the no-slip and the no-temperature jumps at the fluid-solid interface apply, we investigate what happens when microscopic effects start playing a role in determining fluid behavior.

Navier-Stokes Equations Derivation

Continuity Equation

Abbreviated, the derivation of the NSEs follows. The continuity equation can be derived with reference to Figure 6.12, where we apply the conservation of mass to a small control volume $\Delta V = \Delta x \Delta y \Delta z$, through which there is fluid flow. The rate of accumulation of mass in ΔV is:

$$\frac{\partial \rho}{\partial t}\Delta V \tag{6.12}$$

The rate of inflow across a control surface plus the rate of change of mass inside the control volume ΔV equals the rate of outflow across the control surface:

$$\text{Inflow through S} = \rho u \Delta y \Delta z + \rho v \Delta x \Delta z + \rho w \Delta x \Delta y$$

Rate of outflow through S

$$= \left[\rho u + \frac{\partial(\rho u)}{\partial x}\Delta x\right]\Delta y \Delta z + \left[\rho v + \frac{\partial(\rho v)}{\partial y}\Delta y\right] \Delta x \Delta z + \left[\rho v + \frac{\partial(\rho w)}{\partial z}\Delta z\right]\Delta x \Delta y \tag{6.13}$$

and, because mass is neither created nor destroyed, we apply Equation 6.12 and divide by $\Delta V = \Delta x \Delta y \Delta z$, resulting in the earlier introduced continuity Equations 6.5 and 6.7:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{V}) = 0 \text{ (compressible)} \tag{6.5}$$

$$\nabla \cdot \mathbf{V} = 0 \text{ (incompressible)} \tag{6.7}$$

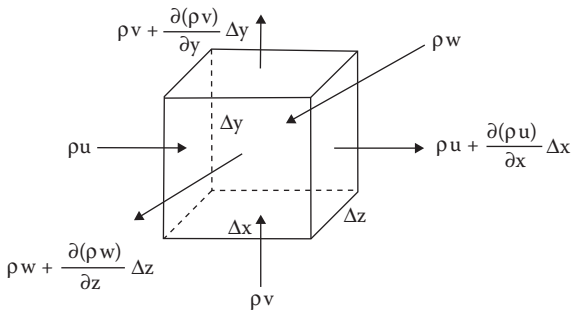


FIGURE 6.12 Mass balance: what goes in must come out.

Momentum Equation

Now we will set up the momentum equation. Even in a steady flow, a fluid particle can change momentum by moving to a position where its velocity is different; this acceleration requires that a force be applied to the fluid particle. This situation is equivalent to that of a change in temperature of a fluid particle over time Δt, during which it also has moved over a distance (Δx, Δy, Δz). This heat transfer problem may be expressed as:

$$dT = \frac{\partial T}{\partial x}\Delta x + \frac{\partial T}{\partial y}\Delta y + \frac{\partial T}{\partial z}\Delta z + \frac{\partial T}{\partial t}\Delta t \Rightarrow \frac{dT}{dt} = \frac{\partial T}{\partial x}u + \frac{\partial T}{\partial y}v + \frac{\partial T}{\partial z}w + \frac{\partial T}{\partial t} \tag{6.14}$$

This can be written as a total substantive derivative (also called the Stokes derivative) or:

$$\frac{DT}{Dt} = \frac{\partial T}{\partial x}u + \frac{\partial T}{\partial y}v + \frac{\partial T}{\partial z}w + \frac{\partial T}{\partial t} = \frac{\partial T}{\partial t} + \mathbf{V}\cdot\nabla T \tag{6.15}$$

The operator in both the heat transfer example and in convective fluid transport is the same, i.e.:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{V}\cdot\nabla \tag{6.16}$$

This equation sums the local rate of change in $\frac{\partial}{\partial t}$ and the convective rate of change in $\mathbf{V} \cdot \nabla$. Momentum is a vector quantity, with the total rate of change of momentum for an incompressible fluid (per unit volume) given as:

$$\frac{D}{Dt}(\rho \mathbf{V}) = \rho(\mathbf{V}\cdot\nabla)\mathbf{V} + \rho\frac{\partial \mathbf{V}}{\partial t} \tag{6.17}$$

We now apply Newton's Second Law to Equation 6.17: for a small fluid element ($F = ma = \rho \Delta x \Delta y \Delta z a$) the rate of change of momentum equals the sum of all applied forces on that fluid element. For the sum of the forces in the x-direction, we may write that $\sum F_x = \rho \Delta x \Delta y \Delta z \frac{Du}{Dt}$ (u being the x-vector of speed $\mathbf{V}$). We first sum up all the body forces in the x-direction: the x-component of the gravity force acting on the fluid element is $\Delta F_g = \rho \Delta x \Delta y \Delta z g_x$

(weight), and the force caused by the pressure gradient in the *x*-direction is $\Delta F_p = -\frac{\partial p}{\partial x}\Delta x\Delta y\Delta z$. Next, we consider the surface forces. The latter are exerted on the fluid body as it interacts with its surroundings, and both normal stresses and shear stresses are involved as shown in Figure 6.13. The forces in the x-direction caused by viscous stresses are given as $\Delta F_s = \frac{\partial \sigma_{xx}}{\partial_x}\Delta x\Delta y\Delta z + \frac{\partial \tau_{yx}}{\partial_y} + \Delta x\Delta y\Delta z + \frac{\partial \tau_{zx}}{\partial z}\Delta x\Delta y\Delta z$. Thus, for an intermediate solution we find for the total change of momentum in the x-direction:

$$\begin{aligned}\sum F_x &= \Delta F_g + \Delta F_s + \Delta F_p \\ &= \rho\Delta x\Delta y\Delta z g_x + \frac{\partial \sigma_{XX}}{\partial x}\Delta x\Delta y\Delta z \\ &\quad + \frac{\partial \tau_{yx}}{\partial y}\Delta x\Delta y\Delta z + \frac{\partial \tau_{zx}}{\partial z}\Delta x\Delta y\Delta z - \frac{\partial p}{\partial x}\Delta x\Delta y\Delta z \\ &= \rho\Delta x\Delta y\Delta z\frac{Du}{Dt}\end{aligned}$$

and after dividing by $\Delta x\Delta y\Delta z$ we obtain

$$\rho\frac{Du}{Dt} = -\frac{\partial p}{\partial x} + \rho g_x + \left(\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z}\right) \quad (6.18)$$

In Chapter 4 we saw that the stress gradients in Equation 6.18 can be rewritten as:

$$\left(\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z}\right) = 2\eta\frac{\partial^2 u}{\partial x^2} + \eta\frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) + \eta\frac{\partial}{\partial z}\left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right) \quad (6.19)$$

where viscosity, η, for the first time enters our derivation. This expression can be simplified by

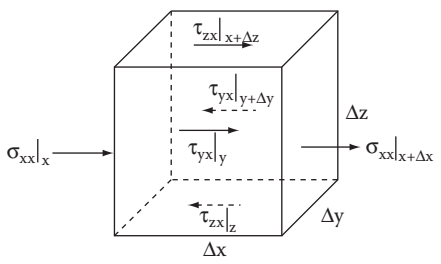


FIGURE 6.13 Surface forces exerted on an infinitesimal control volume $\Delta x\Delta y\Delta z$ include normal stresses and shear stresses.

using the continuum assumption (Equation 6.7, for 1D):

$$\begin{aligned}&= \eta\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) + \eta\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) \\ &\quad\Downarrow \qquad\qquad\qquad\Downarrow \\ &= 0, \text{ since } \nabla\cdot u = 0 \qquad \eta\cdot\nabla^2\cdot u\end{aligned} \quad (6.20)$$

Thus, we obtain the equation for the x-momentum as:

$$\rho\frac{Du}{Dt} = -\frac{\partial p}{\partial x} + \eta\nabla^2 u + \rho g_x \quad (6.21)$$

and similarly for the y- and z-directions. We can write all three momentum equations together, using vector notation, and using Equation 6.17, we obtain Equation 6.8, i.e., the irrotational NSE for constant density and viscosity:

$$\rho\frac{\partial \mathbf{V}}{\partial t} + \rho(\mathbf{V}\cdot\nabla)\mathbf{V} = -\nabla p + \eta\nabla^2\mathbf{V} + \rho\mathbf{g} \quad (6.8)$$

The gravitational acceleration **g** is the negative of the gradient of the potential energy (–∇Ω). We also introduce the reduced NSEs here; the latter formulation will enable us to discuss "similar" situations (see the concept of similarity in Volume III, Chapter 7 on scaling). The nondimensional NSEs will come in handy, for example, when we compare the efficiency of different separation techniques further below. We first define nondimensional variables, according to Table 6.1.

To obtain the dimensionless velocity, we write: $\mathbf{V}^* = \mathbf{V}/V$, for dimensionless length: $\mathbf{x}^* = \mathbf{x}/L$, for dimensionless gravity: $\mathbf{g}^* = \mathbf{g}/g$, and for dimensionless time we get: $t^* = ft$, with f the characteristic frequency. For pressure: $p^* = p/p_0 - p_\infty$, and the nabla or del operator as $\nabla^* = L\nabla$. Substituting these parameters in the NSE results in:

$$\begin{aligned}&\rho V f\frac{\partial \mathbf{V}^*}{\partial t^*} + \frac{\rho\mathbf{V}^2}{L}(\mathbf{V}^*\cdot\nabla^*)\mathbf{V}^* \\ &= -\frac{p_0 - p_\infty}{L}\nabla^* p^* + \frac{\eta V}{L^2}\nabla^{*2}\mathbf{V}^* + \rho g g^*\end{aligned} \quad (6.22)$$

To nondimensionalize Equation 6.22, we multiply all terms by $L/(\rho\mathbf{V}^2)$, which has primary dimensions $\{m^{-1}L^2t^2\}$, so that the dimensions cancel. After rearrangement:

Table 6.1 Scaling Parameters Used to Nondimensionalize the Continuity and Momentum Equations, along with Their Primary Dimensions

Scaling Parameter	Description	Primary Dimensions
L	Characteristic length	{L}
V	Characteristic speed	$\{Lt^{-1}\}$
f	Characteristic frequency	$\{t^{-1}\}$
$p_0 - p_\infty$	Reference pressure difference	$\{mL^{-1}t^{-2}\}$
g	Gravitational acceleration	$\{Lt^{-2}\}$

$$\underbrace{\left[\frac{fL}{V}\right]}_{\text{Strouhal}}\frac{\partial \mathbf{V}^*}{\partial t^*}+(\mathbf{V}^*\cdot\nabla^*)\mathbf{V}^*=-\underbrace{\left[\frac{p_0-p_\infty}{\rho V^2}\right]}_{\text{Euler}}\nabla^* p^*+\underbrace{\left[\frac{\eta}{\rho VL}\right]}_{\text{Inverse Reynolds}}\nabla^{*2}\mathbf{V}^*+\underbrace{\left[\frac{gL}{V^2}\right]}_{\text{Inverse Froude}}\mathbf{g}^*$$

which can be written as:

$$[St]\frac{\partial \mathbf{V}^*}{\partial t^*}+(\mathbf{V}^*\cdot\nabla^*)\mathbf{V}^*=-[Eu]\nabla^* p^*+\left[\frac{1}{Re}\right]\nabla^{*2}\mathbf{V}^*+\left[\frac{1}{Fr^2}\right]\mathbf{g}^* \qquad (6.23)$$

In some textbooks $\mathbf{g}^*$ is replaced by $-\mathbf{k}$ (vector along the *z*-axis); the *x*- and *y*-axes are assumed to be in a horizontal plane relative to the earth's surface, so one can write $\mathbf{g} = -9.81\ \mathbf{k}(m/s^2)$.

Writing the NSEs in dimensionless form not only reduces the number of free parameters, it also shows the appropriate limiting form of the equations if the dimensionless numbers approach extreme values. To normalize Equation 6.23, we must choose scaling parameters *L*, *V*, etc., that are appropriate for the flow being analyzed, such that all nondimensional variables are of order of magnitude unity, i.e., their minimum and maximum values are close to 1.0 (e.g., $t^* \sim 1$, $V^* \sim 1$, $p^* \sim 1$, etc.). If the NSEs are properly normalized, one can compare the relative importance of the terms in the equation by comparing the relative magnitudes of the nondimensional parameters Strouhal (St), Euler (Eu), Froude (Fr), and Reynolds (Re).

With Re tending to zero, we obtain Stokes or creep flow (with St ~ 1 and Fr ~ 1) (see Equation 6.11):

$$\underbrace{[Eu]}_{\text{Pressure forces}}\nabla^* p^*=\underbrace{\left[\frac{1}{Re}\right]}_{\text{Viscous forces}}\nabla^{*2}\mathbf{V}^* \qquad (6.24)$$

and because $p^* \sim 1$ and $\nabla^* \sim 1$:

$$Eu=\frac{p_0-p_\infty}{\rho V^2}\sim\frac{1}{Re}=\frac{\eta}{\rho VL}$$

and:

$$p_0-p_\infty\sim\frac{\eta V}{L} \qquad (6.25)$$

This result is very different from inertia-dominated flow (see below), where $p_0 - p_\infty \sim \rho V^2$. Equations 6.11 and 6.25 are linear equations that are much easier to solve for simple geometries than the full NSEs.

Example 6.1: Falling Sphere (I): Viscid Regime

With a small sphere, of radius *r*, falling through a viscous fluid the inertial forces are small compared with viscous forces (Re < 1). In this case, fluid motion is smooth and transfer is molecular. That is, momentum transfers from molecule to molecule through the fluid, and fluid flows in layers. The analytical solution of the NSEs for creeping flow around the falling sphere gives a drag force that, as predicted by Equation 6.25, is a linear function of velocity *V* and viscosity η:

$$F_D = 6\pi\eta Vr \qquad (6.26)$$

Equation 6.26 is the Stokes' Law (1851, George Gabriel Stokes), where *V* is the terminal velocity. When an object starts to fall (free fall), it accelerates downward until it reaches a terminal, maximal velocity *V*. At this velocity, the forces acting on the body, i.e., drag forces, gravity, and buoyancy, have reached equilibrium, and acceleration stops due to gravity working on the body. In other words:

$$\sum F = 0 = \underbrace{F_D}_{\text{Friction}} + \underbrace{F_g}_{\text{Gravity}} + \underbrace{F_b}_{\text{Buoyancy}} \text{ or}$$

$$0 = 6\pi\eta Vr + \frac{8\pi r^3}{6}\left(\rho_p - \rho_f\right)g \text{ or} \qquad (6.27)$$

$$V = \frac{2\left(\rho_p - \rho_f\right)gr^2}{9\eta}$$

The terminal velocity thus depends on the density (ρ_f) and viscosity (η) of the fluid, as well as the shape, size (*L*), and density of the object (ρ_p).

Viscid (Stokes), Inviscid (Euler), and Froude Flow

Einstein, in his 1905 paper on Brownian motion, established the following unexpected relation:

$$D = \mu k_B T \tag{6.28}$$

linking D, the diffusion constant, and μ, the mobility of the particles, where k_B is Boltzmann's constant, and T is the absolute temperature. The mobility μ is the ratio of the particle's terminal drift velocity to an applied force, $\mu = V/F_D$. Substituting this in Equation 6.26 from Example 6.1 one obtains the Stokes-Einstein formula for the molecular diffusion coefficient—another example of Stokes flow:

$$D = \frac{k_B T}{6\pi\eta r} \tag{6.29}$$

where k_B = Boltzmann's constant ($1.38 \times 10\ m^{-23}\ J\ K^{-1}$)
T = the absolute temperature (K)
η = the absolute viscosity (kg/m·s), and
r = the hydrodynamic radius

It is this very relation that allowed Perrin, using colloid particles as molecules, to measure Boltzmann's constant (Figure 6.14).

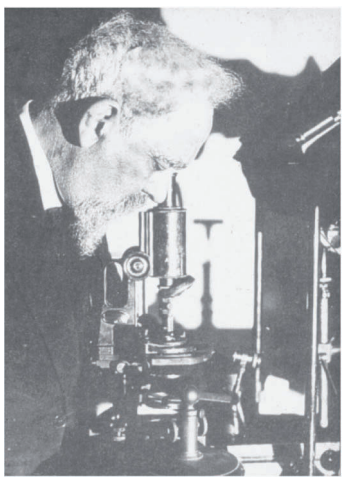


FIGURE 6.14 J. Perrin (1870–1942). Nobel Prize for Physics in 1926. "For his work on the discontinuous structure of matter, and especially for his discovery of sedimentation equilibrium." "I do not believe that it is possible to study Brownian motion with such precision," Einstein wrote to Perrin in 1910.

In Volume III, Chapter 7 on scaling, we derived a drag coefficient C_D for flow around a submerged object as:

$$C_D = \frac{F_D}{\frac{1}{2}\rho V^2 A} \qquad \text{(Volume III, Equation 7.27)}$$

where ρ = the density of the fluid
V = the free stream velocity (far from the object)
A = the characteristic area of the body: the projected area; the projected area is the area "seen" by the fluid area; for a spherical particle with diameter L, the area $A = \pi L^2/4$ (see Figure 6.15)*
F_D = the drag force

Keep in mind that this drag coefficient C_D is analogous to the friction factor in pipe flow (see below).

As derived in Volume III, Chapter 7 on scaling, we may express the drag coefficient C_D and friction force F_D also in terms of the Reynolds number Re:

$$C_D = F_D/\rho l^2 v^2 = f\,(Re) \qquad \text{(Volume III, Equation 7.8)}$$

A plot of drag coefficient versus Reynolds number is shown in Figure 6.16. For a given Reynolds number, the best design for a race car, submarine, or airplane is the one that minimizes the drag coefficient C_D (see also Volume III, Figure 7.8). In the Stokes (viscid)

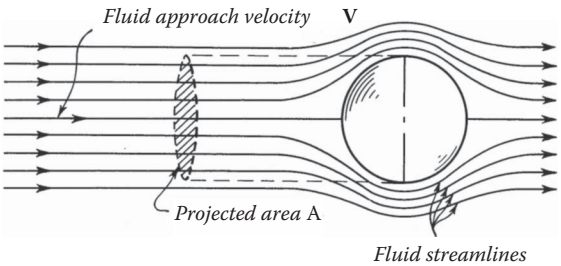


FIGURE 6.15 The projected area is the area "seen" by the fluid area. For a spherical particle $A = \pi L^2/4$ in a fluid stream.

* A = frontal area (thick body, cylinder, cars, projectiles)
= plan area (wide/flat body, wings, hydrofoils)
= wetted area (surface ships, barges)
This is a simplification, as both frictional and pressure drag forces should be considered, and factor A has a different meaning in both. For a more rigorous treatment, see Granger (1995).[6]

regime with values of Re < 1 (see Figure 6.16) one obtains:

$$C_D = \frac{24}{Re} \tag{6.30}$$

In inviscid flow, the net viscous forces are negligible compared with pressure and/or inertia forces:

$$[St]\frac{\partial \mathbf{V}^*}{\partial t^*} + (\mathbf{V}^* \cdot \nabla^*)\mathbf{V}^* = -[Eu]\nabla^* p^* + \left[\frac{1}{Fr^2}\right]\mathbf{g}^* \left[\frac{1}{R}\right]\nabla^{*2}\,\mathbf{V}^* \tag{6.31}$$

With a very large Reynolds number Re (= $\rho VL/\eta$) the inertial effects dominate over the viscous effects and we obtain Euler's expression (Equation 6.10). Elimination of the viscous term changes the partial differential equations from mixed elliptic-hyperbolic to hyperbolic. This affects the type of analytical and computational tools used to solve the equations. This result is very different from viscous flow, with the drag force, $F_D = C_D\,\rho V^2 A/2$, now proportional to ρV^2. The Euler equation often is used in aerodynamics (Figure 6.17).

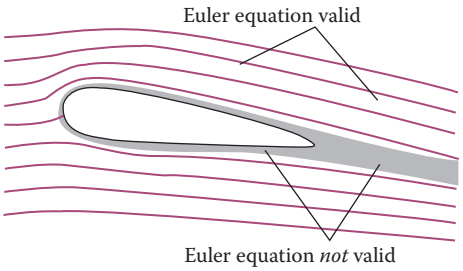


FIGURE 6.17 Regions of validity of the Euler equation in aerodynamics.

Example 6.2: Falling Sphere (II): Inviscid Flow

At high Reynolds numbers (Re > 1000), with negligible viscous forces and inertia-dominated flow, Newton's resistance law predicts that the force on the falling sphere will be proportional to the fluid pushed away and the relative velocity between the sphere and the fluid. In this case the flow is large, fluid motion is chaotic, and transfer is with blocks of molecules moving in all directions, causing eddy currents. For Re ⋙ 1, the flow pattern is very complicated; thus, very little theory exists to predict the relationship between C_D and Re. As the flow increases, wake drag becomes an important factor. The streamline pattern becomes mixed at the rear of the sphere and at very high Reynolds numbers completely separates in the wake. From experimental results shown in Figure 6.16, for $1000 < Re < 10^5$, C_D is about 0.44. In the transition regime, the drag coefficient is:

$$C_D = \frac{24}{Re}\left(1 + \frac{Re^{2/3}}{6}\right) \tag{6.32}$$

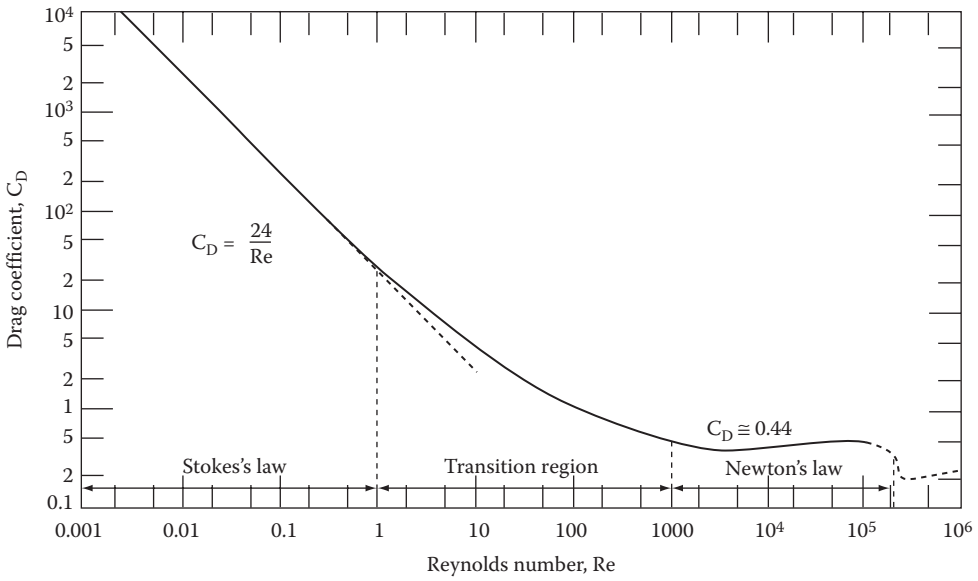


FIGURE 6.16 Drag coefficient C_D of a sphere versus Reynolds number Re.

The Froude number in Equation 6.31 represents the ratio of inertial to gravitational forces. If the Froude number is large, gravitational forces are not important and the $1/Fr^2$ term can be neglected. In many cases, the Froude number is large and gravity is unimportant; exceptions are in modeling ships and waves. Waves involve vertical displacements and therefore, gravity is involved (see also Volume III, Chapter 7, Equations 7.24–7.26) (Figure 6.18).

For a more detailed derivation of the NSEs, including the conservation of energy equation, refer to White,[4] Allen,[5,6] and Denn.[1] The NSEs cannot be solved unless some terms drop out and the equations simplify. There are about 80 known exact solutions to the NSEs. Various computational fluid dynamic (CFD) codes for solving dynamic behavior of fluids in microsystems built on these equations are commercially available (see Volume III, Chapter 6).

Navier-Stokes Applications

Introduction

Here, we present five applications of the NSEs. As warm-up applications we derive the expression for hydrostatics and calculate the Bernoulli expression from the Euler equation for flow. We then move on to pressure-driven flow between two large plates, derive the Couette flow, and finally introduce Hagen-Poiseuille's law for pipe flow.

Hydrostatics

An extreme simplification of the NSEs involves hydrostatics where the velocity $\mathbf{V} = 0$ (Figure 6.19).

FIGURE 6.18 In modeling waves the Froude number is important.

With gravity in the z-direction, only the following NSE term survives:

$$\rho g = \frac{\partial p}{\partial z} \tag{6.33}$$

More generally, this can be interpreted as saying that pressure varies only in the direction of net acceleration. With a constant density, Equation 6.33 can be integrated to Figure out how pressure varies with depth, or:

$$\int dp = p(z) = \int \rho g dz = \rho g \int dz = \rho g z + p_0 \tag{6.34}$$

Because the fluid is not moving, η does not play a role.

Bernoulli's Equation

Remember that there is no flow across a streamline because at every point the streamline is a tangent to the velocity. Therefore, a streamline can be thought of as if it were a solid boundary. The superposition of irrotational, inviscid flow solutions gives us tools to construct velocity distributions but does not yet tell us anything about the pressure distribution in the flow. For that we need to derive the Bernoulli equation connecting pressure, velocity, and elevation. To derive that equation, we start from the

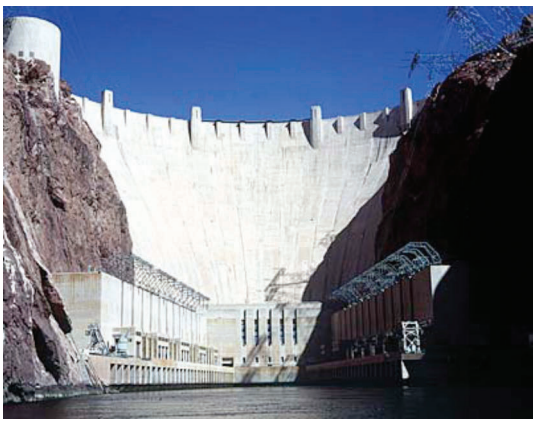


FIGURE 6.19 The Hoover Dam: for an incompressible fluid at rest, pressure increases linearly with depth. As a consequence, large forces can develop on plane and curved submerged surfaces. The water behind the Hoover Dam, on the Colorado River, is approximately 715 ft. deep, and at this depth the pressure is 310 psi. To withstand the large pressure forces on the face of the dam, its thickness varies from 45 ft. at the top to 660 ft. at the base.

NSEs for inviscid flow, i.e., the dimensional Euler expression:

$$\frac{\partial \mathbf{V}}{\partial t}+(\mathbf{V}\cdot\nabla)\mathbf{V}=-\frac{\nabla p}{\rho}+\mathbf{g} \tag{6.10}$$

and we write the gravity vector as:

$$\mathbf{g}=-g\nabla z \tag{6.35}$$

with z-positive upward. We also recall the following vector calculus identity:

$$(\mathbf{V}\cdot\nabla)\mathbf{V}=\frac{\mathbf{1}}{2}\nabla(\mathbf{V}\cdot\mathbf{V})-\mathbf{V}\times(\nabla\times\mathbf{V}) \tag{6.36}$$

With these two last equations we can rewrite the Euler equation as:

$$\begin{aligned}&-\rho g\nabla z-\nabla p=\frac{\rho}{2}\nabla(\mathbf{V}\cdot\mathbf{V})-\rho\mathbf{V}\times(\nabla\times\mathbf{V})\\&\Rightarrow\frac{\nabla p}{\rho}+\frac{1}{2}\nabla(\mathbf{V}^2)+g\nabla z=\mathbf{V}\times(\nabla\times\mathbf{V})\end{aligned} \tag{6.37}$$

Now we take the dot product with differential length $d\mathbf{s}$ along a streamline:

$$\frac{\nabla p}{\rho}\cdot d\mathbf{s}+\frac{1}{2}\nabla(\mathbf{V}^2)\cdot d\mathbf{s}+g\nabla z\cdot d\mathbf{s}=\mathbf{V}\times(\nabla\times\mathbf{V})\cdot d\mathbf{s} \tag{6.38}$$

Because $d\mathbf{s}$ and $\mathbf{V}$ are parallel:

$$\begin{aligned}&\mathbf{V}\times(\nabla\times\mathbf{V})\text{ is perpendicular to }\mathbf{V}\\&\mathbf{V}\times(\nabla\times\mathbf{V})\cdot d\mathbf{s}=0\end{aligned} \tag{6.39}$$

Thus, the term on the right in Equation 6.38 = 0. The $d\mathbf{s}$ term can be expanded as:

$$d\mathbf{s}=dx\,\mathbf{i}+dy\,\mathbf{j}+dz\,\mathbf{k}$$

$$\text{So: }\nabla p\cdot d\mathbf{s}=\frac{\partial p}{\partial x}dx+\frac{\partial p}{\partial y}dy+\frac{\partial p}{\partial z}dz=dp \tag{6.40}$$

Putting this last equation in the expression for flow results in:

$$\frac{dp}{\rho}+\frac{1}{2}d(\mathbf{V}^2)+gdz=0 \tag{6.41}$$

Integration along a streamline then results in:

$$\int\frac{dp}{\rho}+\frac{1}{2}d(\mathbf{V}^2)+gdz=\text{constant} \tag{6.42}$$

For an incompressible fluid this can be rewritten as:

$$\frac{p}{\rho}+\frac{V^2}{2}+gz=\text{constant} \tag{6.43}$$

What happens if we impose the additional restriction that the flow be irrotational? In the above derivation we relied on the fact that the restriction in Equation 6.39 holds, but now we also add to this that $\nabla\times\mathbf{V}=\mathbf{0}$, so that the direction of $d\mathbf{s}$ is irrelevant. What this means is that in an irrotational flow Bernoulli's equation can be applied at any two points in the flow, not just at two points on a streamline! A more popular formulation of Bernoulli's equation is the following (where $\gamma=g\times\rho$):

$$\begin{aligned}&\frac{p}{\gamma}+z+\frac{\mathbf{V}^2}{2g}=\text{constant}\\&\frac{p}{\gamma}+z=\text{piezometric head}\\&\frac{\mathbf{V}^2}{2g}=\text{velocity (dynamic) head}\\&\frac{p_1}{\gamma}+z_1+\frac{\mathbf{V}_1^2}{2g}=\frac{p_2}{\gamma}+z_2+\frac{\mathbf{V}_2^2}{2g}\end{aligned} \tag{6.44}$$

Example 6.3

Given: The velocity in an outlet pipe from a reservoir is 6 m/s and h = 15 m (see Figure 6.20).
Find: Pressure at point *A*.
Solution: Bernoulli's equation

$$\begin{aligned}&\frac{p_1}{\gamma}+z_1+\frac{V_1^2}{2g}=\frac{p_A}{\gamma}+z_A+\frac{V_A^2}{2g}\\&\frac{0}{\gamma}+h+\frac{0}{2g}=\frac{p_A}{\gamma}+0+\frac{V_A^2}{2g}\\&p_A=\gamma\left(h-\frac{V_A^2}{2g}\right)=9810\left(15-\frac{18}{9.81}\right)\\&p_A=129.2\text{ kPa}\end{aligned}$$

Unidirectional Flow between Two Infinite Plates

In Figure 6.21 we consider the unidirectional flow of a fluid between two infinite parallel plates. The plates are stationary, and the steady, incompressible flow is driven by a pressure gradient in the x-direction. We first decide what velocity components are nonzero

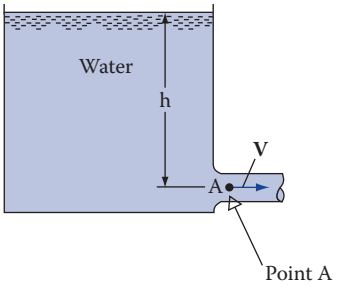


FIGURE 6.20 Velocity in outlet pipe from reservoir is 6 m/s and h = 15 m.

in this situation: w and v are = 0 and u is nonzero [u = u(y)].

This situation again simplifies the NSEs considerably:

$$0 = -\frac{\partial p}{\partial x} + \eta\frac{\partial^2 u}{\partial y^2}$$
$$0 = -\frac{\partial p}{\partial y} - \rho g$$
$$0 = -\frac{\partial p}{\partial z} \qquad (6.45)$$

After integrating the first expression twice and applying the no-slip conditions at the plates (this sets u = 0 at z = +h and at –*h*), we obtain:

$$u = \frac{1}{2\eta}\left(\frac{\partial p}{\partial x}\right)\left(y^2 - h^2\right) \qquad (6.46)$$

This solution is parabolic, as illustrated in Figure 6.22. To obtain the volumetric flow rate, we integrate:

$$Q = \int_{-h}^{h} u\,dy = \int_{-h}^{h} \frac{1}{2\eta}\left(\frac{\partial p}{\partial x}\right)\left(y^2 - h^2\right)dy$$
$$Q = -\frac{2h^3}{3\eta}\left(\frac{\partial p}{\partial x}\right) \qquad (6.47)$$

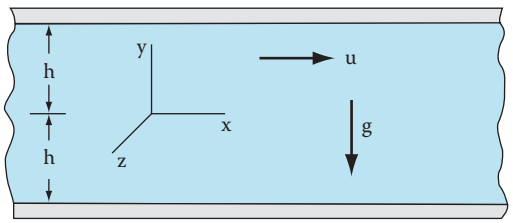


FIGURE 6.21 Unidirectional flow of a fluid between two large parallel plates. The plates are stationary, and the steady, incompressible flow is driven by a pressure gradient.

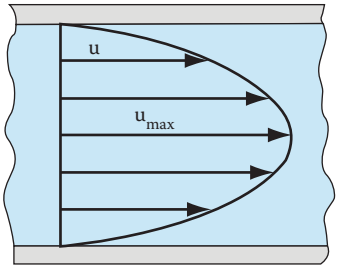


FIGURE 6.22 Parabolic velocity profile for flow of a fluid between two large parallel plates. Adjacent horizontal layers flow at different speeds, sliding over one another.

Fully Developed Couette Shear Flow

To illustrate Couette flow the same arrangement shown in Figure 6.22 is used; however, this time one plate is moving at a velocity U with respect to the other fixed plate (Figure 6.23). We simplify the Navier-Stokes equations with the same assumptions used above, except that here the no-slip boundary condition at the upper boundary is u(b) = U. This results in:

$$u = U\frac{y}{b} + \frac{1}{2\eta}\left(\frac{\partial p}{\partial x}\right)\left(y^2 - by\right) \qquad (6.48)$$

Making this expression dimensionless leads to:

$$\frac{u}{U} = \frac{y}{b} + \frac{b^2}{2\eta U}\left(\frac{\partial p}{\partial x}\right)\left(\frac{y}{b}\right)\left(1 - \frac{y}{b}\right) \qquad (6.49)$$

We can define $p = -\frac{b^2}{2\mu U}\left(\frac{\partial p}{\partial x}\right)$, which determines the effect of the pressure gradient as illustrated in

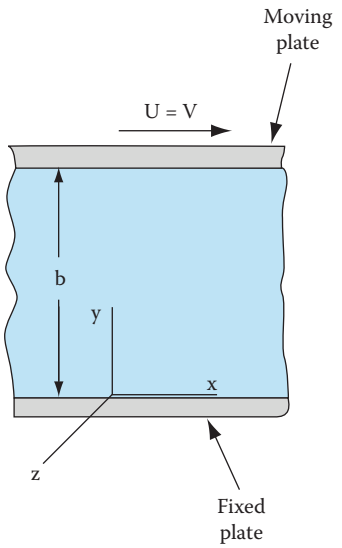


FIGURE 6.23 Couette flow between two infinitely large parallel plates.

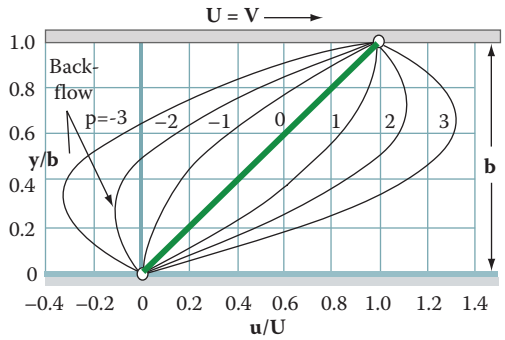


FIGURE 6.24 The dimensionless velocity as a function of reduced height and *p* value. Green slanted line is without pressure gradient.

Figure 6.23. In this case there is no pressure gradient $\left[\left(\frac{\partial p}{\partial x}\right)=0\right]$, and Equation 6.49 reduces to:

$$u = U\frac{y}{b} \tag{6.50}$$

The latter result corresponds with the slanted green line in Figure 6.24, where we show the dimensionless velocity as a function of reduced height and p value.

Hagen-Poiseuille's Law for Steady Duct Pipe Flow

Let us now consider pressure-driven flow in a pipe as shown in Figure 6.25. We assume steady and laminar flow, with flow only in the *z*-direction, so that $V_z = f(r)$ and $v_r = 0$ and $v_\theta = 0$ (we are working with cylindrical coordinates here):

$$\begin{aligned} 0 &= -\rho g\sin\theta - \frac{\partial p}{\partial r} \\ 0 &= -\rho g\cos\theta - \frac{1}{r}\frac{\partial p}{\partial \theta} \\ 0 &= -\frac{\partial p}{\partial z} + \eta\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_z}{\partial r}\right)\right] \end{aligned} \tag{6.51}$$

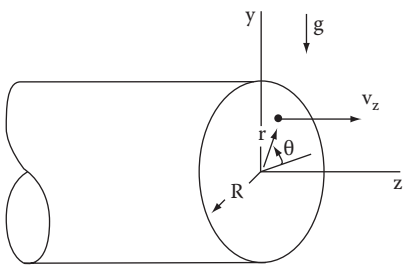


FIGURE 6.25 Pipe flow and cylindrical coordinates.

Solving these equations with the no-slip condition applied at r = R (the walls of the pipe), we obtain a parabolic velocity profile:

$$v_z = \frac{1}{4\eta}\left(\frac{\partial p}{\partial z}\right)\left(r^2 - R^2\right) \tag{6.52}$$

Experimentally, fluid viscosity may be determined based on this Hagen-Poiseuille law for laminar flow of a fluid through a capillary. Either the pressure difference over the capillary is measured for a known flow rate or the flow rate may be measured for a fixed pressure drop. The average velocity, *V*, often is used for convenience instead of flow rate, *Q*; the two are related through $V = Q/\pi R^2$. With reference to Figure 6.26, we calculate the volumetric flow rate *Q* as:

$$\begin{aligned} Q &= 2\pi\int_0^R v_z r\,dr \\ Q &= -\frac{\pi R^4}{8\eta}\left(\frac{\partial p}{\partial z}\right) \end{aligned} \tag{6.53}$$

The pressure gradient in macroscopic laminar flow in the capillary is given as $\Delta p/l$, where *l* is the length of the capillary. Substituting *Q* for average velocity and using $\Delta p/l$, we rewrite Equation 6.53 as the Hagen-Poiseuile equation or:

$$V = -\frac{R^2\Delta p}{8\eta l} \tag{6.54}$$

This clearly predicts the pressure drop over narrow capillaries to be very high and scaling as the inverse of R^2. The fact that the volumetric flow rate *Q* reduces as the fourth power of the radius (Equation 6.53) can be understood as follows. A gas or a liquid is at rest at the capillary wall and the fluid velocity, which reduces as the square of the channel radius (from Equation 6.54), is maximal at the center of the tube. In a wide tube, larger velocities can thus be achieved. This, coupled with the trivial factor of the larger diameter, results in a *Q* proportional to R^4.

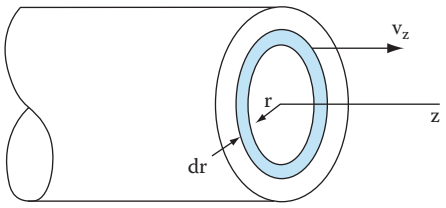


FIGURE 6.26 Volumetric flow rate in pipe flow.

What these scaling laws mean is that a 10 times reduction in the radius of the tube will mean a 10^4 reduction in volumetric flow and a 10^2 increase in pressure. Consequently, distributed micropumping, with each pump responsible for only a fraction of the total fluid movement, is expected to be important in microsystems. The required high pressures associated with mechanical pumping are a good reason to look at alternative propulsion techniques. Piezoelectric, electro-osmotic, electrowetting, and electrohydrodynamic pumping are all surface forces and scale more favorably than pressure pumping, which is a volume force, so they represent some considerable advantages for moving fluids in narrow capillaries.

Heat Conduction

Introduction

Many microdevices are actuated by thermal means (see Volume III, Chapter 8 on actuators). We are interested in learning about scaling laws in heat conduction, such as the total amount of heat required to start a thermal actuator, its heating and cooling rate, the thermal stresses or distortions expected in thermal sensors and actuators, and possible damage to delicate MEMS components. The three fundamental modes of heat transmission are conduction, convection, and radiation. They are governed by, respectively, Fourier's Law, Newton's Law of Cooling, and Stefan-Boltzmann's Radiation Law.

Fourier's Law of Heat Conduction

The total amount of heat, Q, flowing through a rectangular slab of material held at temperatures T_h and T_c (Figure 6.27) is proportional to the surface area A of the slab, the temperature difference between the two surfaces, and the time t during which heat flows, and it is inversely proportional to the distance the heat must travel ($\sim L^{-1}$), or:

$$Q = -\kappa \frac{A(T_h - T_c)t}{L} \tag{6.55}$$

where κ is the bulk heat conductivity in W/m K, and L is the plate surface separation. The value of κ is a measure of how good a heat conductor a given material is—for solids, κ generally increases with temperature. However, for most engineering materials at

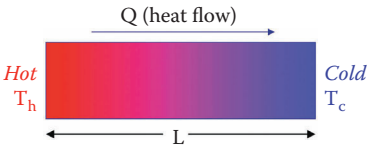


FIGURE 6.27 Heat conduction in a solid held at two different temperatures.

common operating temperatures, κ can be regarded as constant. Its value is 2.6 10^{-2} W/mK for air, 385 W/mK for copper, 50 W/mK for steel, 0.01 W/mK for Styrofoam, and 0.59 W/mK for water. The second law of thermodynamics requires the minus sign in Equation 6.55: thermal energy transfer from a thermal gradient must be from a warmer to a colder region. Heat transfer by conduction, from Equation 6.55, scales as l^{-1}, or a 10 times reduction in size of the device results in the same reduction in heat transfer.

Heat conduction or heat flux, q, is the heat flow per unit area and time, that is, $q = Q/At$ or, according to Fourier's Law in 1D Cartesian coordinates:

$$q = -\frac{\kappa}{L}(T_h - T_c) = -\kappa \frac{dT}{dx} \tag{6.56}$$

Conservation of energy for a small control volume within a larger structure requires that the stored energy in that volume (E_{st}) can be written as the sum of energy in (E_{in}) minus energy out (E_{out}) and the energy generated or lost in it ($E_{internal\ conversion}$):

$$E_{st} = E_{in} - E_{out} + E_{internal\ conversion}$$

$$\rho c_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(\kappa \frac{\partial T}{\partial x}\right) + q_m \tag{6.57}$$

where c_p is the heat capacity under constant pressure, ρ is the mass density, and q_m is the rate of internal energy conversion ("heat generation or source" or "heat sink"). The energy storage term (E_{st}) represents the rate of change of thermal energy stored in the matter in the absence of a phase change. The internal energy conversion term ($E_{internal\ conversion}$) in the case it acts as a source might be the result of chemical, electrical, or nuclear energy. This heat diffusion equation assumes there is no heat loss through the system boundary. The solution to this differential equation provides the temperature distribution in a stationary medium.

More generally, Fourier's Law of Heat Conduction may be written in vector notation as:

$$\mathbf{q} = -\kappa \nabla T \tag{6.58}$$

This equation determines the heat flux vector **q** for a given temperature profile T and thermal conductivity κ (a 3×3 tensor).

Just as we introduced kinematic viscosity in Equation 6.4, it is useful to introduce thermal diffusivity λ_d, in m^2/s, which is related to thermal conductivity via:

$$\lambda_d = \frac{\kappa}{\rho c_p} \tag{6.59}$$

where ρ and c_p are, respectively, the mass density and the specific heat under constant pressure. Thermal diffusivity is a measure of how fast heat conducts. Its value is 2.24×10^{-5} m^2/s for air and 10^{-7} m^2/s for water.

We can now apply Fourier's Law to the generalized version of Equation 6.57, assuming κ is constant, and introduce the thermal diffusivity to yield the generalized heat diffusion equation as:

$$\frac{1}{\lambda_d}\frac{\partial T}{\partial t} = \nabla^2 T + \frac{q_m}{\kappa} \tag{6.60}$$

A common example for heat generation is the resistance heating of an electrical conductor. From Equation 6.60, scaling of the rate of heat conduction is measured by λ_d, and from Equation 6.59, we observe that $\lambda_d \sim \rho^{-1}$ and, since density ~ volume ~ l^3, this leads to the simple conclusion that a 10 times reduction in size results in 1000 times faster heat exchange.

Thermal diffusivity and kinematic viscosity (defined as the ratio of viscosity and density) have the same dimensions, and their ratio defines the important dimensionless Prandtl number:

$$\mathrm{Pr} = \frac{\nu}{\lambda_d} \tag{6.61}$$

The interpretation of this number is the ratio of the fluid velocity boundary layer thickness to the fluid temperature boundary layer. This fluid property enables one to choose an optimum thermal transfer medium. For water, Pr is 10, and for air it is 0.71. For the most efficient heat transfer between two components at different temperatures, while minimizing viscous energy losses, one chooses media with a $\mathrm{Pr} \lll 1$. This can be accomplished, for example, by using liquid metals ($\mathrm{Pr} = 10^{-3}$ to 10^{-1}). On the other hand, for lubrication of parts at different temperatures and keeping low thermal transfer between them, one chooses fluidic media with a $\mathrm{Pr} \ggg 1$ such as silicone oils ($\mathrm{Pr} = 10$ to 10^7).

Convection: Newton's Cooling Law

Fourier's Law, as explained above, describes heat transport in solids through conduction. Convection heat transfer is the mode of energy transfer between a solid surface and an adjacent liquid or gas that is in motion, and it involves the combined effects of conduction and fluid motion. In Figure 6.28 we show velocity and temperature variation of air flowing over a hot substrate and illustrate the difference between forced and natural convection.

In Newton's Cooling Law, the heat flux between a solid surface at temperatures T_s and the bulk temperature of a fluid, T_∞, is proportional to the temperature difference, the exposed area A, and the convective heat transfer coefficient h:

$$q = hA\left(T_s - T_\infty\right) \tag{6.62}$$

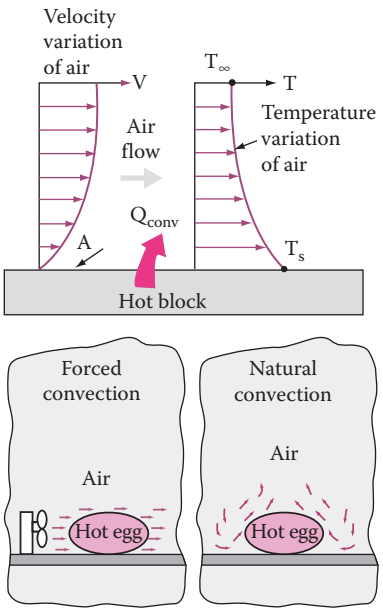


FIGURE 6.28 Velocity and temperature variation of air flowing over a hot substrate, and the difference between forced and natural convection.

The units of h are W/m^{2}K. For most experiments, h is related to the fluid velocity u, or h ~ u, which in turn is proportional to l^{-1}. As in the case of conductive heat transfer, 10 times reduction in size will give 10 times reduction in heat transfer. It is important to keep in mind that the energy exchange at the solid/fluid interface is by conduction and that this energy is then convected away by fluid flow.

Convective heat transfer also can be discussed in terms of the Nusselt number (William Nusselt, 1882–1957), the dimensionless heat transfer coefficient:

$$Nu = hL/\kappa \tag{6.63}$$

with L (in W/mK) the characteristic length and κ the thermal conductivity of the solid, so this number may be viewed as the ratio of convection to conduction for a layer of fluid. With Nu = 1, we have pure conduction, and higher values mean that heat transfer is enhanced by convection. In a boundary layer situation, L is the thickness of the boundary layer. To determine a numerical value for Nu in a forced convection situation, one uses the relationship $Nu = \alpha (Re)^{\beta} (Pr)^{\gamma}$, and for a free convection at low velocity $Nu = \alpha(Re)^{\beta}(Pr)^{\gamma}(Gr)^{\delta}$. The parameters α, β, γ, and δ in this last expression are determined by the experimental setup. From the dimensionless numbers—the Reynolds number (Re), the Prandtl number (Pr), and the Grashof number (Gr)—Re has the dominant effect on h. The Grashof number Gr is defined as:

$$Gr = \frac{L^3 \rho^2 g \beta \Delta T}{\eta} \tag{6.64}$$

with L the characteristic length (m), ρ the density (kg/m^3), g the acceleration resulting from gravity (9.080665 m/s^2), β the coefficient of volumetric expansion (K^{-1}), ΔT the temperature difference between the wall and the surrounding bulk (°C), and η is the viscosity (Pa·s). The Grashof number is a ratio between the buoyancy forces and viscous forces.

Typical values of the convection heat transfer coefficient h (W/m^{2}K) are shown in Table 6.2.

Radiation

The third mode of heat transmission is the result of electromagnetic wave propagation, which occurs both in fluidic media and in an absolute vacuum (Figure 6.29).

TABLE 6.2 Typical Values of the Convection Heat Transfer Coefficient h [W/m^{2}K]

Free convection	**h**
Gases	2–25
Liquids	50–1,000
Forced convection	
Gases	25–250
Liquids	50–20,000
Convection with phase change	
Boiling or condensation	2,500–100,000

Radiation is the energy emitted by matter that is at a finite temperature. Radiant heat transfer is proportional to the fourth power of the absolute temperature, whereas we learned that conduction and convection are linearly proportional to temperature differences. The fundamental Stefan-Boltzmann's law $E = \sigma T^4$ (Chapter 3):

$$E = 5.67 \times 10^{-8} T^4 \tag{3.71}$$

where T is the absolute temperature, and the constant is independent of surface, medium, and temperature. The ideal emitter, or blackbody, is one that gives off radiant energy according to Equation 3.71. All other surfaces emit somewhat less than this amount, and the thermal emission from many surfaces (gray bodies) can be represented by:

$$E = \varepsilon\ 5.67 \times 10^{-8} T^4 \tag{6.65}$$

where ε, the emissivity of the surface, ranges from zero to one.

Example 6.4: Thermal Problem in Miniaturization

Let us apply the insights gathered so far to a simple practical problem, i.e., the calculation of

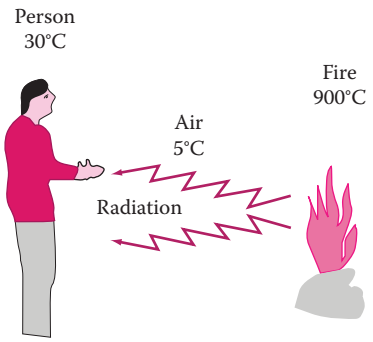


FIGURE 6.29 Example of radiation heat.

the thermal time constant associated with the sudden transfer of an object from one temperature to another. The heat flow into a material of volume V_0, at a uniform temperature T_0, through boundary surface A, when the volume is suddenly immersed in an environment of temperature T_∞, is equated to the rate of increase of the internal energy of the material in volume V_0, or:

$$q = hA\left(T_0 - T_\infty\right) = -\rho c_p V_0 \frac{dT}{dt} \qquad (6.66)$$

Solving this differential equation gives us the temperature of the solid as a function of time:

$$\frac{T - T_\infty}{T_0 - T_\infty} = e^{-\left(\frac{hA}{\rho c_p V_0}\right)t} \qquad (6.67)$$

The term on the left represents the dimensionless temperature, and it can be verified that the temperature in the solid decays exponentially with time and that the shape of the curve is determined by the time constant τ_c given by:

$$\tau_c = \frac{\rho c_p V_0}{hA} \qquad (6.68)$$

To develop a feeling for the above expression, we compare how fast a Ni rod of 0.5-in. diameter and one with a 50-μm radius come to equilibrium with their surroundings. For a cylinder, with $V_0 = \pi r^2 l$ and $A = 2\pi rl$, the above equation transforms to:

$$\tau_c = \frac{\rho c_p r}{2h} \qquad (6.69)$$

Assuming $h = 10\ W/m^2K$, $\rho = 8900\ kg\ m^{-3}$, and $c_p = 0.44\ J\ kg^{-1}/K$ gives a $\tau_c = 1.24$ s for the bigger cylinder. If we assume equilibration with the surrounding temperature, T_∞, in 4 τ_c, equilibration is reached in 5 s. For the 50-μm Ni rod, equilibrium is reached after 0.039 s.

Biot's Number

When a solid body is placed in a surrounding of a different temperature, the body heats up or cools down, and internal temperature gradients that cause thermal stresses are set up. The dimensionless number that characterizes the magnitude of the established thermal gradient is the Biot number (Bi), the ratio of the average surface heat/transfer coefficient for convection from the entire surface (h), to the conductivity of the solid (κ), for a characteristic dimension L, or:

$$Bi = \frac{hL}{\kappa} \qquad (6.70)$$

As L becomes smaller, the Biot number goes down. For Bi << 1, internal temperature gradients become small, and the body can be treated as having a uniform temperature. Consequently, with small devices there is less worry about thermal stresses induced by thermal gradients, and systems with larger internal heat generation capacity by volume can be built.[7] A representative application of the described effect is the micromachined planar Taguchi gas sensor shown in Volume II, Figure 9.11. Not only is the power required to operate at 350°C reduced by more than a factor of 4 (225 mW instead of 1 W), it also can be heated very quickly to high temperatures without breaking.[8]

With small things, we can heat and cool down many times a second (~20 Hz in the example of the 50-μm Ni rod) while thermal stress is minimized. This finds its application in fast thermal detectors such as thermocouples, high-performance heat sinks for high-density microelectronic devices, and fast actuators such as thermopneumatic valves in μ-PCR chambers, fast gas sensors, etc.

Reynolds Number, Friction Factor, and Hydrodynamic and Thermal Boundary Layers

Introduction

Three important concepts need some further dusting off before we comfortably can analyze fluidics in the microdomain and nanodomain, i.e., the Reynolds number, the friction factor, and the hydrodynamic and thermal boundary layers. The hydrodynamic boundary layer is compared with the electrical double layer and Nernst diffusion layer in Chapter 7.

Reynolds Number

The Reynolds number introduced in Equation 6.23 characterizes fluid flow as laminar or turbulent and includes fluid properties ρ and η, the characteristic length L, and average velocity V:

$$Re = \frac{\rho VL}{\eta} \qquad (6.71)$$

Characteristic length, as the term implies, is the length that best represents the body under consideration. In a capillary, for example, the diameter d, where it is called the hydraulic diameter (in m), is far more important in determining the nature of the flow than the total length of the capillary, L_t. The Reynolds number Re characterizes a liquid flow within a conduit of characteristic length L or around an object of characteristic length L. It may be viewed as the ratio of shear stress resulting from turbulence (i.e., inertial forces or forces set up by acceleration or deceleration of the fluid) to shear stress caused by viscosity. Similarly, it can be viewed as the ratio of convective forces over diffusional forces. For Re <<< 1, viscous and diffusional forces dominate, whereas for Re >>> 1, convective and inertial forces do. Similarity of flow will always be maintained if the Reynolds number has an identical value for the bodies being considered.[6] In laminar flow the resistance, caused by intermolecular viscous forces, is proportional to velocity, with turbulence the resistance, generated along the channel perimeter and related to channel shape, particle size, and concentration is proportional to the square of the velocity. Laminar flow is characterized by smooth flow of fluid in layers that do not mix: such flow also is called streamline flow. Turbulent flow is characterized by eddies and swirls that mix layers of fluid together. At sufficiently high velocity, a fluid flow can change from streamline to turbulent flow. The onset of turbulence can be found from the Reynolds number. If Re = 2000 or below, the flow is streamline; if 2000 < Re < 3000, the flow is unstable; and if Re = 3000 or above, the flow is turbulent.

Example 6.5:

Equation 6.71 may be used to calculate the Reynolds number for the flow of air ($\eta = 1 \times 10^{-6}$ m^2s^{-1}) and water ($\eta = 15 \times 10^{-6}$ m^2s^{-1}) in a 50-μm-diameter capillary. Assuming a velocity $V = 500 \times 10^{-6}$ ms^{-1} (10 times the diameter of the pipe per second), we obtain 0.0016 for air and 0.025 for water. Both are considerably less than 1, indicating that if fluids behave as a continuum, as is assumed in classical fluids, microsystems will operate in a viscous-dominated Stokes regime.

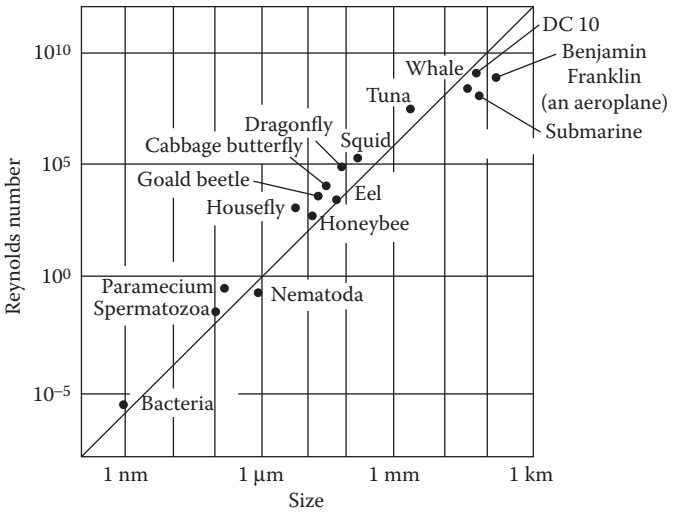


FIGURE 6.30 Reynolds number for mobile machines in their fluidic environment as a function of their size.

A tiny rotor blade in an air-filled, 10-μm-wide capillary will only be able to rotate in milliseconds rather than microseconds. The consideration of moving small things in water is even more daunting, because water is like syrup to the submerged micromachine. Fluid systems in the microdomain are damped heavily and exhibit slow response times. In Figure 6.30, the Reynolds number for mobile machines in their fluidic environment is plotted as a function of size (length), and Table 6.3 lists some additional example Reynolds numbers. Bacteria must be seen as truly autonomous micromachines. Swimming in an aqueous environment, such small vessels are up against the tide with an Re of 10^{-5} (see Table 6.3). Mechanical engineers have avoided designing such machines, but that is exactly how nature has done it, and bacteria are the most abundant life form on Earth.

Friction Factor

We touched upon frictional resistance when discussing a falling sphere and when deriving pipe flow. When a fluid flows through a fluidic channel it experiences frictional resistance, and as a consequence, useful mechanical energy is degraded into unusable heat energy. Knowledge of frictional characteristics is vital for the design of any fluidic system. For a tube of circular cross-section, friction loss or head loss can be determined in terms of shear stress at the wall, τ:

$$\tau = \frac{\Delta p L}{41} \tag{6.72}$$

TABLE 6.3 Example Reynolds Numbers and Hydrodynamic Layer Thickness

Description	Re	
Earth's tectonic plates (100 km at 10^{-8} m/s, kinematic viscosity is 10^{20} m^2/s)	10^{-23}	
Glacier	10^{-11}	
Bacteria in water (1 μm at 10 μm/s and kinematic viscosity ν in water = 10^{-6} ma St)	10^{-5}	
Sperm cells in semen	10^{-3}	
A small marble falling in honey	10^{-2}	
Fish in a tropical aquarium	10^{2}	
Human swimmer	10^{5}	
Bird	10^{6}	
Car (1.5 m long at 90 km/h, kinematic viscosity ν in air = 15×10^{-6} ma St)	2.5×10^{6}	Size of δ, hydrodynamic boundary layer (mm)
		~0.9
Airplane taking off at 300 km/h, 9-m wingspan	50×10^{6}	1.2
Pleasure boat (10-m long at 21.6 km/h or 1,166 knots, and kinematic viscosity υ in water = 10^{-6} ma St)	60×10^{6}	1.3
Large fish in the ocean	10^{8}	

where l is the length of the tube in meters and L is the internal diameter of the pipe.

With reference to Figure 6.31, Equation 6.72 is easily derived from the fact that the force balance is given by:

$$p\frac{\pi L^2}{4} = (p - \Delta p)\frac{\pi L^2}{4} + \tau\pi Ll \quad \text{or}$$
$$\tau = \frac{\Delta pL}{4l}$$

The shear stress, τ, at the wall is a function of five independent parameters: $\tau = f(\eta, \rho, V, L, k)$. The surface roughness of the pipe is given as k and can be described as a length. With five independent variables, the solution is very complex. However, the complexity can be reduced by use of only two nondimensional variables. The nondimensional shear stress, called the friction coefficient or Fanning friction factor, C_D, can be tabulated as a function of the Reynolds number and surface roughness k only. The Fanning friction factor C_D is defined as $2\tau/\rho V^2$ and

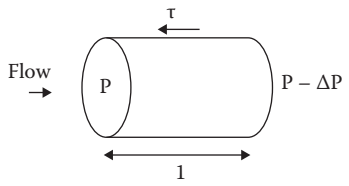


FIGURE 6.31 Friction in pipe flow ($\tau = F_D/A$).

it is used mostly for heat transfer problems in pipe flow. In other words (see also Volume III, Equation 7.27 with $\tau = F_D/A$):

$$C_D = \frac{F_D}{\frac{1}{2}\rho V^2 A} = \frac{\tau}{\frac{\rho V^2}{2}} = f\left(\frac{\rho VL}{\eta}, \frac{k}{L}\right) \qquad (6.73)$$

with k/L the relative surface roughness. Civil engineers replace the Fanning friction factor frequently with the Moody friction factor, C_M, related to C_D as:

$$C_M = 4C_D = \frac{8\tau}{\rho V^2} \qquad (6.74)$$

Yet differently, Coulson and Richardson define the friction factor as $\tau/\rho V^2$.[9] The friction factor for pipe flow of all incompressible Newtonian fluids is a unique function of the Reynolds number. This is a very important insight. Only two groups of variables need to be studied experimentally to obtain a relation that is universally valid for a wide class of fluids, geometries, and flow parameters. For a smooth pipe, Equation 6.73 reduces to $C_D = f(Re)$ only. In laminar flow, we calculate from Equation 6.54 that the pressure drop, Δp, is given as:

$$\Delta p = -\frac{Q8\eta l}{\pi R^2}\left(\text{with } L = 2R \text{ and } V = \frac{Q}{\pi R^2}\right)$$
$$= -\frac{32\eta lV}{L^2} \qquad (6.75)$$

Q is the volumetric flow rate, and from the balance of forces (see Equation 6.72):

$$\tau = \frac{\Delta pL}{4l} = \frac{8\eta V}{L} \qquad (6.76)$$

For macroscopic laminar flow in long tubes, the friction factor, C_D, thus correlates to the Reynolds number as:

$$C_D = \frac{\tau}{\frac{\rho V^2}{2}} = f\left(\frac{\rho VL}{\eta}\right) = 16\frac{\eta}{\rho VL} = \frac{16}{R_e} \qquad (6.77)$$

where C_D is a constant (for Re < 2200), also called the *Poiseuille number*, that depends on the cross-sectional shape but typically has a value of 16 for tube flow (see linear part of curve in Figure 6.32). Compare this with a Fanning friction factor for a sphere of 24 (see Figure 6.16). In case the Moody friction factor, C_M, for pipe flow is plotted, the slope is 64 instead.

The friction factor can be regarded as the ratio of the net imposed external force to the inertial force. Figure 6.32 represents Re versus friction factor C_D for different pipe sizes, relative pipe roughness, densities, and a broad range of viscosities.[1] Data from all pipe sizes and the entire viscosity range overlap and cover more than four decades in Re. At Reynolds numbers below 2200, the viscous shear effects are so large they balance the driving effect of the pressure, and a laminar, creeping, or Stokes flow condition results. For this type of flow, acceleration is negligible. Stokes flow is obtained either by a slow flow, that is, V is small, L is small, or η is large. It is in this regime that Equation 6.54, the Hagen-Poiseuille equation, is valid. At large enough Reynolds numbers, above 4000, a laminar flow becomes unstable and eventually turns turbulent. The region between 2100 and 4000 is a transition region, where transition depends on the size and frequency of disturbances inherent to the type of flow, the roughness of

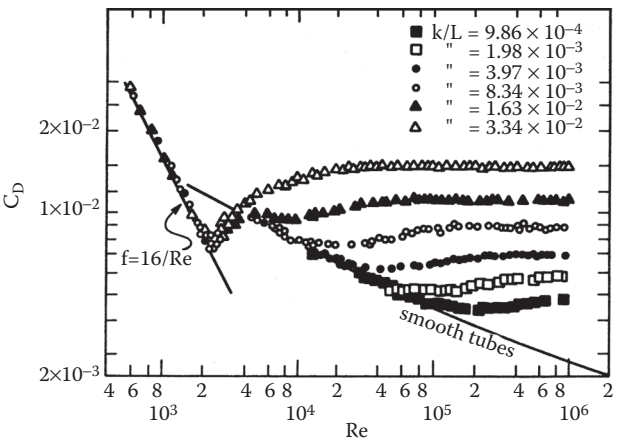


FIGURE 6.32 The Reynolds number, Re, is plotted versus friction factor, C_D, for different pipe sizes, densities, and a broad range of viscosities. Data from all pipe sizes and the entire viscosity range overlap. Different relative roughness factors *k/L* (caused by sand-roughened pipes) cause different onsets for turbulence. (From Denn, M. M. 1980. *Process fluid mechanics.* Englewood Cliffs, NJ: Prentice Hall. With permission.)

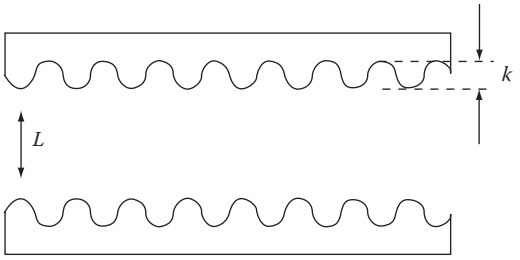


FIGURE 6.33 Schematic of a pipe wall of uniform, regular roughness.

the boundaries, temperature, boundary flexibility, etc. For turbulent flow, the friction factor/Reynolds number dependence is weaker, but no simple theory exists and we must rely on experimental results. An empirical equation for turbulent flow is $f = 0.079\ Re^{-1/4}$.[1,5] The onset of turbulent flow, as is obvious from Figure 6.32, is related to the relative roughness of the pipe wall. As illustrated in Figure 6.33, the relative roughness is represented by the dimensionless group k/L. The larger that ratio, the earlier turbulence starts. For micromachined tubes, one might expect a large k/L and turbulence to kick in quickly. But some simple calculations show that the Reynolds numbers are so low that turbulence is not likely to be observed in microsystems even with a large k/L.

Boundary Layers

The primary resistance to heat transfer and mixing by convection is controlled within a thin layer of stagnant fluid adjacent to the immersed body. In this hydrodynamic boundary layer, the flow velocity, V, varies from zero at the surface, that is, the no-slip condition, to the value in the bulk of the fluid, that is, V_∞. A quantity δ, the upper limit of the boundary layer, corresponds to the layer thickness when V has reached 0.99 V_∞. For laminar flow around an object of length L, δ is given by:

$$\delta \approx \frac{L}{\sqrt{\frac{VL}{\nu}}} \approx \frac{L}{\sqrt{Re}} \tag{6.78}$$

As the Reynolds number gets larger, the boundary layer gets thinner. From Figure 6.34 we can deduce how Equation 6.78 comes about. From the continuity equation it follows that:

$$\frac{\partial v}{\partial y} = -\frac{\partial u}{\partial x} \tag{6.79}$$

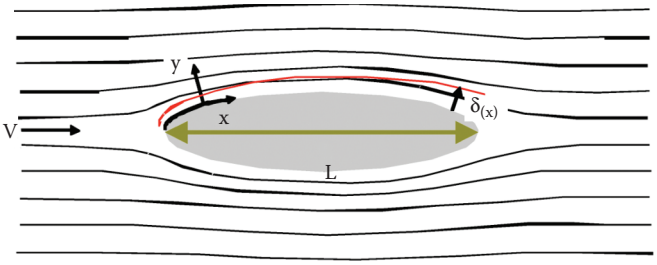


FIGURE 6.34 Boundary layer dimensions.

The scaling of the term, $\partial u/\partial x$, goes as V/L. If the length scale in the direction normal to the body is denoted as δ, i.e., the boundary layer thickness, and the normal velocity scale is denoted as U, it follows from the estimate $\partial v/\partial y \sim U/\delta$ that:

$$U \sim \frac{\delta}{L} V \tag{6.80}$$

The stream-wise equation of momentum balance, with p the pressure, and the scaling of each term annotated, is then given as:

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + \nu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

$$\frac{V^2}{L}, \quad \frac{V^2}{L}, \quad \frac{V^2}{L}, \quad \nu\frac{V}{L^2}, \quad \nu\frac{V}{\delta^2}$$

multiply by L/V^2 to obtain:

$$1, \quad 1, \quad 1, \quad \frac{\nu}{VL}, \left(\frac{L}{\delta}\right)^2 \frac{\nu}{VL} \quad \text{or also,}$$

$$1, \quad 1, \quad 1, \quad \frac{1}{Re}, \left(\frac{L}{\delta}\right)^2 \frac{1}{Re} \tag{6.81}$$

The case of interest here is that of a large Reynolds number (but not so large as to cause turbulence), so that $1/Re \ll 1$. The viscous terms in the Navier-Stokes equation in general are small compared with the other terms when the Reynolds number is large. However, this cannot be true everywhere; there must be at least some small region where viscosity is powerful enough to bring the tangential velocity to zero at the boundary. The small zone where viscosity remains important, even at larger Re, is in the boundary layer, δ. This means that in Equation 6.81 the last term remains, and we obtain Equation 6.78:

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{1}{\rho}\frac{\partial p_d}{\partial x} + \nu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \text{ or}$$

$$\left(\frac{L}{\delta}\right)^2 \frac{\nu}{VL} \sim 1 \text{ or } \delta \sim \frac{L}{\sqrt{Re}}$$

Example 6.6

1. For a car being driven at 90 km/h (with $\nu = \pm 15 \times 10^{-6}$ ma St), the thickness of the hydrodynamic boundary layer is about 1 mm (see Table 6.3 for other examples).
2. In Volume II, Chapter 7, in the section on chemical vapor deposition (CVD), we calculated the average boundary layer thickness over a substrate on which a thin film is deposited, using the boundary layer model from Prandtl (Volume II, Equation 7.13) and obtained:

$$\delta = \frac{2L}{3\sqrt{Re}} \qquad \text{(Volume II, Equation 7.13)}$$

When a fluid at one temperature flows along a solid surface that is at another temperature, a thermal boundary layer develops in addition to the hydrodynamic boundary layer. The thickness of the thermal boundary of a solid object in a fluid medium depends on the velocity of the fluid over the solid surface and the properties of the fluid medium. The equation to estimate the thermal boundary layer thickness, δ_T, is given by:

$$\delta_T \approx \frac{L}{\sqrt{\frac{VL}{\lambda_d}}} \approx \frac{L}{\sqrt{Pe}} \tag{6.82}$$

where λ_d is the mean free path. This equation is equivalent to Equation 6.78 for the estimation of hydrodynamic boundary layer thickness, where the Péclet number ($Pe = VL/\lambda_d$) replaces the Reynolds number. The Péclet number represents the ratio of convection over diffusion forces. The higher the Péclet number, the more the influence of flow dominates over molecular diffusion. Temperature and velocity profiles are identical when the Prandtl number equals 1,

as then Pe and Re are identical. This is approximately the case for most gases ($0.6 < Pr < 1.0$). The Prandtl number for liquids, however, varies widely, ranging from very large values for viscous oils to very small values (on the order of 0.01) for liquid metals. One can deduce that $\delta/\delta_T \sim Pr^{1/2}$. In air (Pr = 0.71) the hydrodynamic boundary layer is thinner than the thermal boundary layer ($\delta = 0.84\ \delta_T$). In water (Pr = 10) the situation is reversed; the hydrodynamic boundary layer is thicker than the thermal one ($\delta = 3.12\ \delta_T$).

In Chapter 7 we learn that the relation between δ_d, the thickness of a steady-state diffusion layer, and δ_h, the thickness of the stagnant hydrodynamic layer on an electrode in a stirred solution, is given by $\delta_d = (D/\nu_k)^{1/2}\delta_h$. For $D = 10^{-5}$ cm^2 s^{-1} and $\nu_k = 10^{-2}$ cm^2 s^{-1}, this leads to $\delta_d = 0.1\ \delta_h$ (see note on page 581 "Nonlinear Diffusion Effects on Ultramicroelectrodes"). Obviously, the thickness of the diffusion boundary layer is considerably smaller than that of the hydrodynamic boundary layer. To prevent random convective motion from affecting transport to and from the sensing electrode, one wants the diffusion layer thickness to be smaller than the hydrodynamic boundary layer and δ_h to be regular. The latter explains the use of rotating electrodes, stirring bars, etc. The diffusion layer is in turn considerably thicker than the electrical double layer at an electrode surface.

Breakdown of Continuum Theory in Fluidics

Introduction

Microfluidics is the control of small volumes of liquids (fL–μL, 10^{-15}–10^{-6} L). In a short period of time, analytical chemists moved from flasks, to test tubes, to microfluidic chips and learned about some of the amazing physics in microfluidics (Table 6.4). Whenever an application demands reproducible handling of small amounts of liquid, microfluidics comes in handy to reduce liquid consumption, for a faster response, to afford disposable hardware, efficient automation (e.g., several functions on the same chip and parallel systems), reduction of power budget, portable devices, etc. In the micro- and nanodomain, Navier-Stokes equations (NSEs), with no-slip boundary conditions at the fluid interface, often seem to fail, and slip flow, creep flow, rarefaction, viscous dissipation, compressibility, intermolecular forces, etc., start demanding the introduction of molecular-based flow models. The breakdown of macroscale physics in gases readily is recognized, and some progress in molecular-based modeling has been made. For liquids, with molecules much more closely packed, evidence of deviation from macroscale fluidics laws is more recent, and modeling has proven to be much more elusive. For gases, breakdown of continuum theory starts being noticed

Table 6.4 Some Physics in Microfluidics

Channel of 100 μm × 100 μm × 1 cm	**Volume = 0.1 μL**		
Typical protein sample: concentration 0.1 mg/mL; average molecular weight = 50 kDa = 50,000 g/mol	Concentration = 2 μM	Number of molecules in microchannel, 2 pmol = 1.2×10^{11} molecules	
Channel of 100 nm × 100 nm × 1 μm	**Volume = 10 aL**		
Protein solution at 2 μM in nanochannel	Number of moles in nanochannel = $2\ 10^{-22}$	Number of molecules in nanochannel = 12	
Pressure vs. flow rate	A few inches of water gives 100s of μm/s in a typical channel (l = 1 cm, r = 50 μm)		
Reynolds number ($Re = \rho VL/\eta$)	Small, Re < 100, no turbulence, viscous and diffusional forces dominate		
Péclet number ($Pe = VL/D$)	Large, $10 < Pe < 10^6$; slow diffusion		
Diffusion coefficients	Small molecule	D = 500 μm^2/s	25 μm in 1 s
	Protein	D = 70 μm^2/s	8 μm in 1 s
	0.5-μm sphere	D = 0.5 μm^2/s	0.7 μm in 1 s
Surface/volume area	Huge: at small size scales the ratio of molecules in solution to molecules adhering to surfaces goes down and evaporation of droplets goes up		
Slip condition	Slip may occur at the fluid/solid boundary		

typically around 1 µm, whereas for liquids this happens around 10 nm.

Once we understand the behavior of fluids at the micro- and nanoscale better, we might be able to exploit many new phenomena. For example, viscous flow in nanochannels could be used to orient and assemble nanowires, separation without membranes is already possible in so-called diffusion-based *H*-separators, and electro-osmotic pumping results in attractive plug flow profiles for better analytical separation (see below).

The No-Slip and Slip Condition

Introduction

It is at an interface between solid and fluid where macroscopic meets microscopic and where microfluidics turns complex: near the contact, the physical properties of the fluid are not the same as in the bulk, and the influence of nanoscale interactions at the interface extends to macroscopic distances (Figure 6.35; see also van der Waals forces in Chapter 7). The no-slip condition means that at a boundary, fluids have zero velocity relative to the boundary. The fluid velocity is equal to that of the solid boundary. Conceptually, one can think of theoutermost molecule of a fluid sticking to the surfaces past which it flows. Stokes, in "On the effect of the internal friction of fluids on the motion of pendulums" (1851),[10] showed that such no-slip condition led to remarkable agreement with a wide range of experiments, including the capillary tube experiments of Poiseuille (1840)[11] and Hagen (1839).[125] However, the no-slip assumption at the interface is an idealization, assuming moderately strong attractive forces between fluid and wall. At the micro- and nanoscale, new phenomena arise because surface forces, and all types of microeffects, ignored at the macroscale, become important. These include friction, surface tension, air bubbles, liquid evaporation, porosity, osmotic effects, van der Waals forces, electrostatic forces, etc. In the case of slip, a slip length, L_s, must be introduced. Slip length is defined as the distance behind the interface at which the liquid velocity extrapolates to zero (see Figure 6.36a and b). In the case of electro-osmosis, with liquids moving in microchannels because of an electrical field, we will see that the electro-osmosis flow velocity V_{EOF} may be regarded as the slip velocity. Phenomena that can be explained assuming slip include a nonconstant pressure gradient in a long microduct and a flow rate higher than predicted from nonslip boundary conditions, electric currents needed to move micromotors that are a lot higher than expected, and micromachined accelerometers operating at atmospheric pressure that are overdamped.

Slip in Gases

Gas flow in devices with dimensions that are on the order of the mean free path of the gas show significant slip. When the mean free path of a gas, λ, approaches the separation, *L*, of the container plates (conditions easily met in a rarefied gas), the laws of high-pressure behavior begin to err. As the

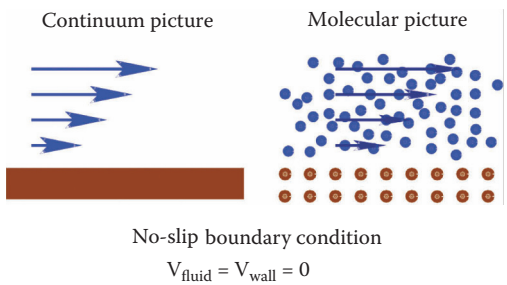


FIGURE 6.35 From continuum theory to the molecular picture at the interface between a solid and a fluid.

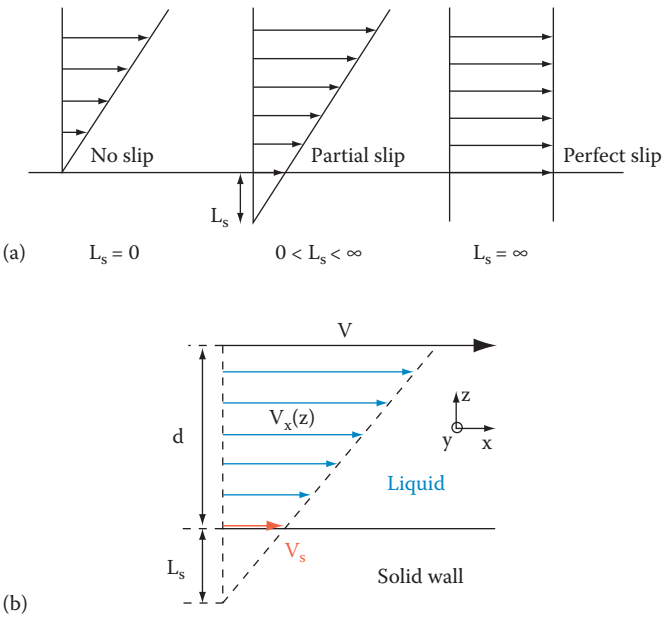


FIGURE 6.36 (a) To slip or not to slip. (b) Determination of slip length L_s.

density is lowered, the gas seems to lose its grip, so to speak, on solid surfaces. In gas flow it begins to slip over the surface, and in the conduction of heat a discontinuity of temperature develops at the boundary of the gas (see below). The same phenomena can be seen at high pressure with fluidic vessels of microdimensions. At the standard atmospheric conditions, the mean free path of air, λ, is about 0.06 μm, and with micrometer-sized channels in air, the no-slip condition does not apply; the friction factor starts to decrease with diameter reduction. The Knudsen number is a dimensionless number given by the ratio of the mean free path of the molecules to the characteristic dimension L:

$$\mathrm{Kn} = \frac{\lambda}{\mathrm{L}} \tag{6.83}$$

which measures the deviation of the state of a gas from continuum behavior. As the value of the Knudsen number increases, noncontinuum effects become more important. If the characteristic dimension of the system, L in this case, is much larger than the mean free path λ, continuum equations apply (Kn ≪ 1). A typical way to determine the conditions of flow is as follows:

For Kn < 0.001, continuum flow regime
For 0.001 < Kn < 0.1, slip flow regime
For 0.1 < Kn < 10, transition regime
For 10 < Kn, free molecular flow regime

These different Knudsen number regimes are summarized in Figure 6.37.

For tube flow (Equation 6.77), the corrected Navier-Stokes result for the friction coefficient is written as:

$$\mathrm{C_D} = \left(\frac{16}{\mathrm{Re}}\right)\left(\frac{1}{1+8\dfrac{\lambda}{\mathrm{L}}}\right) \tag{6.84}$$

For λ of a size approaching L, C_D is approximately 1/9 the value expected from n analysis assuming that the velocity at the wall equals 0. Experiments of flow resistance in 0.5-μm channels by Pfahler et al.[12] show a higher flow rate. This phenomenon is important for gas damping of oscillating microstructures, as they almost always operate in the size regime where λ approaches L.[13] Liu et al.[14] measured the pressure distribution of gaseous flow in microchannels by lining the channels with a series of pressure sensors. This group confirmed the Navier-Stokes equations with slip boundary conditions (first used by Arkilic et al.[15]).

Example 6.7

The macroexpression for the pressure drop through a tube is given in Equation 6.54 with the pressure gradient ~1/R². Deviations in gas flow start showing up already in channels with a hydraulic diameter of 100 μm. This is the case, for example, in a gas chromatography column of that diameter and a length of 1 m (making the length-to-diameter ratio 10,000). Because the channel is very narrow, the viscous effect becomes dominant over the inertial effect. For a long and narrow channel, the pressure and density changes between inlet and outlet are very large. This causes the flow to accelerate near the outlet and leads to a higher mass flow rate than predicted based on incompressible flow. As a consequence, when modeling this so-called compressible creep flow, one needs to include the compressibility effect by incorporating a large change in pressure and density, even though the gas is traveling at a very low subsonic speed.[16]

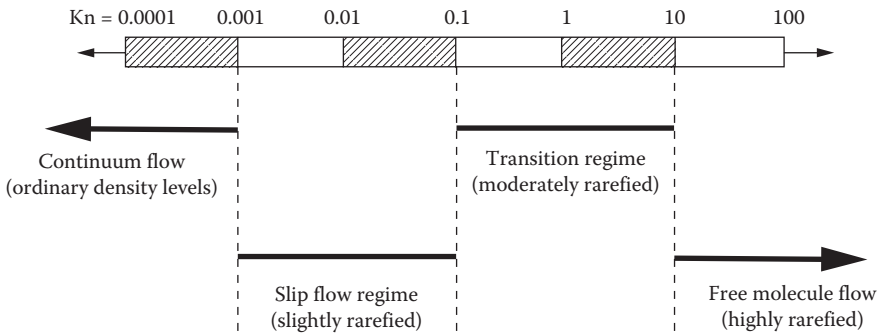


FIGURE 6.37 Knudsen number regimes. (From Gad-el-Hak, M., ed. 2001. *The MEMS handbook*. Boca Raton, FL: CRC Press.)

Corresponding to Equation 6.84, there also is a corrected friction factor for the gas velocity at the surface of a small aerosol particle in the presence of slip (the so-called Cunningham slip correction factor), given as:

$$C_D = \frac{24}{Re\left(1+\frac{2.52\lambda}{L}\right)} \qquad (6.85)$$

This constitutes the corrected Equation 6.30, derived from continuum theory and valid for particles larger than 0.1 μm. For particles less than 0.1 μm, the primary transport mechanism is through Brownian motion and diffusion, with a diffusion coefficient D calculated from Equation 6.29, the Stokes-Einstein expression.

Let us analyze also how slipping of gas molecules at a solid surface influences the effectiveness of heat transport. As fluidic channels become smaller, the effect of boundary layers becomes dominant, and classical theories break down. In the case of a gas, the effectiveness of heat transport from one plate to another is controlled by the mean free path λ of the gas molecules, that is, the average distance before the next collision. The mean free path λ from the kinetic theory of gases is derived as:

$$\lambda = \frac{1}{\sqrt{2}\pi\rho a^2} = \frac{kT}{\sqrt{2}\pi p a^2} \qquad (6.86)$$

where ρ = the density (the number of molecules per unit volume)
T = the temperature
k = the Boltzmann constant
p = the pressure
a = the size of the molecule

In Chapter 3 we calculated the bulk heat conductivity κ of a mole of solid (Equation 3.59); for a gas the same equation holds, except for the additional term for the gas density (ρ):

$$\kappa = \frac{1}{3}\rho c_v v\lambda \qquad \text{(3.59 adapted for gases)}$$

and with the mean velocity v given as $v = \sqrt{\frac{3kT}{m}}$ (3.20)

where c_v = the constant volume-specific heat capacity
λ = mean free path of the molecules
ν = mean velocity, and
m = mass of the gas molecule

From Equation 6.86 we recognize that $\lambda \sim 1/\rho$, so that an increase in density decreases the mean free path with the product ρλ in Equation 3.59 remaining constant. Consequently, κ is not a function of density. We also confirm, from another inspection of Equation 3.59, that κ will increase with increasing temperature. What happens, though, when λ approaches *L*? As the density ρ decreases, no corresponding increase in λ occurs because λ is limited by *L*. In other words, boundary conditions start limiting the average value of λ within the gas, and the thermal conductivity goes down; the linear relation between heat flux and temperature gradient and the no-jump temperature condition at the solid-fluid interface are no longer accurate. A qualitative corrective factor to the heat flux in Equation 6.56, considering the effect on both plates, sometimes is used for small gaps:

$$q = -\frac{\kappa}{L+2g}(T_h - T_c) = -\kappa\frac{dT}{dx} \qquad (6.87)$$

where, for conduction in air, oxygen, nitrogen, carbon dioxide, methane, and helium, experimental g values satisfy $2.4\lambda < g < 2.9\lambda$. The macroscopic prediction for the heat flux exceeds the microscale prediction by less than 5% when $L > 12\lambda$. For hydrogen, however, the experimental value for g is 11.7λ, and the condition $L < 47\lambda$ must be satisfied.[17] Equation 3.59 also determines the boundary for microscale conduction in solids and liquids.

For solid-state thermal conduction in metals Equation 3.59 holds, with λ the electron mean free path, C_v the electron-specific heat, and v the Fermi velocity (see also Chapter 3). In dielectrics, fluids, and semiconductors, λ is the phonon mean free path, C_v the phonon-specific heat, and v the average speed of sound. Experimental g values for solids are not yet available. However, calculations indicate that macroscale exceeds microscale results by less than 5% for $L > 7\lambda$. Consequently, the boundary between the microscale and macroscale

regimes equals $L = 7\lambda$ for the conduction across a layer, where L is the layer thickness and λ the mean free path of the dominant carrier of heat. Given the dimensions of a phonon at least 10 times smaller than λ of a gas at atmospheric pressure, we might not observe deviations from the Navier-Stokes equations in liquids until the dimensions of the containing vessels are in the 100-Å range.

Maps for heat transfer regimes in microstructures, immediately showing whether a certain heat transport in a given device can be analyzed with macroscale theory, have been developed.[17]

Slip in Liquids

Deviations from the Navier-Stokes equations in liquid are more difficult to demonstrate than in a gas. Liquids do not have a well-advanced molecular-based theory as dilute gases; for example, there is no Knudsen number for liquids, and the concept of a mean free path in a liquid is very different than for a gas. Gravesen et al.[13] in 1993 reviewed some of the literature describing deviations from Navier-Stokes equations in liquid flow in narrow channels but could not find enough consistency in the data to conclude that there was a breakdown of the macrotheory. More clarity has emerged since 1993. In controlled experiments, generally with dimensions of microns or smaller, apparent violations of the no-slip boundary condition clearly have been demonstrated in liquids. We usually assume that viscosity in liquids is independent of the dimensions of the flow channel. However, experiments by Harley et al.[18] suggest that in microchannels liquid flow takes on values as much as an order of magnitude greater than that expected from a constant viscosity. Pfahler et al.[12] similarly measured a flow significantly higher than expected from macrotheory when using isopropanol and silicone oil. Israelachvili[19] showed that the apparent viscosity of a liquid is the same as that in the bulk once the film is thicker than a dozen molecular layers (~5 nm); below that thickness the viscosity grows by several orders of magnitude. Tretheway et al.,[20] using a 30×300-μm channel made hydrophobic with a 2.3-nm-thick monolayer of octadecyltrichlorosilane, measured fluid slip velocities within 450 nm of the wall of about 10% of that in the center instead of zero. This slip condition has the effect of increasing the total volume flow. In the case of a hydrophilic surface, velocity profiles remained consistent with Stokes flow and the well-accepted no-slip boundary (see Figure 6.38). Slip of liquids was also measured with hexadecane flowing over an atomically smooth surface, such as mica, that again was made nonwettable by deposition of a hydrophobic monolayer.[21]

From the above, we conclude that slip causes a "shift" in the velocity profile. It also has been established that slip increases with the contact angle,[22] decreases with increasing surface roughness,[22] and increases with viscosity and driving rate.[23]

In flow channels approaching molecular diameters, "molecular" effects turn yet more dramatic. When the channel is less than 10 molecular diameters, of the order of a few nanometers, the fluid loses its liquid-like behavior and assumes some solid-like characteristics.[24] The structure of the molecules changes from a random order to a discrete number of ordered layers, with an "apparent" viscosity of the fluid 10^5 times the "regular" viscosity. The molecular time scale for molecules in these nanotubes slows down by a factor of 10^{10}.[25] In Figure 6.39, the friction force between two parallel plates, separated by a few

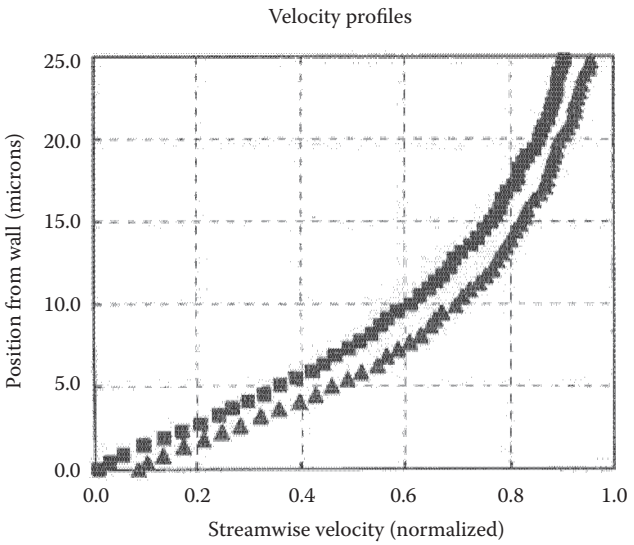


FIGURE 6.38 Measurement of slip length. Square symbols: hydrophilic coating ($L_s = 0$), triangle symbols: hydrophobic coating ($L_s = \sim 1$ μm). The velocity at 450 nm from the wall was measured at 10% from that in the middle of the flow (instead of 0). (Tretheway, D. C., and C. D. Meinhart. 2002. Apparent fluid slip at hydrophobic microchannel walls. *Phys Fluids* 14: L9.)

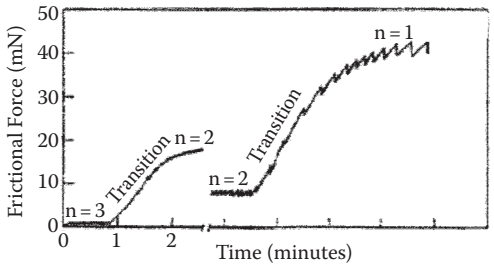


FIGURE 6.39 Friction forces between two parallel plates. (Ho, C. M. 2001. 14th IEEE International Conference MEMS, Interlaken, Switzerland, 375.)

molecular layers of OMCTS (octamethylcyclotetrasiloxane), sliding with respect to each other while the distance between them decreases (horizontal axis), is shown. Two very striking features are apparent. First, the frictional forces increase in a stepwise fashion as the distance between the plates is reduced and settles on a stable value corresponding to an integer number of molecular layers (1, 2, or 3). Second, a saw-tooth behavior is observed, corresponding to a stick-stick motion of the plates when the molecular layers are rearranging themselves while moving with respect to each other. When working with macromolecules such as DNA and proteins in nanochannels, effects as described here are to be expected.

Instead of continuum theory, molecular dynamics methods are used to study interfaces at the molecular level. In molecular dynamics one integrates Newton's equations of motion for each individual particle. A commonly used force between the interacting molecules is the Lennard-Jones force introduced in Chapter 7 (Equation 7.22).

Forces at Interfaces

Introduction

Molecules condense when attractive intermolecular bond energies are comparable with or greater than their thermal (i.e., kinetic) energy. These condensed materials have a surface: a transition region between condensed phase [S(olid) or L(iquid)] and a gas phase. An interface is a region between two condensed phases. At interfaces, atoms and molecules possess energies and reactivities significantly different from those of the same species in the bulk. That imbalance leads to surface free energy. The imbalance of forces at a gas/liquid interface usually is termed the interfacial tension (in units of N/m or also J/m^2), whereas for interfaces involving solids, it is called "surface energy." We start by illustrating the surface forces at the liquid/gas interface, i.e., surface tension. Next we study the solid/liquid interface to elucidate beading of droplets and capillary action.

Surface Tension

Many of the properties of a liquid, including surface tension, capillary action, and beading, are caused by cohesion and/or adhesion of molecules. In liquids, the rearrangement after creating a new interface is very fast, and the surface tension is an equilibrium value. Liquids and gases are separated by a meniscus and differ only in density, not structure. Molecules at the surface of a liquid experience forces around and below, whereas molecules in the middle are attracted from all sides, so the net attraction on molecules at the top is downward (Figure 6.40). Increasing the surface area results in more molecules that are exposed and a higher overall energy, so the surface area tends toward the smallest possible area to reduce the overall energy of the system. A liquid thus arranges itself to have less surface area. Water forms spherical drops because a sphere exhibits less surface to volume.

The energy required to break through the surface of a substance or to spread a drop of the substance out into a film is the surface tension. Fluids behave as though they have a surface composed of an elastic skin that is always in tension. Surface tension, with reference to Figure 6.41, is the amount of energy

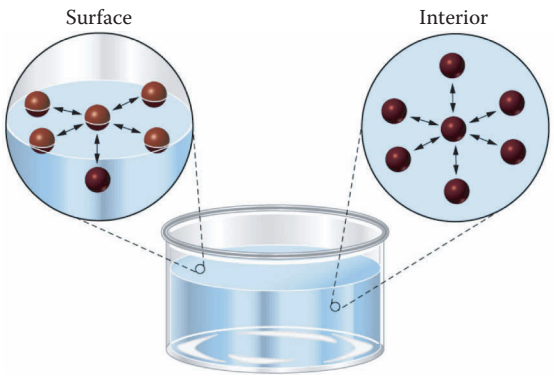


FIGURE 6.40 The molecular origin of surface tension. Molecules at the surface of the liquid experience forces around and below, whereas molecules in the middle are attracted from all sides, so the net attraction on molecules at the top is downward.

required to stretch or increase the surface area of a liquid by a unit area. The force exerted by a planar film supported on a rectangular frame with one movable bar of length l, as shown in Figure 6.41, is given by:

$$F_s = 2\gamma l \tag{6.88}$$

The factor of two is introduced because the liquid film has two surfaces. The work required to create new surface area is then:

$$W = 2\gamma l d \text{ and } 2\gamma = \frac{dW}{dA} \tag{6.89}$$

An interfacial energy γ is associated with any interface between two phases (J m^{-2}) (also called a surface tension: Nm^{-1}). An increase in area increases the system's free energy, which is not thermodynamically favorable. Water has the highest surface tension of all common liquids.

Drops and Bubbles

The forces on a liquid sphere (say, a drop of water) of radius r can be calculated as follows. Balance of forces in Figure 6.42 requires that $\sum F = 0$. So we can write:

$$F_{in} - F_{out} - F_{surface} = 0$$

$$p_{in}\pi r^2 - p_{out}\pi r^2 - \gamma 2\pi r = 0$$

$$(p_{in} - p_{out})\pi r^2 = \gamma 2\pi r$$

$$p_{in} - p_{out} = \frac{2\gamma}{r} \text{ or also } \Delta p = \frac{2\gamma}{r} \tag{6.90}$$

In this example there is only one liquid/gas surface. For a soap bubble, with two liquid/gas surfaces, the force will be twice as great.

Atoms of liquid on the surface of a small droplet are held less tightly compared with atoms on a flat liquid surface. High curvature effectively reduces the coordination number of the surface atoms, making them easier to evaporate. This is manifested as an increase in the vapor pressure p of a liquid droplet compared with that of a bulk liquid p_0 and is described in the Kelvin equation. In the case of a positive curvature, the liquid in drops has a higher vapor pressure than in the bulk. In the case of negative curvature, it has a lower vapor pressure than in the bulk (e.g., liquid in pores). The Kelvin equation for a droplet is given as:

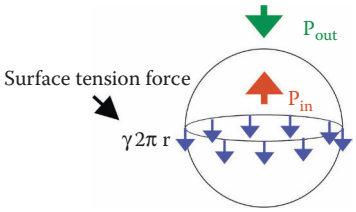


FIGURE 6.42 Pressure difference across a curved surface.

$$\ln\frac{p}{p_0} = \frac{2\gamma\bar{V}}{RTr} \tag{6.91}$$

In this equation:

r = the drop radius
γ = the surface tension
$\bar{V}$ = the molar volume of the liquid
R = the gas constant
T = the absolute temperature

From Equation 6.91 the vapor pressure p relative to the bulk p_0 can be calculated as:

$$p = p_0 \exp\left(\frac{2\gamma\bar{V}}{RTr}\right) \tag{6.92}$$

Evaporation of a droplet of diameter d at time t is given as:

$$d^2 = d_0^2 - \beta t \tag{6.93}$$

where d_0 is the initial diameter, and β is a constant independent of drop size. The time it takes for the drop to disappear is then:

$$t = \frac{d_0^2}{\beta} \tag{6.94}$$

The latter scales as l^2, and clearly miniaturization favors faster evaporation. Fast evaporation of small

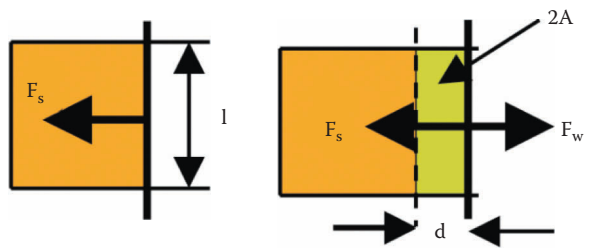


FIGURE 6.41 The work involved in expanding a liquid film by 2A (=2dl).

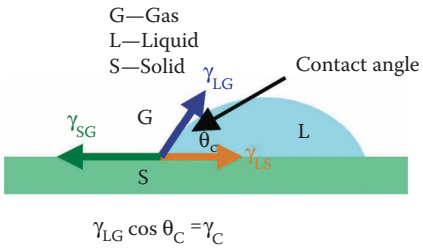


FIGURE 6.43 The horizontal force balance between these three interfacial tensions represents Young's equation.

$$\sum \gamma = 0$$

$$\gamma_{LS} + \gamma_{LG} \cos\theta_c - \gamma_{SG} = 0$$

$$\gamma_{LS} + \gamma_{LG} 2\pi r \cos\theta_c - \gamma_{SG} = 0 \qquad (6.95)$$

$$\frac{\gamma_{LS} - \gamma_{LG}}{2\pi r} + \gamma_{LG} \cos\theta_c = 0$$

$$\frac{-\gamma_{LS} + \gamma_{LG}}{2\pi r} = \gamma_{LG} \cos\theta_c = \gamma_c$$

liquid droplets often poses a considerable problem in MEMS. A droplet ejected from a small nozzle might evaporate before it reaches the substrate, which makes storing small amounts of chemicals on a liquid platform very cumbersome for the same reason.

Wetting and Contact Angle–Beading on Solid Surfaces

We now introduce a solid surface in a liquid and define adhesion forces and wetting of a solid surface. From the previous section we know that increasing the surface area of any liquid requires work, and this effort per unit surface area is the liquid's surface tension (with a unit of N/m and symbol γ). When liquid of a sessile drop comes in contact with both a solid and a gas, as shown in Figure 6.43, a contact angle θ_c of between 0 and 180° is set up. Three interfacial tensions come into play: γ_{SG}, the solid/gas interface tension; γ_{SL}, the solid/liquid interface tension; and γ_{LG}, the liquid/gas interface tension. The horizontal force balance between these three interfacial tensions represents Young's equation:

where θ_c is the contact angle. The critical surface tension γ_c is an intrinsic characteristic of the surface. Liquids with $\gamma < \gamma_c$ completely wet the surface ($\theta = 0°$). Liquids with $\gamma > 90°$ are said to not wet the surface ($\gamma_{LS} > \gamma_{SG}$).

The surface of plant leafs, especially of the lotus flower (Nelumbo nucifera), the symbol of purity in some Asian religions, can show extreme hydrophobicity or superhydrophobicity to water. This makes the deposition of water drops on their surface almost impossible (large water contact angles $\theta > 150°$) (Figure 6.44). Such hydrophobic surfaces also show a remarkable self-cleaning effect. Even after emerging from mud, the lotus leaves do not retain dirt when they unfold. The prospect of self-cleaning surfaces is obviously of significant commercial interest.

Two botanists, Barthlott and Neinhuis, from the University of Bonn, discovered the reason for this self-cleansing effect.[27] The general opinion used to be that the smoother a surface, the less dirt and water adhere to it. However, when inspected with an SEM, lotus leaves show a combination of nano- and microstructures that make for a rough surface at the micro- and nanoscale. The explanation for the resulting effect lies in two physical characteristics of the

FIGURE 6.44 Extreme hydrophobicity of lotus plant leafs. The far right photo represents a commercial application of the Lotus Effect; i.e., a BASF aerosol spray for paper, wood, leather, etc., makes those materials superhydrophobic.

lotus leaf, i.e., the properties of these microstructures repel water, and the nanostructures found on top of the microstructures are made of waxy materials that are poorly wettable. The combination of the chemistry, the nanostructures, and the non-adherence of dirt and water to the micro-rough surface is what Barthlott and Neinhuis named the Lotus Effect. The roughness of a surface improves the wettability for hydrophilic surfaces ($\theta < 90°$). A drop on such a surface will appear to sink into the hydrophilic surface. For hydrophobic surfaces ($\theta > 90°$) the wettability decreases as it gets energetically too expensive to wet a rough hydrophobic surface. The result is an increased water-repellency. Energetically the best configuration for the drop is then on top of the roughness spikes like "a fakir on a bed of nails" (Figure 6.45).

A droplet on an inclined superhydrophobic surface rolls off instead of sliding. When the droplet rolls over a contamination, the particle is picked up from the surface, and the energy of absorption into the rolling droplet is higher than the surface stiction force between the dirt particle and the surface. Usually the force needed to remove a particle is very low because of the minimized contact area between the particle and the surface. As a result, the droplet cleans the leaf by rolling off the surface (Figure 6.46). See also Volume III, Chapter 3, page 144 "Superhydrophobicity and the Lotus Effect".

Capillary Action

The beading, illustrated in Figure 6.44, occurs when a polar substance is placed on a nonpolar surface and cohesive forces compete with adhesive forces. There is also a competition between adhesive and cohesive forces in capillary action. If adhesive forces are greater than cohesive forces, the liquid surface is attracted to

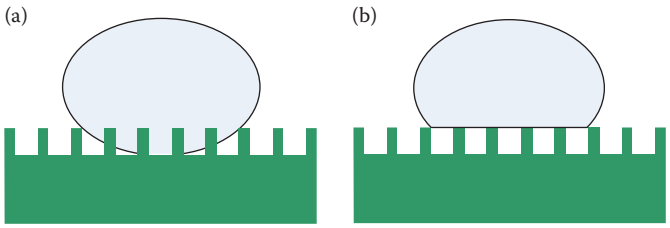


FIGURE 6.45 (a) A droplet on a hydrophilic rough surface seems to sink into the gaps. (b) A droplet on a rough hydrophobic surface sitting on the spikes. (From http://lotus-shower.isunet.edu/the_lotus_effect.htm.)

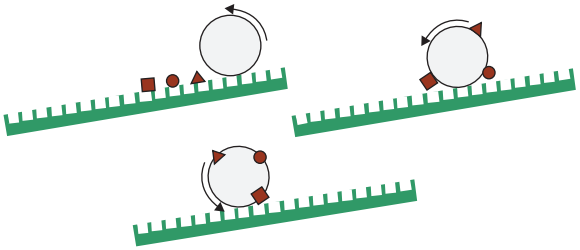


FIGURE 6.46 Self-cleaning behavior of a superhydrophobic surface by droplets rolling off instead of sliding off. (From http://lotus-shower.isunet.edu/the_lotus_effect.htm.)

its container more than the bulk molecules. Therefore, the meniscus is U-shaped (e.g., water in glass) (Figure 6.47Aa left). If cohesive forces are greater than adhesive forces, the meniscus is curved downward (Figure 6.47Aa right). In capillary action, liquids rise spontaneously in a narrow tube (Figure 6.47Ab).

We are all familiar with water being sucked up by a hydrophilic capillary (e.g., in a glass capillary) as a result of positive capillarity. Less evident from daily experience is that water may have to be pushed up in the case of a hydrophobic capillary as a result of negative capillarity (e.g., in a Teflon tube). Figure 6.47B illustrates the two types of capillarity schematically. In the positive capillarity case, liquid in a capillary rises as to seek a lower interfacial energy. The liquid will rise until it lifts a column of liquid balancing the energy reduction gained by rising. The energy reduction by rising over a distance l in the capillary is given by $(\gamma_{SG} - \gamma_{SL})\, 2\pi rl$. In this expression, r is the capillary radius, and the work performed in raising the liquid column is given by $\Delta p \pi r^2 l$, with Δp the pressure difference between the top and bottom of the raised column. By equating these two energies, we can deduce the pressure drop across a liquid/gas interface in a small capillary as:

$$\Delta p = \frac{(\gamma_{SG} - \gamma_{SL})}{r} = \frac{2\sigma_L \cos\theta_c}{r} \qquad (6.96)$$

where it has been assumed, as is generally the case, that $\gamma_{LG} = \sigma_L$ (the surface tension of the liquid). The pressure produced by wetting, calculated from this expression for typical values of the gas/liquid surface tension ($\gamma_{LG} = \sigma_L = 0.073$ N/m at room temperature and 0.058 N/m at 100°C) and a capillary radius of 10 μm, is in the range of 0.01 MPa. This is roughly the same as achieved with piezoelectric pumps, which are much larger in size.

Capillary forces scale with l^1, and as we will see in Volume III, Chapter 7 on scaling, the smaller the exponent of l, which is 1 in this case, the more favorable the scaling. The Kelvin equation for a liquid in a cylinder (say, in a pore of a solid) as shown in Figure 6.47 is, in analogy with Equation 6.91 for a droplet, given as:

$$\ln \frac{p}{p_0} = \frac{\gamma \bar{V}}{RTr} \tag{6.97}$$

In this equation:

r = the cylinder radius
γ = the surface tension
$\bar{V}$ = the molar volume of the liquid
R = the gas constant
T = the absolute temperature

The difference is a factor of 2 because for a liquid column there is only 1 interface.

An air bubble can completely block a microchannel, and Colgate et al.[28] report that the pressure difference, Δp, needed to move a gas bubble in a straight channel is empirically given by:

$$\Delta p = \frac{2\gamma_f}{r} \tag{6.98}$$

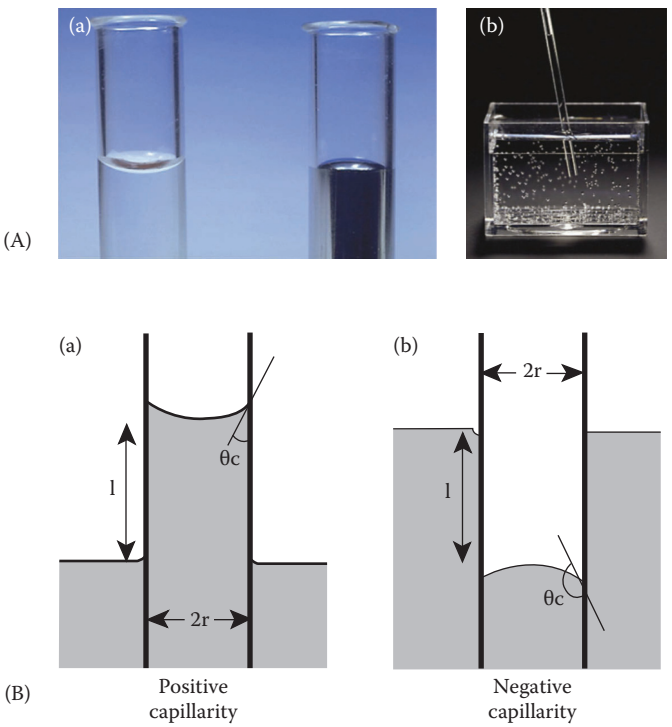


FIGURE 6.47 (A) (a) Positive (left) and negative (right) capillarity; (b) Liquid rises in a capillary. (B) Capillary phenomena: positive (a) and negative (b) capillarity.

Mathematically this equation, not surprisingly, has the same form as the expression derived in Equation 6.90 or a liquid sphere with γ_f now a frictional surface tension. The pressure differences in Equations 6.90, 6.96, and 6.98 all increase with decreasing channel dimension or bubble size.

Mixing, Stirring, and Diffusion in Low Reynolds Number Fluids

The Péclet Number

Stokes or low Reynolds number flow, we saw above, occurs when Re << 1; this might be the case because ρ, V, or L are very small, e.g., in microorganisms, MEMS, particles, bubbles, etc., or it might happen because η is very large, e.g., lava flow, honey, etc. In these regimes viscous forces dominate, and inertial strategies do not help for moving microsystems in a medium (swimming) nor are they any help in pushing microvolumes of fluid (mixing). The question we address in this section is how to mix reactants in a small microreactor in the absence of turbulence? The primary resistance to mixing by convection, we saw earlier, is controlled by a thin layer of stagnant fluid adjacent to a solid surface. In this hydrodynamic boundary layer, the flow velocity, V, varies from zero at the surface, that is, the no-slip condition, to the value in the bulk of the fluid, that is, V_∞. Laminar flow around an object of length L and boundary layer δ is given by:

$$\delta \sim \frac{L}{\sqrt{Re}} \tag{6.78}$$

As the Reynolds number becomes smaller, the boundary layer gets larger. So one appreciates that mixing in the microdomain (< 1 mm) is difficult because flows are predominantly laminar, and mechanical stirring is difficult in restricted fluidic channels. Stated differently: with very low Re numbers, inertial convective mass transport effects (like turbulence) are not feasible. Or also: at the microfluidic level, mixing is like trying to stir syrup into honey, and two liquids, traveling side-by-side through a narrow channel, only become mixed after several centimeters because mass transport in the microdomain is traditionally limited to simple diffusion (Figure 6.48). To decide which transport

FIGURE 6.48 Laminar flows do not mix. Taking advantage of low Reynolds number (Re) and high Péclet number (Pe).

type dominates, diffusion or convection, one must inspect the Péclet number.

The Péclet number represents the ratio of mass transport by convection to mass transport by diffusion (see also Table 6.4):

$$\mathrm{Pe} = \frac{\mathrm{VL}}{\mathrm{D}} \tag{6.99}$$

The higher the Péclet number, the more the influence of flow dominates over molecular diffusion (see also Equation 6.82). In liquids, the diffusion coefficient of a small molecule typically is about 10^{-5} cm^2/s. With a velocity of 1 mm/s, in a channel of 100-µm height, the Péclet number is on the order of 100. This elevated value suggests that the diffusion forces are acting more slowly than the hydrodynamic transport phenomena: for mixing by diffusional forces one must have Pe < 1.

Definitions of Mixing, Stirring, and Chaotic Advection

Mixing means to homogenize a solution and involves convection and diffusion. In mixing, concentration gradients are smoothened until they disappear (Figure 6.49).

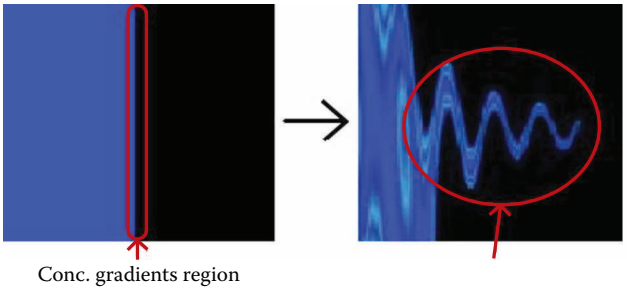


FIGURE 6.50 Stirring is a mechanical process that stretches and folds fluid elements, increasing regions with concentration gradients.

Stirring is convection meant to reduce the mixing time. It involves a mechanical process that stretches and folds fluid elements, increasing regions with concentration gradients (Figure 6.50).

At low Reynolds numbers, stirring is like kneading dough for making bread, with stretching of fluid elements to increase the diffusional interface and folding to decrease distance over which species have to diffuse. Creating chaotic pathlines for dispersing fluid species effectively in smooth and regular flow fields is called "chaotic advection."[30] Chaotic advection results in rapid distortion and elongation of the fluid/fluid interface, increasing the interfacial area across which diffusion occurs, which increases the mean values of the gradients driving diffusion, leading to more rapid mixing. Designing for chaos refers to devising mixer configurations that produce chaotic advection. The concept is illustrated in Figure 6.51.

An example chaotic advection structure is shown in Figure 6.52.

Implementing Micromixers

Introduction

From Table 6.4 we gleaned that, assuming a diffusion constant D of 10^{-5} cm^{2}s^{-1}, a molecule will take 500 s to diffuse over a distance of 1 mm but only 0.5 ms to cross a distance of 1 µm. In the case one is working with a macromolecule such as myosin, the diffusion coefficient, D, becomes very small (1.2×10^{-11} cm^2/s

FIGURE 6.49 Mixing homogenizes concentration gradients (t_a is some intermediate time).

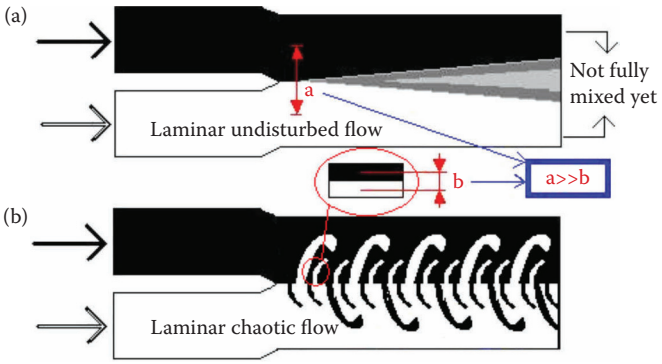


FIGURE 6.51 The concept of chaotic advection. (a) Two flows come together in a small conduit and only limited mixing takes place. (b) Laminar chaotic advection reduces the time it takes to mix in a microchamber.

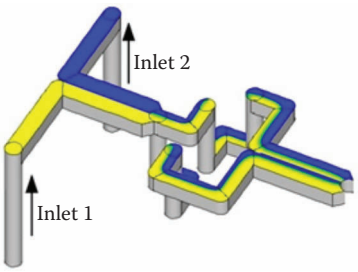


FIGURE 6.52 Mixer configuration for chaotic advection in the microdomain. Two fluids entering the inlet ports laminate at the first horizontal junction, producing two side-by-side fluid streams. Successive vertical separation and horizontal reuniting of fluid streams increase the number of laminates with each stage and, thus, the contact area between the two fluids.

compared with 3×10^{-6} cm^2/s for fluorescein), and one might need to look for alternatives for inducing fast mixing because pure diffusional mixing is now too slow. This is an important, current technological challenge, for example, in the design and manufacture of fast DNA and protein arrays for molecular diagnostics: assays in arrays need to be married with a small, simple fluidic manifold, whereas reaction rates in the microfluidic chambers holding the arrays must remain high, despite low Reynolds numbers and high Péclet numbers.

Passive Micromixers

Several passive micromixer designs aim at reducing the diffusion path length. We start this analysis from the Einstein-Smoluchowski equation:

$$x^2 = 2Dt \qquad (6.100)$$

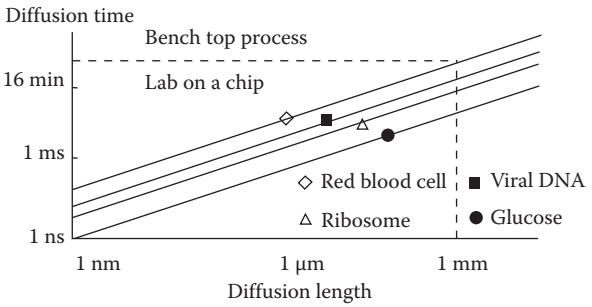


FIGURE 6.53 Diffusion times and lengths in the lab and on a chip.

where D is the diffusion coefficient, x the distance over which a species has diffused, and t is the average time between collisions. From this equation, the diffusional mixing time is of the order x^2/D. Miniaturization, as illustrated in Figures 6.53, can thus be expected to lead to smaller diffusive mixing times (scales with l^2). To help things along, in passive mixer designs the interfacial area is increased to reduce diffusion length. This may be accomplished using sinusoidal, square-wave, or zigzag channels. As illustrated in Figure 6.54, the smaller the distance between regions of different concentration, the larger the diffusive flux. Multi-laminated flow down-scales the diffusion length.

Mixing as illustrated in Figure 6.55 can also be improved on by using a fractal approach, a design akin to how nature uses passive mixing (Figure 6.55). In nature, only the smallest animals rely on diffusion for transport; animals made up of more than a few cells cannot rely on diffusion anymore to move materials within themselves. They augment transport with hearts, blood vessels, pumped lungs, digestive tubes, etc. These distribution networks

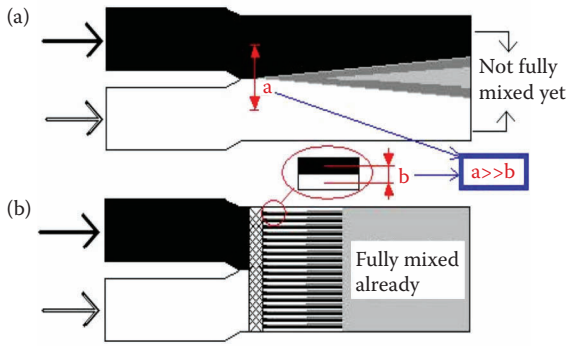


FIGURE 6.54 (a) Two flows come together in a small conduit and only limited mixing takes place. (b) The smaller the distance between regions of different concentration, the larger the diffusive flux. Multi-laminated flow down-scales the diffusion length.

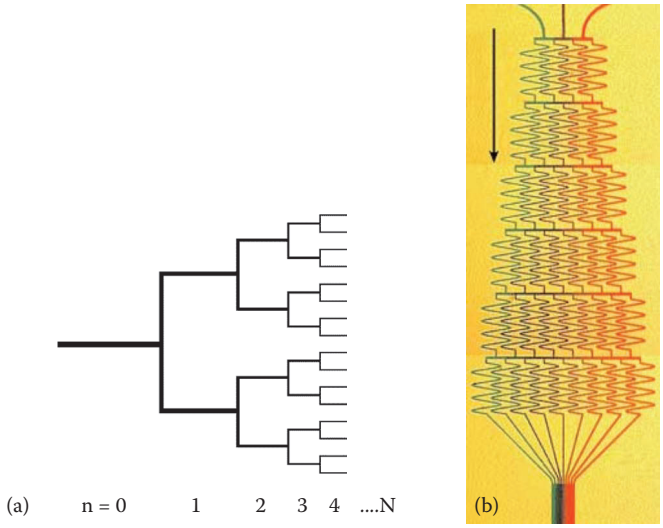


FIGURE 6.55 (a) The branching structures found in mammalian circulatory and respiratory systems have, through natural selection, evolved over many millennia to their current state of minimizing the cost function associated with the amount of biological work required to operate and maintain the system. (b) Biomimetic microfluidic device used for studying concentration gradients in a network of branching serpentines, in which three dyes are injected at the top before combining in a single channel at the bottom. (Whitesides, G. M. 2003. The "right" size in nanobiotechnology. *Nat Biotech* 21:1161–65.)

typically constitute fractals. Fractals are an optimal geometry for minimizing the work lost as a result of the transfer network while maximizing the effective surface area (see Volume III, Chapter 9).

In Figure 6.56 we demonstrate a T-shaped diffuser/mixer, where two fluids A and B enter a T-shaped intersection. Beyond the T, the width of the collection channel is reduced. In the narrowed rectangular cross-section of the fluidic structure, with width w and height b, the diffusive mixing time is proportional to:

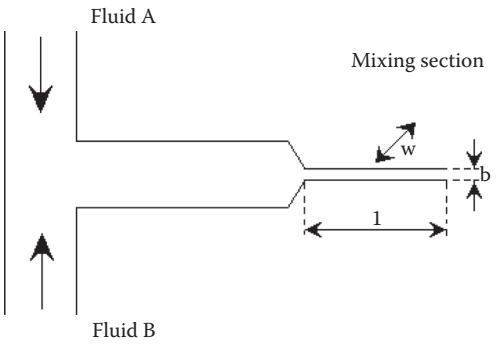


FIGURE 6.56 T-shaped diffuser/mixer, forcing fluids to mix by diffusion in a constriction of the microchannel. The length of the constriction is *l*, the width is *w*, and the height is *b* (unchanged from the wide channel section).

$$\tau \sim \frac{w^2}{D} \tag{6.101}$$

The necessary mixing length of the mixer is estimated by assuming that the two fluids at a velocity of Q/wb, with Q the total flux, require a length L_m to achieve mixing, or:

$$L_m \sim \frac{Qw}{bD} \tag{6.102}$$

At a fixed height b, narrowing the channel width w can reduce the length L_m to achieve mixing in a shorter time.

Two adjoining laminar flows A and B in a T-diffuser/mixer may be used as a filter as we illustrate in Figure 6.57. As the two parallel laminar flows contact, diffusion may extract one compound from one stream into the other. This way, components with higher diffusivity can be extracted from a sample stream (say, A) into a second diluent flow (say, B). This simple system does not require a membrane and is somewhat akin to field flow fractionation. Example applications include competitive immunoassays, determination of diffusion coefficients, and reaction kinetics parameters.

Diffusive extraction is also the underlying principle of the H-filter, illustrated in Figure 6.58. This continuous filter consists of two T-diffuser/mixers: one for mixing and one for separation. In the H-filter a proper design of a microfluidic channel allows for controlled extraction of analytes with high diffusion coefficients from components of the sample with lower diffusion

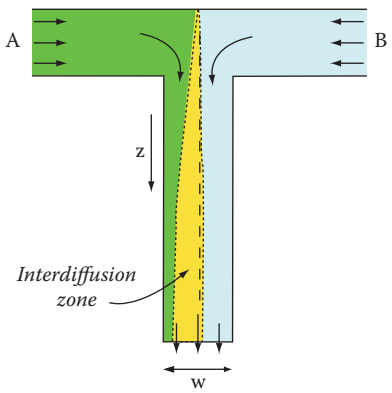


FIGURE 6.57 T-diffuser/mixer used as a filter. (Ismagilov, R. F., A. D. Stroock, P. J. A. Kenis, and G. Whitesides. 2000. Experimental and theoretical scaling laws for transverse diffusive broadening in two-phase laminar flows in microchannels. *Appl Phys Lett* 76:2376–78.)

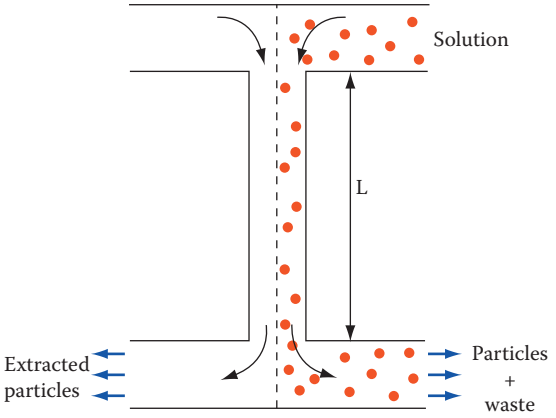


FIGURE 6.58 The H-filter is based on separation by diffusion between laminar flows. (Brody, J. P., P. Yager, R. E. Goldstein, and R. H. Austin. 1996. Biotechnology at low Reynolds numbers. *Biophys J* 71:3430–41.)

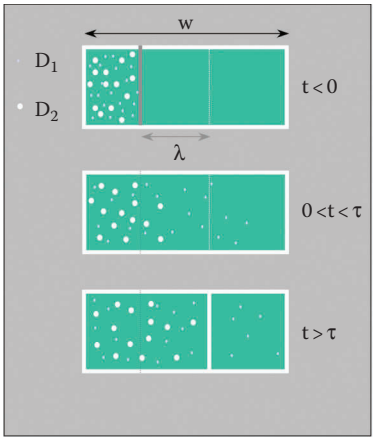


FIGURE 6.59 Design parameters of the H-filter. The fixed residence time in the transverse channel provides a "diffusion window" for the fastest components in one stream to diffuse to the other parallel stream and thus be separated.

coefficients. In the center of the device shown in Figure 6.58, streams move in parallel, and diffusion causes equilibration of small molecules across the channel, whereas larger particles do not equilibrate during the transit time of the device. The fixed residence time in the transverse channel provides a "diffusion window" for the fastest components in one stream to diffuse to the other parallel stream and thus be separated. Simplifying, the design rules of such filter reillustrated in Figure 6.59, the width of the conduit, *w*, accomodates two flows parallel to each other and the design obeys the following relation $w^2 \ll D_1\tau$ and λ the overlap region (with fast and slow diffusing particles mixed the smaller particles come with the largest diffusion coefficient $D_1 > D_2$) obeys: $\lambda^2 \gg D_2\tau$. Typical values are $\lambda = 10\ \mu m$, $w = 40\ \mu m$, and $t = 5$ s.

Applications of the H-filter include cell lysis, nonmotile sperm separation, extracting a wide range of analytes from the mucins in saliva, etc.

For a final example of a passive mixer in low Reynolds number fluids, we return to chaotic advection mixing. A first example of this type of mixing was illustrated in Figure 6.52. As a second example we use the herringbone micromixer developed by Stroock et al. (2002),[34] shown in Figure 6.60.

In this mixing approach, staggered herringbone microgrooves in the bottom of a flow channel in the

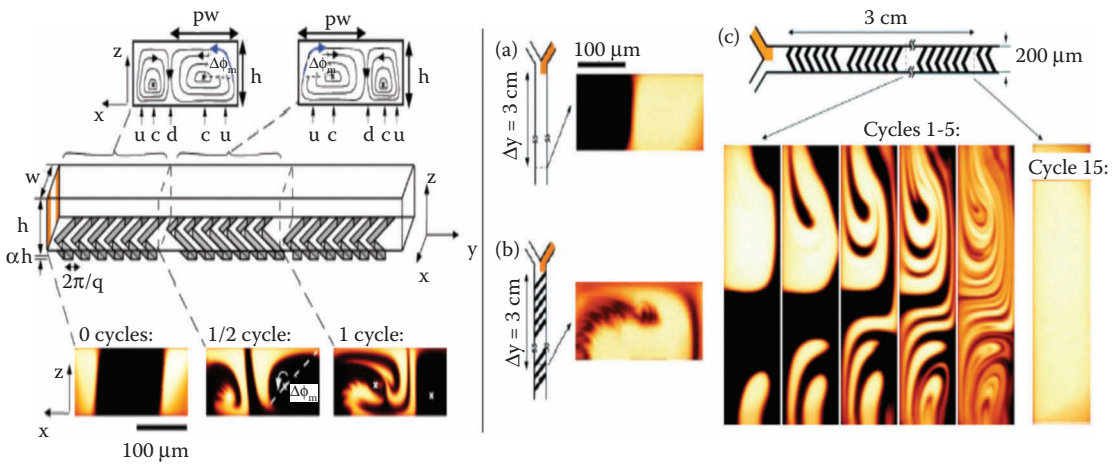


FIGURE 6.60 Micromixing by generating chaotic advection: the staggered herringbone mixer. By patterning grooves in the bottom of a microchannel, secondary flows are promoted that stretch and fold fluid elements, increasing areas with concentration gradients and so increase diffusive mass transport and mixing. In (a) no grooves are patterned, and parallel-flowing separated fluids undergo almost no mixing. By patterning grooves in several configurations, different secondary flows are promoted with smaller (b) or larger (c) mixing effects. (Stroock, A. D., S. K. W. Dertinger, A. Ajdari, I. Mezi, H. A. Stone, and G. M. Whitesides. 2002. Chaotic mixer for microchannels. *Science* 295:647–651.)

FIGURE 6.61 Mixing and stirring in low Reynolds number fluids is like kneading dough.

y-direction promote secondary rotating velocities (v_x, v_z). The underlying process for this type of efficient mixing, even in very viscous flow, is depicted in Figure 6.61 and is called the "baker's transformation. The main property of this transformation is that fluid pair elements will separate exponentially in "chaotic advection."

As illustrated in Figure 6.62, the width of the fluid folds (Δx) decreases exponentially, and the length of the interface (l) grows exponentially:

$$\Delta x(n) = \Delta x_0 (2)^{-n}$$

and

$$l(n) = l_0 (2)^n \qquad (6.103)$$

The mixing length L_m was estimated as[34]:

$$L_m \approx w \log Pe \qquad (6.104)$$

where w is the channel width, and Pe, the Péclet number, is calculated from the width and average flow velocity. This estimation, confirmed by experiments, leads to mixing lengths several tens of times larger than the width w for Péclet numbers on the order of 10^5.

Active Mixers

One also can use active perturbation to create time-dependent flows so that chaotic advection can occur. Different sources can be employed: pressure,

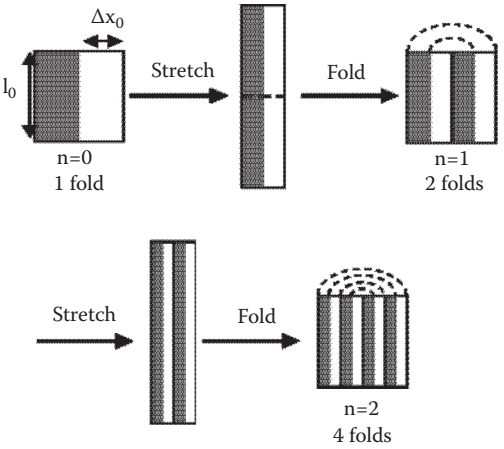


FIGURE 6.62 The baker transformation. Chaotic advection.

ultrasonic/piezoelectric, electrohydrodynamic, electro-osmosis, and mechanical stirring.

The mixing in the narrow section of the T-diffuser/mixer in Figure 6.56, for example, can be dramatically improved through hydrodynamic focusing, the technique used in cell and particle flow sorting. Two side flows "squeeze" or "hydrodynamically focus" an inlet flow. By using hydrodynamic focusing, the diameter of a central liquid jet can be reduced to just a few nanometers, which allows for mixing within this kind of beam in the tens of microseconds. This work, by Knight et al.,[35] is illustrated in Figure 6.63. The diameter of the jet (W_f) in the collection channel (W_c) is controlled by the ratio of the flow rates coming out of the three channels that bring the fluids together at the intersection. Knight et al. showed that with flow ratios of >100, W_f can be made to be around 30 nm.

Applications of hydrodynamic focus-based mixing include the study of fast reaction kinetics (like protein folding), which depend on fast mixing of reactants.

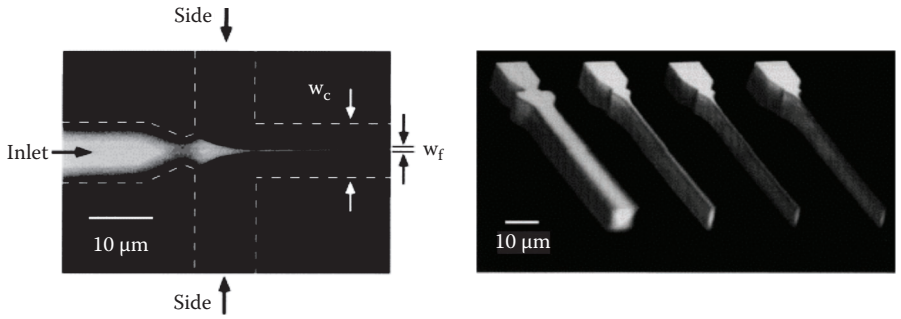


FIGURE 6.63 Geometry of a hydrodynamic focusing structure allowing for the reduction of the size of the jet and thereby enabling effective mixing transverse to the jet. (Knight, J. B., A. Vishwanath, J. P. Brody, and R. H. Austin. 1998. Hydrodynamic focusing on a silicon chip: mixing nanoliters in microseconds. *Phys Rev Lett* 80:3863–66.)

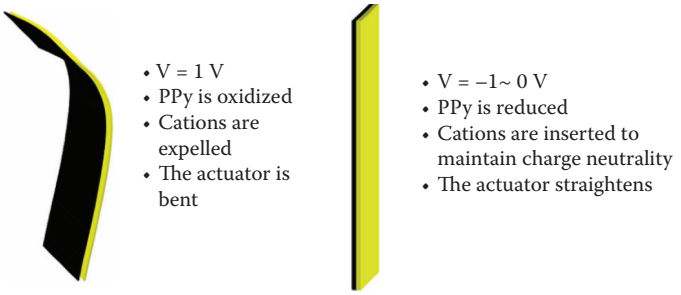


FIGURE 6.64 Principle of PPy/Au actuator. Mixing is achieved by activation of PPy-based actuators in microchannels.

Another active mixing approach pursued by the Madou group at the University of California, Irvine, involves the use of gold/polypyrrole actuators. Small rectangular gold/polypyrrole bilayer structures make for little flaps that can be made to move by biasing the polypyrrole with respect to a counter electrode in the same solution (see Figure 6.64). These actuators have been used to promote flows around a sensor in close proximity to the flapping bilayers. In doing so, convective mass transport is achieved, and the response in the sensor is increased over diffusive limits. Arrays of these mixing flaps have been fabricated as well with individual actuators ranging in dimensions from several millimeters to a few micros (see Casadevall i Solvas, X. 2009. *Active micromixing and in-vivo protection of sensors with gold/polypyrrole actuators*. PhD diss., University of California, Irvine, Irvine, CA; and Casadevall i Solvas, X., R. A. Lambert, L. Kulinsky, R. H. Rangel, and M. J. Madou. 2009. Au/PPy actuators for active micromixing and mass transport enhancement. *Micro and Nanosystems* 1:2–11).

The idea of the flexible flaps presented above is based on biomimetics. In 1977 Purcell, discussing swimming of microorganisms in fluids at very low Reynolds numbers,[36] presented the physical requirements that would allow microsystems to move (or propel fluids) in the microdomain (Figure 6.65). A simple single-hinged, rigid system (Figure 6.65a) would permanently move back and forth from its initial to its final position without accomplishing any real progress. To circumvent this limitation, multiple-hinged systems (Figure 6.65b) and flexible propelling structures (Figure 6.65c) can be used. Both configurations can be implemented easily with

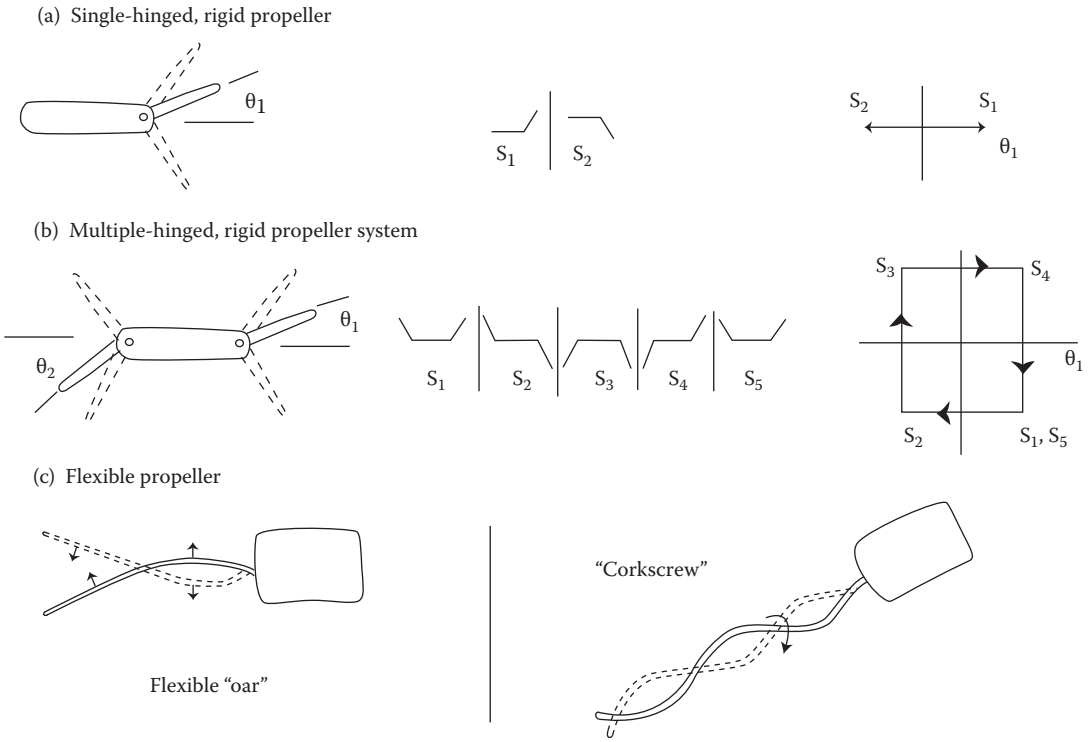


FIGURE 6.65 Swimming of microorganisms in fluids at very low Reynolds numbers. A simple single-hinged, rigid system (a) would permanently move back and forth from its initial to its final position without accomplishing any real progress. To circumvent this limitation, multiple-hinged systems (b) and flexible propelling structures (c) can be used. (Purcell, E. M. 1977. Life at low Reynolds number. *Am J Phys* 45:3–11.)

the flexible gold/polypyrrole actuators shown in Figure 6.64. In microvolumes of fluid, because of the flexible nature of these actuators and the possibility to actuate several of them simultaneously, these devices could be used to promote flows of sufficient complexity for chaotic advection to develop, and so enhance mixing in a targeted, isolated fashion, minimizing device and reagent volumes.

For a good review and an excellent comparison of active and passive micromixing, see Nguyen and Zhigang.[37] The summarizing comparison between active and passive mixing in Figure 6.66 is borrowed from this review.

Chemical Reactions in Microchambers—Microreactors

Introduction

The ultimate objective of passive and active micromixers in many instances is to speed up a chemical reaction by enhancing the mass transport to reaction sites in a chemical microreactor. As pointed out earlier, this finds a key application in the development of faster DNA and protein arrays for molecular diagnostics. These array assays need to be accommodated in a small reaction chamber, whereas reaction rates must remain high, despite low Reynolds numbers and high Péclet numbers. Because of these chemical reactions, the earlier described equations for fluid motion now need to be expanded with terms that take into account the generation (source) or consumption (sink) of reagents dissolved in the fluid. The proper design of chemical microreactors leads to faster and more sensitive assays.

Damköhler Numbers

Diffusive transport is captured by Fick's second law (see Equation 4.21 in Chapter 4), stating that the flux gradient $\left[-\frac{\delta F(x,t)}{\delta x}\right]$ is proportional to the change of the concentration with time. For a diffusion coefficient D, which is independent of position, this equation can be written for 3D using the ∇ operator as:

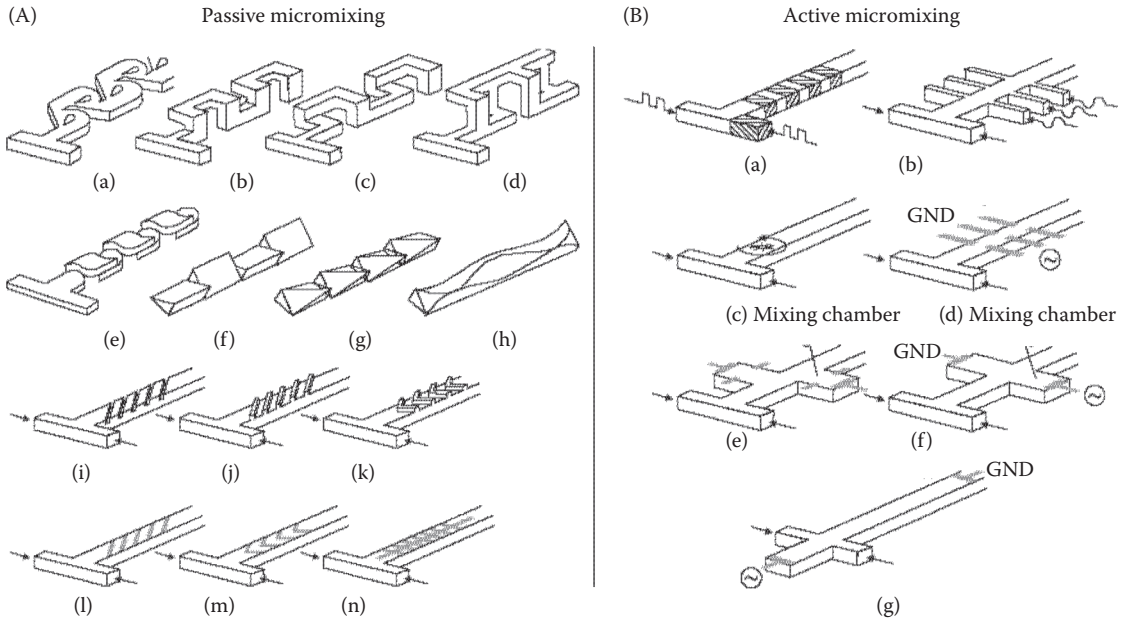


FIGURE 6.66 Comparing passive (A) and active (B) micromixing. Passive micromixers: modified Tesla structure (A) (a), C-shape (A) (b), L-shape (A) (c), connected out-of-plane L-shapes (A) (d), twisted microchannel (A) (e), other designs of twisted channel (A) (f–h), slanted ribs (A) (i), slanted grooves (A) (j), staggered-herringbone grooves (A) (k), and patterns for surface modification in a micromixer with electrokinetic flows (A) (l–n). Active micromixers: serial segmentation (B) (a), pressure disturbance along the mixing channel (B) (b), integrated microstirrer in the mixing channel (B) (c), electrohydrodynamic disturbance (B) (d), dielectrophoretic disturbance (B) (e), electrokinetic disturbance in the mixing chamber (B) (f), and electrokinetic disturbance in the mixing channel (B) (g). (Nguyen, N. T., and W. Zhigang. 2005. Micromixers—a review. *J Micromech Microeng* 15:R1–R16.)

$$\frac{\partial C}{\partial t} = D\nabla^2 C \tag{6.105}$$

For steady-state conditions the concentrations are constant in time, and there is no change in the mass stored with time, or $\nabla^2 C = 0$ (Laplace's equation). If, besides diffusion, we also consider convective or advective transport (mass flux due to the velocity V of the flow), we need to add a term:

$$\frac{\partial C}{\partial t} = D\nabla^2 C - \nabla CV \tag{6.106}$$

Expanding the advective term we obtain:

$$\frac{\partial C}{\partial t} = D\nabla^2 C - V\nabla C - C\nabla V \tag{6.107}$$

For transport in a steady-state flow field $\nabla V = 0$, so that:

$$\frac{\partial C}{\partial t} = D\nabla^2 C - V\nabla C \tag{6.108}$$

In case the velocity of the fluid is zero, we recover the diffusion equation. Finally, when reactions take place, we add a term for production (source) or removal (sink) of chemicals. This results in:

$$\frac{\partial C}{\partial t} = D\nabla^2 C - V\nabla C + r\ (= kC) \tag{6.109}$$

where r is the net mass produced or consumed per unit volume per unit time (moles/L^3t). For each species in the system an Equation 6.109 is required. The general form of the r term is $r = kC$, that is, r is the product of concentration and a reaction rate coefficient, k.

Depending on the type of reaction, r takes on different forms; in a first order–type reaction (e.g., encountered in radioactive decay), $r = -\lambda C$, where λ is a decay constant. In the case of second order kinetics, with reagents A and B, $r = -kC_AC_B$.

One can distinguish between two extreme cases depending on the Damköhler numbers. The Damköhler numbers (Da) are dimensionless numbers used to relate chemical reaction time scales to other phenomena occurring in a system. They are named after the German chemist Gerhard Damköhler, 1908–1944. There are several Damköhler numbers, and their definition varies according to the system under consideration. A global definition for the Damköhler number is:

$$Da = \frac{\tau_C}{\tau_M} \tag{6.110}$$

where τ_C is the characteristic time of the chemical reaction, and τ_M is the mixing time that can be controlled by hydrodynamics or by diffusion. So a small Damköhler number refers to a very slow chemical reaction compared to all other processes. Introducing a dimensionless concentration C^* (dividing C by an arbitrary concentration C_r); a dimensionless time t^* (dividing t by an arbitrary time t_r); and dimensionless lengths x^*, y^*, and z^* (by dividing by the characteristic length L), we nondimensionalize Equation 6.109 as:

$$\left[\frac{L^2}{Dt_r}\right]\frac{\partial C^*}{\partial t^*} = \nabla^2 C^* - \left[\frac{VL}{D}\right]\nabla C^* + \left[\frac{kL^2}{D}\right]C^*$$

Fourier Péclet Damköhler

$$\text{or Fo}\frac{\partial C^*}{\partial t^*} = \nabla^2 C^* - \text{Pe}\nabla C^* + \text{PeDa}_I C^*$$

$$\text{where Da}_I = \frac{kL}{V}$$

$$\text{and also Fo}\ \frac{\partial C^*}{\partial t^*} = \nabla^2 C^* - \text{Pe}\Delta C^* + \text{Da}_{II} C^*$$

$$\text{where Da}_{II} = \frac{kL^2}{D} \tag{6.111}$$

The Fourier number Fo (L^2/Dt_r) is a dimensionless quantity that can be used to characterize a diffusion length for a particular time. For example, if $D = 10^{-9}$ m^2/s and $t_r = 1$ year, then the diffusion length is about 0.18 m. The first Damköhler number ($Da_I = kL/V$) measures the tendency for reaction versus advective transport. The second Damköhler number ($Da_{II} = kL^2/D$) measures the tendency of reaction relative to the tendency for diffusive transport. With a small Da_I or Da_{II}, convection and/or diffusion is much faster than chemical reaction, and the concentration in the reactor is uniform. The system is reaction rate-controlled, and convection and/or diffusion have no influence. With a large Da_{II}, the chemical reaction is very fast so that the concentration, say

on an electrode, might fall to zero, and the system is transport-limited.

Since the Biot number (see above) will tend to decrease for microdevices, it is expected that microchemical reactors will have nearly uniform temperatures. Thus, the wall effects will be very important from the point of view of keeping a high-temperature reaction going. The tendency will therefore be for the presence of small Damköhler numbers (weak reaction) due to heat losses through the reactor walls that reduce the reaction rate and consequently increase the chemical time, while the small dimensions reduce the diffusion time. For a general chemical reaction A → B of n-th order, the Damköhler number is defined as:

$$Da = kC_0^{n-1}t \tag{6.112}$$

where k = kinetics reaction rate constant
C_0 = initial concentration
n = reaction order
t = time

Electrochemical Reactions

Here we set up an expression for the current on an electrochemical sensor with transport of electroactive species to and from the electrode by diffusion, migration, and convection. From Fick's first law (Equation 4.20) we derive for the flux or current resulting from electroactive species *i*:

$$J_i(x) = -D_i \frac{\delta C_i(x)}{\delta(x)} - z_i \mu_i F C_i \frac{\delta\phi}{\delta x} + C_i V(x)$$

$$\Uparrow \qquad \Uparrow \qquad \Uparrow$$

Diffusion Migration Convection (6.113)

where J_i = flux of species *i* in mol/cm² s
z_i = charge number of species *i* in eq/mol
μ_i = mobility of species *i* in cm²mol/Js
C_i = concentration of species *i* in mol/cm³
ϕ = electrostatic potential in *V*
F = the Faraday constant
D_i = diffusion coefficient of species *i* in cm²/s
V = fluid velocity in cm/s

In this expression for an electrochemical reaction, the new term we consider is migration transport. For migration of species, say negatively charged DNA molecules to a positively charged electrode, the solution must have a low conductivity as otherwise the field does not penetrate the liquid (see Chapter 7). As we learn in Chapter 7, to obtain reliable analytical data on amperometric electrochemical sensors, the migration term needs to be minimized in favor of the diffusion and convection terms. This, we will learn, is accomplished by the addition of an indifferent electrolyte suppressing the migration of the electro-active species of interest.

In Table 6.5 we summarize the important transport phenomena determining the response of an electrochemical sensor.

Electrochemicals are studied in detail in Chapter 7 where they are also compared with optical sensors.

Fluid Propulsion

Introduction

Many options present themselves for fluid propulsion, including mechanical means, dc and ac electrokinetics, acoustic pumping, electrowetting, centrifugal, electrohydrodynamic and magnetohydrodynamic, etc. Acoustic, electrohydrodynamic, and magnetohydrodynamic pumping techniques are reviewed in Volume III, Chapter 8 on actuators, and in Volume III, Chapter 1 we compare and summarize the most important new pumping methods and their MEMS integration (see Volume III, Table 1.8). In the current section we detail the theory behind dc and ac electrokinetics, electrowetting, and centrifugal fluidics.

Electrokinetic Effects—Direct Current (DC)

Introduction to Electrophoresis and Electro-Osmosis

An electrical field works on both charged particles in a fluid and on the fluid itself when the fluid is placed in a narrow capillary with electrically charged walls. The force on the particles in the fluid leads to electrophoresis; the force on the fluid in a narrow column leads to electro-osmosis. Both phenomena are of great importance in miniaturization science: electrophoresis for separation of compounds and electro-osmosis for movement of fluids. Mathematically, it does not make a difference whether the fluid passes a charged wall or a charged particle moves through a fluid.

Table 6.5 Transport Phenomena on an Electrochemical Sensor

Transport	Migration	Diffusion	Convection
Response to	Potential gradient	Concentration gradient	Pressure gradient
Acts on	Ions only	All solutes	Solutes + solvent
Flux density (from Equation 6.113) J =	$-zF\mu C\frac{\partial\phi}{\partial x}$ (mobility)	$-D\frac{\partial C}{\partial x}$ (diffusivity)	CV
Relationship between μ and D	Einstein derived a simple relation between D and μ as: $D = \mu RT$		
Combining migration with diffusion flux	$J = -D\left[\frac{\partial C}{\partial x} + \frac{zF}{RT}C\frac{\partial\phi}{\partial x}\right]$ (Nernst-Planck flux equation using the Einstein relation between D and μ.)		
Conservation law	$\frac{\partial C}{\partial t} = -\frac{\partial J}{\partial x}$ for planar transport; $\frac{\partial C}{\partial t} = -\frac{\partial J}{\partial r} - \frac{2J}{r}$ for spherical transport		

Electrophoresis

Electrophoresis is the migration of ions in a separation medium under the influence of an electric field. An ion with charge q (Coulombs) in an electric field E (Vm^{-1}) feels an acceleration force qE (Newtons). A steady-state speed is reached when the accelerating force (qE) equals the frictional force generated by the separation medium, or:

$$qE = fV_E \tag{6.114}$$

where f = the friction coefficient or friction factor
V_E = the electrophoretic steady-state speed of the ion

The expression for the electrophoresis velocity V_E of a charged species in a separation medium can then be written as:

$$V_E = \mu_E E \tag{6.115}$$

with μ_E ($m^2/V/s$) the electrophoretic mobility (= q/f) and E the applied electrical field. The friction coefficient for a spherical particle of radius r moving through a fluid of viscosity η, we learned from Equation 6.26, is given by $6\pi\eta r$, and its electrophoretic mobility consequently equals:

$$\mu_E = \frac{q}{6\pi\eta r} \tag{6.116}$$

The frictional coefficient f depends on the size of the molecule in free solution. For a solid sphere f is given by $6\pi\eta r$, so that $f \sim MW^{1/3}$ because $r \sim MW$ (where MW is molecular weight). Other models are more appropriate for differently shaped molecules, as summarized in Table 6.6.

According to the free-draining coil model, a randomly coiled polymer chain in motion hardly interacts with the solvent within it at all. DNA is a good example of a free-draining coil with each unit of the polymer contributing equally to the friction created by its movement through a fluid. Because the charge q is proportional to N (number base pairs), and $f \sim (MW)^1 \sim N$, we conclude that:

$$\mu_E = \frac{q}{f} = \text{constant} \tag{6.117}$$

As a consequence, DNA strands of varying size cannot be separated by electrophoresis in solution.

TABLE 6.6 Friction Factor f for Variously Shaped Molecules

Model	Proportionality
Sphere	$f \sim (MW)^{1/3}$
Random coil	$f \sim (MW)^{1/2}$
Long rod	$f \sim (MW)^{4/5}$
Free-draining coil	$f \sim (MW)^{1}$
MW = molecular weight.	

This is why we need to perform electrophoresis in a porous or tortuous material such as a gel. In this case, the solute moves snake-like through the polymer pores. In this model, the frictional term is $f \sim N^2$. This leads to:

$$\mu_E = \frac{q}{f} \approx \frac{N}{N^2} \approx \frac{1}{N} \tag{6.118}$$

which is the type of behavior normally observed in DNA electrophoresis.

The difference in ion mobility gives rise to a separation in bands of the various solutes, and the narrower the bands, the better the separation. Traditionally, electrophoresis was carried out on columns packed with gels (e.g., a polyacrylamide gel) or on gel-covered plates (Figure 6.67).

Electrophoresis in narrow capillaries, as shown in Figure 6.68, was pioneered by researchers such as Virtanen[38] and Jorgenson,[39] who recognized important advantages of this approach. Performing electrophoresis in small-diameter capillaries allows the use of very high electric fields because the small capillaries efficiently dissipate the heat that is produced. Increasing the electric fields produces very efficient separations and reduces separation times. Capillaries are typically of 50-μm inner diameter and 0.5 to 1 m in length. The applied potential is 20 to 30 kV. Because of electro-osmotic flow (see next section), all sample components migrate toward the negative electrode. A small volume of sample (10 nL) is injected at the positive end of the capillary, and the separated components are detected near the negative end of the capillary. CE detection schemes include absorbance (see the example in Figure 6.68), fluorescence, electrochemistry, and mass spectrometry. The capillary also can be filled with a gel, which eliminates the electro-osmotic flow. Separation is accomplished as in conventional gel electrophoresis, but the capillary allows higher resolution, greater sensitivity, and on-line detection.

To cool the capillary during electrophoresis, a Peltier device, forced air convection, or a flowing liquid bath may be used.

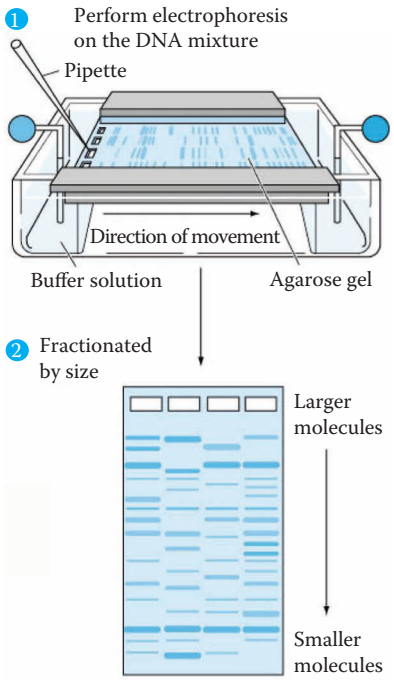


FIGURE 6.67 Gel electrophoresis.

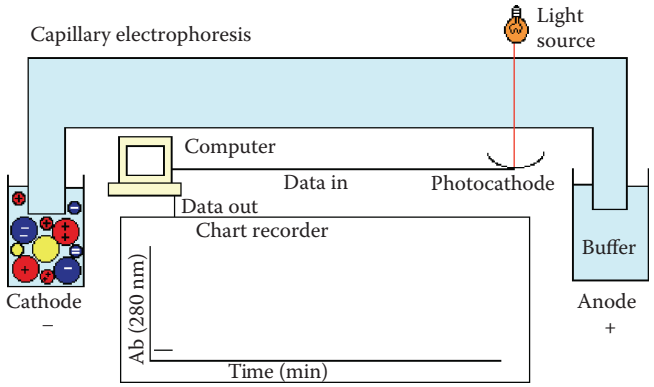


FIGURE 6.68 Capillary electrophoresis schematic. A sample may be injected in the capillary through pressure application (e.g., a syringe) on the sample vial or suction at the outlet of the capillary. A more elegant solution for sample introduction is through electro-osmotic pumping, discussed below. The chart recorder in this case plots absorbance (Ab at 280 nm) versus time.

Electro-Osmosis

Capillary electrophoresis to some degree always is accompanied by electro-osmosis. The inside wall

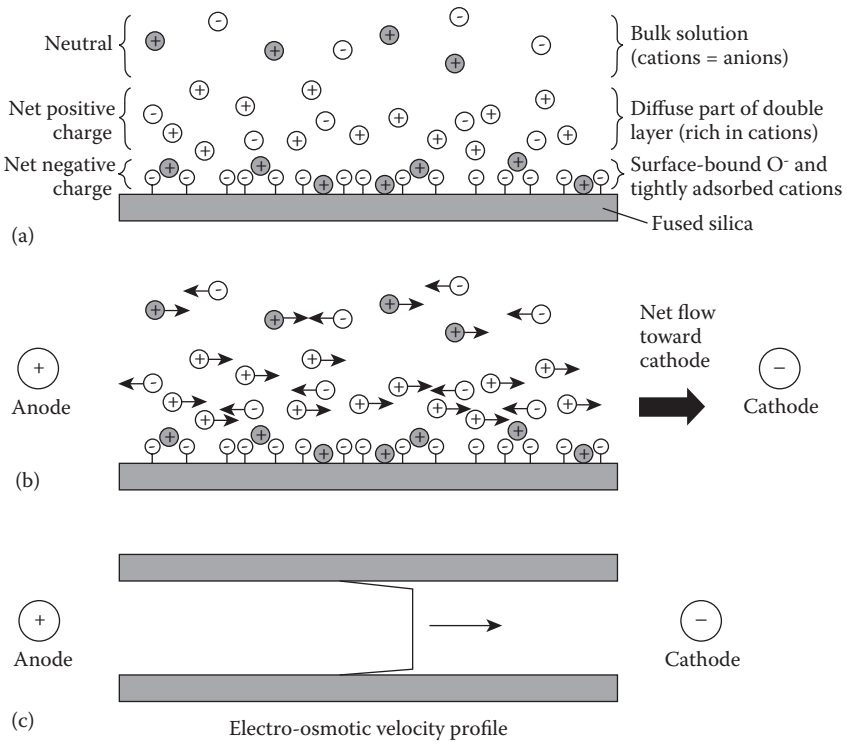


FIGURE 6.69 A schematic diagram showing the origin of electro-osmotic flow. (a) Electric double layer is created by negatively charged silica surface and excess cations in the diffuse part of the double layer in the solution near the wall. The wall is negative, and the diffuse part of the double layer is positive. (b), Predominance of cations in diffuse part of the double layer produces net electro-osmotic flow toward the cathode when an external field is applied. (c) Electro-osmotic velocity profile is uniform over more than 99.9% of the cross-section of the capillary. Capillary tubing is required to maintain constant temperature in the liquid. Temperature variation in larger diameter tubing causes bands to broaden.

of a fused silica capillary is covered with silanol groups (Si-OH). Silanol groups have a pK_a near 2 and consequently carry a negative charge ($Si\text{-}O^-$) at pH > 2. Positive compensating charges (cations) are either in an immobile layer immediately adjacent to the glass surface, the Stern layer, or in the diffuse part of the double layer, referred to as the Gouy-Chapman layer. The Gouy-Chapman layer stretches a bit further out into the solution, ~10 nm further away when the ionic strength of the medium in the capillary is 1 mM.* In a tangential electric field, excess cations in the mobile part of the diffuse double layer are attracted to the cathode, and this imparts a pumping action onto the whole fluid column, which moves toward the cathode (Figure 6.69).[40] This pumping describes electro-osmotic or also electroendosmotic flow.

The electro-osmotic velocity, V_{EOF}, is given by:

$$V_{EOF} = \mu_{EOF} E \tag{6.119}$$

where μ_{EOF} = the electro-osmotic mobility
E = the applied electrical field

The electro-osmotic mobility (q/f) is given by:

$$\mu_{EOF} = \frac{\varepsilon \zeta}{4\pi\eta} \tag{6.120}$$

where ε = the dielectric constant
ζ = electrokinetic or zeta potential
η = is the fluid viscosity

The structure of the electrical double layer (EDL) is detailed in Chapter 7. Here we list some of the most salient features of the EDL model. A thin layer of liquid along the capillary wall never moves; liquid movement begins only at a shear plane in the diffuse double layer at a distance, x_0, away from the wall. The electrokinetic potential ζ (from which all electrokinetic effects can be derived) represents

* For a binary electrolyte in aqueous solution, the double-layer thickness ranges from 3 to 300 nm for electrolyte concentrations of 10^{-2} to 10^{-6} M, respectively.

the value of the electrostatic potential, ψ, at x_0. Its value at any point is determined from the profile of the electrostatic potential as a function of distance from the capillary wall into the moving fluid (see Figure 7.25). The thickness δ of the diffuse layer itself is defined as the distance from the immobile Stern layer to a point at which the electrostatic potential has dropped to 37% of the zeta potential. The amount and type of ionic species in the solution can greatly influence the zeta potential; for example, the larger the concentration of added indifferent electrolyte in the moving solution, the smaller the ζ (see Figure 7.25). The zeta potential of an aqueous solution in contact with glass can have a magnitude as high as 100 mV.[3] Besides pH and ionic strength, surface impurities also might affect the electrokinetic potential, and all conspire to make electro-osmosis a rather difficult pumping mechanism to control. Despite these barriers, several new miniaturized bioanalytical products based on electrokinetic effects are now on the market (e.g., visit www.nanogen.com and www.calipertech.com). Most product applications concern genomic analysis; medical and environmental diagnostics have not been able to garner the same amount of funding today.

Generalizing, Equations 6.115 and 6.119 may be expressed as:

$$V = \mu_a E \quad \text{or} \quad V = L_d / t \cdot V / L_t \qquad (6.121)$$

where V = the velocity of ions or fluid column
μ_a = the apparent mobility (i.e., the vector sum of electro-osmotic and electrophoretic mobilities)
E = the electric field strength
L_d = the length of the column from injection to detector
t = the time required for solute to migrate from the injection point to the detector
V = the applied voltage between the two ends of a capillary with total length L_t

As pointed out earlier, mathematically, electro-osmosis and electrophoresis are indistinguishable: either the charge moves through the medium or the medium moves through the charged capillary. In the case of electro-osmosis, ζ is the electrokinetic potential at the capillary wall; in the case of electrophoresis, it is the electrokinetic potential at the charged particle. In absolute terms, electro-osmotic mobility generally is larger than electrophoretic mobility. Under typical experimental conditions, say for DNA separations (300 V/cm), the electro-osmotic mobility of DNA molecules within fused silica capillaries is 1.25 to 3 times larger than their electrophoretic mobility.

In some cases electro-osmosis is undesirable; for instance, in gel-filled capillaries the gel may ooze out of the capillary or positively charged proteins may adsorb onto the capillary walls. Electro-osmosis may be suppressed by chemical modification of the capillary surface with a neutral EOF-suppressing polymer (e.g., by coating with polyacrylamide[41] or methyl cellulose). By the adjustment of the buffer pH and/or buffer concentration, or by the application of a radial electric field, the EOF also may be modulated by the user.[40] The latter may change both the magnitude and the polarity of the zeta potential.

In the section below on scaling in analytical separation techniques, we will further detail advantages and disadvantages of miniaturizing electrophoresis and electro-osmosis and separation equipment in general.

Flow Profiles

The flow through a separation media such as a capillary may be a consequence of an applied pressure (i.e., hydrodynamic) or of an applied field (i.e., electro-osmotic). In a pressure-driven system, an external pump provides the pumping force. As shown in Figure 6.70a, a cross-section of a hydrodynamic velocity flow profile driven by a pressure difference between the ends of a column is parabolic (see Equation 6.46 and Figure 6.22). It is the fastest at the center and slows to zero at the walls. When injecting a solute, the parabolic flow profile makes for a less attractive analytical tool, as it causes band broadening (see also the section below on scaling in separation techniques). Electro-osmosis, in which the driving force is generated very close to the capillary wall, gives a uniform flow over more than 99.9% of the cross-section of a capillary column, as shown in Figure 6.70b. The speed of the flow falls off immediately adjacent to the capillary wall. To illustrate

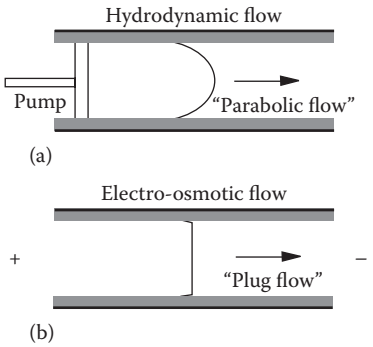


FIGURE 6.70 A hydrodynamic velocity profile (a) compared with an electro-osmotic velocity profile (b).

this with a practical example, in a typical DNA electrophoresis experiment, the double layer δ is around 30 Å thick, and x_0 (see Figure 7.25) is even thinner. Compared with the inner diameter of a capillary of 50 μm, the region of velocity variation is negligible. Flat flow profiles are expected when the capillary radius is greater than about 7 times the double-layer thickness. The rapid flow of buffer and analyte with a flat velocity profile is attractive, as it can be used to pull all analytes, regardless of their mobility, past a detector without contributing significantly to band broadening.

In hydrodynamic injection of a sample in a capillary, the injected volume is derived from the volumetric flow rate, Q, in m^3s^{-1}. The flow rate Q, of an incompressible fluid with viscosity η ($N/s/m^2$) in a capillary with inner diameter d (m) and total length of the capillary L_t (m), resulting from an applied pressure Δp ($N\,m^{-2}$), is calculated from Hagen-Poiseuille's Law (Equation 6.54) as:

$$Q = \frac{\Delta p \pi d^4}{128 \eta L_t} \qquad (6.54)$$

This equation applies for laminar flow of a fluid through a capillary. To obtain the volume, V, the volumetric flow rate Q is multiplied by t, where t is the injection time:

$$V = Q \cdot t \qquad (6.122)$$

Conventional pumps must generate extremely high pressures when narrow capillary diameters are used, and they are not well suited for the delivery of small samples. This is where electro-osmotic pumping comes in.

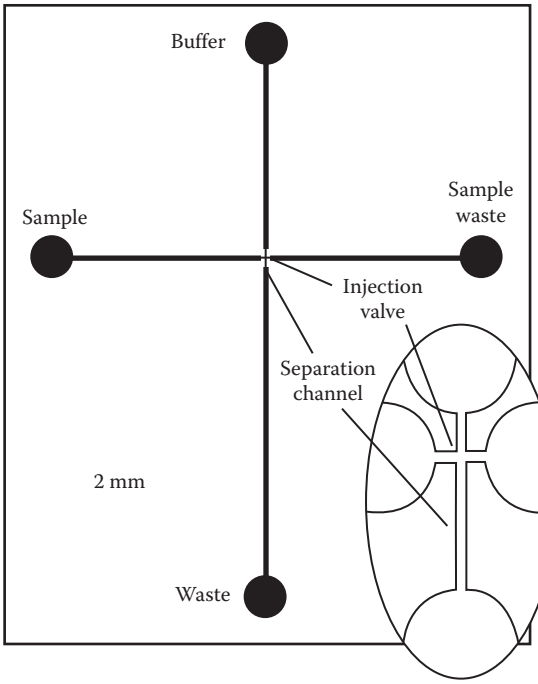


FIGURE 6.71 Microchip used for high-speed electrophoretic separations. Inset: Enlargement of the injection valve and separation channel. (Based on Jacobson, S. C., C. T. Culbertson, and J. M. Ramsey. 1998. *Solid-state sensor and actuator workshop*. Hilton Head Island, SC: Transducers Research Foundation.)

Electrokinetic Injection

Electrokinetic injection is popular in micromachined fluidics (see Figure 6.71), but it is more generally applied in capillary gel electrophoresis, in which the liquid is too viscous to allow for hydrodynamic injection. In electrokinetic injection, a capillary is dipped in sample solution, and a voltage is applied between the ends of the column. The moles of each ion, mol_i, absorbed into the capillary in t seconds are given by:

$$mol_i = \mu_a \left(E \frac{\kappa_b}{\kappa_s} \right) t \pi r^2 C_i \qquad (6.123)$$

where μ_a = the apparent mobility of the analyte
E = the applied electric field
r = the radius of the capillary
C_i = the sample concentration ($molm^{-3}$)
κ_b/κ_s = the ratio of the conductivities of the buffer and the sample

Electrokinetic injection leads to sampling bias, because a disproportionately large quantity of the species with higher electrophoretic mobility migrates into the tube and can cause problems for

quantitative analysis. Moreover, electro-osmotic flow rates tend to change over time because the zeta potential changes over time, thus leading to changes in μ_a in Equation 6.123. Below, we will see that certain types of electro-osmotic injection methods in miniaturized flow channels avoid this type of sample biasing and aging effects.

Micromachined Capillary Electrophoresis Equipment

Planar Fluidic Networks In free-flow, traditional capillary electrophoresis work by Jorgenson et al.,[42] electro-osmotic flow rates of 1.7 mm s^{-1} were reported, certainly fast enough for many analytical purposes. The technology for generating this flow rate does not involve moving parts and is easily implemented in micromachined channels. Using micromachined channels in a substrate instead of glass capillaries offers the potential for mass production, the choice of many materials (e.g., different plastics besides glass), and the integration of many processing steps on one chip. One needs only one electrode in the reservoirs at each end of the different flow channels, and these electrodes may be thin films deposited on the substrate or needle electrodes inserted into the reservoirs from above (bed-of-nails approach). Harrison et al.[43] achieved electro-osmotic pumping with flow rates up to 1 cm s^{-1} in 20-μm capillaries micromachined in glass. Compared with conventional capillary electrophoresis, where separations usually are in the order of 10–20 min, microfabricated electrokinetic devices are capable of producing high quality separation in seconds or even submilliseconds as first demonstrated by Jacobsen et al.[44,45] In Figure 6.71 we show a schematic of the microchip used by Jacobson et al.[46] for submillisecond electrophoresis using a field strength of 53 kV cm^{-1} and a separation length of 200 μm. Two dyes, rhodamine B (RB) and dichlorofluorescein (DCF), were resolved in 0.8 ms using this system.[47]

The earliest patents describing electrophoresis and electro-osmosis in a micromachined channel with side branches in a planar substrate are by Pace (filing date June 1988)[48] and by Soane et al.[49] (priority filing date Feb 28, 1990). The latter is the more generic patent of the two, as it covers all insulating substrates for making fluidic channels, whereas the former patent is confined to Si. Si, besides being expensive, is a semiconductor and needs to be coated with an insulator before it can be used in an electrokinetic application. When using SiO_2 as the insulator, the oxide tends to break down and age rapidly in contact with water, making Si far from ideal as a substrate for an electrokinetic device. As far back as 1991, Manz et al.[50] demonstrated the injection, mixing, separation, and reaction of fluids in a manifold of micromachined flow channels without the use of mechanical valves.

Manz et al.[51] described a set of reservoirs connected by an H-like pattern of capillary channels etched into a glass substrate. This setup is schematically reproduced in Figure 6.72. Application of a voltage between reservoir 1 (sample) and reservoir 4 (injection waste) results in flow of sample across the center of the H. Subsequent application of a potential between reservoir 2 (carrier) and reservoir 5 (waste) directs the geometrically confined sample plug from the center bar of the H into the elongated leg of the H, that is, into the separation channel and toward the detector. By applying a potential between reservoirs 1 and 4 for an extended period of time, the composition of the fluid plug in the center of the H will be identical to that of the sample, thus avoiding the bias errors described under electrokinetic injection above. In addition, because the injected volume is geometrically defined, the amount of injected sample is reproducible and independent of changes in the electro-osmotic flow over time. The key to the tightest possible valving of liquids at intersecting capillaries in a manifold, as shown in Figure 6.72, is the suppression of convection and diffusion effects. Harrison et al.[43] demonstrated that these effects can be well controlled by appropriate and simultaneous application of voltages to the intersecting channels. Fast separations in micromachined fluidics are possible by confining the smallest possible sample plug at an intersection.[52,53] Ramsey achieved this by controlling the voltages at all four terminals of a simple cross structure, as shown in Figure 6.73, and pinching the plug at the intersection. In this way, the injected sample plug is of the order of the channel width. In the pinched injection mode demonstrated in Figure 6.73, analyte is pumped electrophoretically and electro-osmotically from reservoir 1 to 2 (left to right) with mobile phase from reservoir 3 (top) and

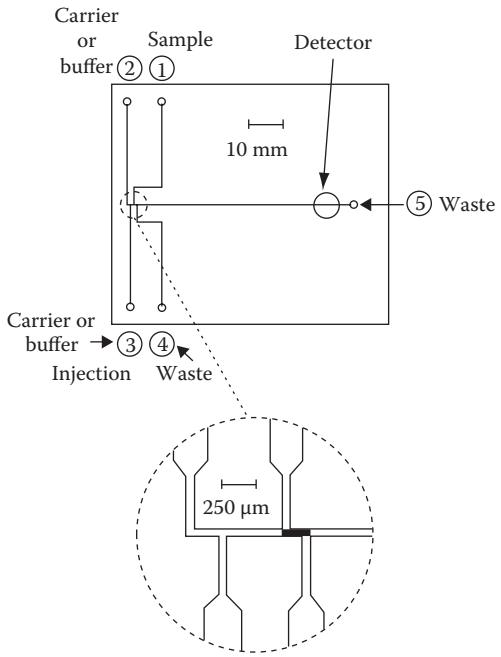


FIGURE 6.72 Schematic drawing of an electrokinetic injection process using a branched microflow system. Layout of the glass chip with integrated sample injector. Channel cross-sections: 50 × 12 (thin channels) and 250 × 12 μm (broad channels), respectively. After application of a high voltage between reservoirs 1 (sample) and 4 (injection waste), the geometrically defined injection volume (double-"T" injection, circled area) is filled by electrophoretic migration of the sample ions. After loading is completed, application of a high potential between reservoirs 2 (buffer) and 5 (waste) drives the injected sample plug into the separation channel and causes electrophoretic separation of the sample components. The detection is based on laser-induced fluorescence (LIF). The T-intersection where the selected sample zones are withdrawn is enlarged in the circled image. (Based on Effenhauser, C. S., A. Manz, and H. M. Widmer. 1995. Manipulation of sample fractions on a capillary electrophoresis chip. *Anal Chem* 67:2284–87.)

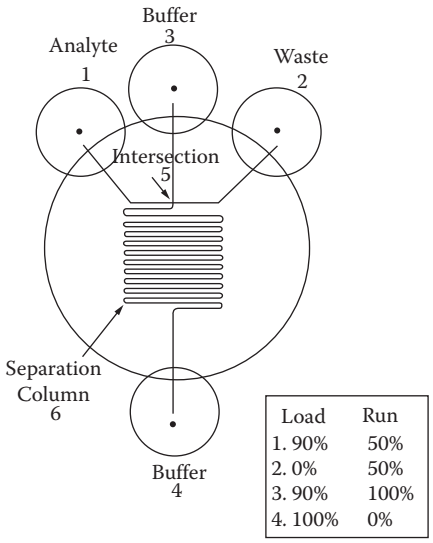


FIGURE 6.73 Pinched injection. Load and run modes (see text). (Based on Ramsey, J. M. 2000. Lockheed Martin Energy Research Corporation. US Patent 6,010,607.)

reservoir 4 (bottom) traveling to reservoir 2 (right). The appropriate voltages, as indicated in Figure 6.73, are applied to reservoirs 1, 3, 2, and 4, whereas reservoir 2 is grounded, thereby spatially constricting the analyte in the injection cross or intersection 5. This plug can be injected into the separation column 6 by applying a bias between electrodes 3 and 4 (4 is grounded now) while the potential of reservoirs 1 and 2 is kept at half the value of reservoir 3.

The above plumbing examples are simple; more complex manifolds exist that enable serial and more complex analytical procedures. These procedures include flow injection analysis, metering, mixing, reaction, dilution, aliquoting, pre- and postcolumn derivatization, addition of masking agents, and preconcentration (perhaps using isoelectric focusing techniques; see below). An interesting manifold configuration by Burggraf et al.[55] involves synchronized cyclic capillary electrophoresis (SCCE). In this technique a sample, through a sequence of voltage applications, is cycled through a loop made up of intersecting capillaries; after each pass the sample gets separated better into its constituent compounds.[55] A single electrokinetic channel can only be optimized to a point; once an optimum critical channel length and a minimal dead volume detector and injector have been designed (see scaling in analytical separations), other means to improve throughput must be implemented. For example, to obtain high throughput with integrated fluidics, one important need is for serial-to-parallel converters.[56]

Optimal Size of Fluidic Channels—Scaling With 1-mm-wide conduits and flow velocities not greater than a typical centimeter-per-second range, the Reynolds numbers for flow are so low that all flow is laminar. Mixing then occurs solely by diffusion, and the time required for molecules to travel a distance x is given by $\tau = x^2/2D$ (see Equation 6.100). For homogeneous flow to be established, bringing two streams of different compositions together at a Y-junction requires that the molecules diffuse over

500 µm (one-half of the channel width). For a molecule such as fluorescein with a D of 3×10^{-6} cm^2/s, this process will take 400 s, and with a flow velocity of 1 cm/s, the channel must be 4 m long before the mix has homogenized. Reducing the channel width to 70 µm and using the same flow rate, mixing is established in 2 s over a distance of 2 mm only. This suggests an upper limit of about 100 µm for the channel width, based on the need for fast mixing by diffusion. At very small dimensions, another limitation lurks. Assume that one wants to detect an analyte at 1 nM concentration in a cubic micrometer volume. The number of molecules in that small volume is only 0.6 (see also Volume III, Table 7.5). In other words, about half of the time nothing there will be detectable. Single-molecule detection is possible, for example, with fluorescence burst counting as demonstrated by Haab et al.[57] and Wahl et al.,[58] but it is far from common; 1000 molecules is a much more comfortable number with which to work. To have about 1000 molecules available in a 1-nM solution (a quantity readily detected), a volume of 12 µm^3 is needed. So, for optimal detection chamber dimensions should not be much smaller than 10 µm. With a channel cross-section of 2×10^{-3} mm^2 and a flow velocity between mm/s and cm/s the flow rate is in the nL/s range. Often, flow control must be within 1% or 1 pL/s. This tight tolerance is impossible to reach with today's micropumps, which cannot provide accurate control in the nL/s range. With electrokinetics flow control better than 1% is easily achieved.[19]

Isoelectric Focusing Isoelectric focusing (IOF) is the migration of charged particles (electrophoresis) under pH gradients to a location in the buffer, where they have zero net charge (isoelectric point). The pH gradient is set up between a cathode and an anode with the cathode at the higher pH.[59] This ingenuous process simultaneously concentrates and separates. Individual charged species, such as proteins, are rendered immobile as electrophoresis pushes them into the pH gradient until they arrive at the pH that renders them neutral, that is, at their isoelectric point (Ip) or isoelectric pH (IpH). Small amphoteric molecules (called *ampholytes*), which comprise a multitude of varying Ips, form the pH gradient. The ampholytes in an agarose or polyacrylamide gel carry a net charge, and a linear

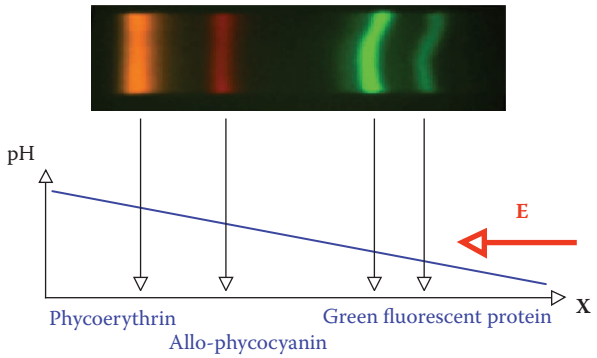


FIGURE 6.74 Isoelectric focusing (IOF).

pH gradient is built up when an electric field is applied in the so-called prefocusing step (see Figure 6.74).

In this brief treatment of the subject, we are particularly interested in the advantages and disadvantages of downscaling a static IOF system compared with downscaling a traditional electrophoresis system.

The focusing effect of the electrical force is counteracted by diffusion, which is directly proportional to the analyte (e.g., proteins) concentration gradient in the zone. At steady state, a balance between electrophoresis and sample diffusion holds and can be described by:

$$C\mu E = D\frac{dC}{dx} \tag{6.124}$$

where C = the sample concentration
dC/dx = the concentration gradient
E = the applied field strength
D = the diffusion coefficient
µ = the electrophoresis mobility

At steady state, this differential equation yields a Gaussian [$C(x) = C_0 \exp(-x^2/2\sigma^2)$] for the concentration distribution, resulting in individual bands with bandwidth σ:

$$\sigma = \pm\sqrt{\frac{D\dfrac{dx}{dpH}}{E\left(\dfrac{d\mu}{dpH}\right)}} \tag{6.125}$$

As the values of E and $d(\text{pH})/dx$ increase, the distribution width decreases. At a separation of 3σ, bands are considered resolvable, and the minimum pH separation between adjacent zones is described as:

$$\Delta pH_{min} = \Delta L_{min} \frac{dpH}{dx} = 3\sigma \frac{dpH}{dx} \qquad (6.126)$$

The resolving power is deduced by setting ΔpH equal to ΔpI and replacing σ with Equation 6.125:

$$\Delta pI_{min} = 3 \sqrt{\frac{DdpH}{\left(\frac{d\mu}{dpH}\right)V}} \qquad (6.127)$$

To arrive at Equation 6.127, we also substituted the electric field *E* by *V*/*L*, with *V* the applied potential and L (= dx) the length of the separation medium. From this equation, the minimum resolvable isoelectric point difference between adjacent focused bands is independent of *L* and dependent only on individual protein characteristics, the total pH difference, and the applied potential.[60] With pH and voltage gradient held constant, a sharper pH gradient (i.e., a shorter column) thus is expected to lead to a better separation. Exceptionally short columns might enable shorter analysis time and imaging of the entire separation with an inexpensive line array charge-coupled device (CCD).[60] In principle IOF downscales better than capillary electrophoresis, where the capillary length determines the resolution (see further below under scaling in analytical equipment). The maximum allowable temperatures dictate the minimum IOF separation length. Higher temperatures decrease the viscosity in an exponential fashion, thereby increasing the diffusivity of the sample. As the separation channel is shortened below 1 mm, Herr et al.[60] calculate a 23% increase in diffusivity compared with macroscale IOF (the macroscale device in the comparison is 20 cm long, and both have a 50-μm diameter channel). As a compromise, this research group chose an 8-mm-long channel. In Figure 6.75 capillary electrophoresis and capillary isoelectric focusing are compared.

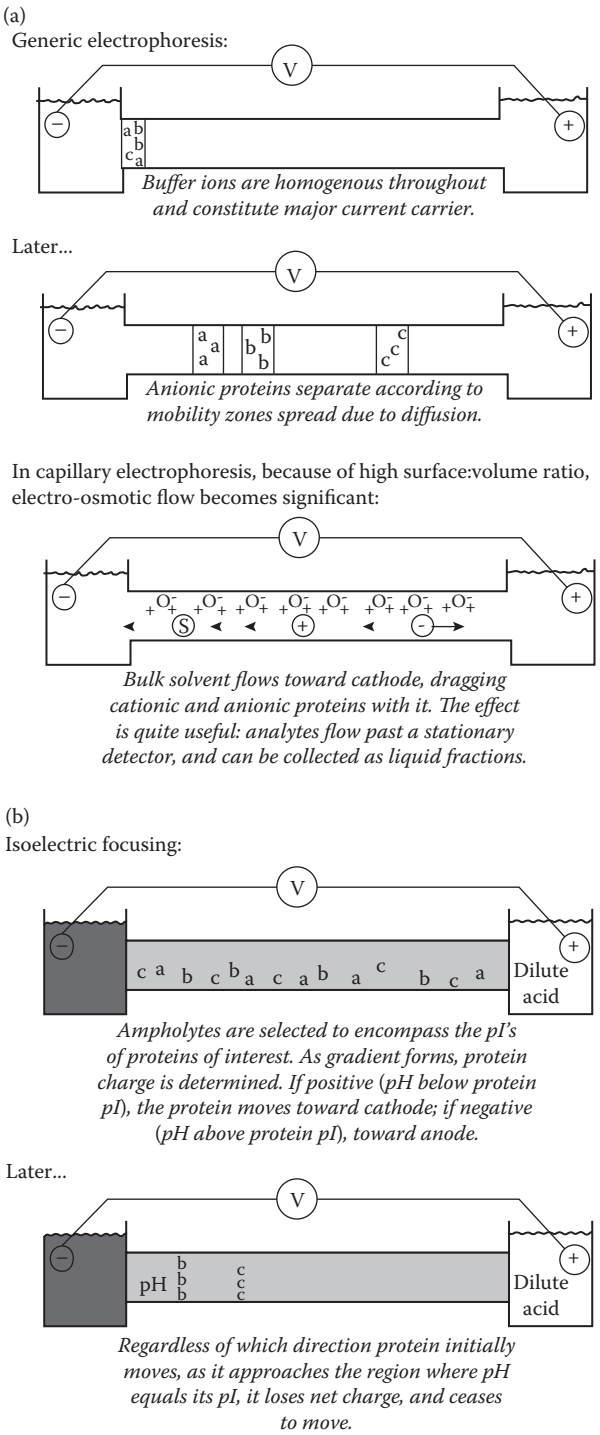


FIGURE 6.75 (a) Generic electrophoresis. (b) Isoelectric focusing (IOF).

Open Electrophoresis Electrokinetics can be exploited to guide and manipulate charged particles (DNA, RNA, PCR amplicons, polynucleotides, proteins, enzymes, antibodies, nanobeads, and even micron scale semiconductor devices) on a planar chip equipped with an array of closely spaced planar metal film electrodes.[61] The ability to produce well-defined electric fields with such electrode arrays (see, e.g., the Nanogen DNA array; Volume II, Chapter 8, Example 8.2) enables one to electrophoretically transport those charged particles to or from any microlocation on the surface of the device submerged in the analyte solution. This so-called open electrophoresis technique does not use fluidic conduits but closely spaced, individually addressable

electrodes. We explore this topic further below when using the Nanogen DNA array for dielectrophoresis (DEP) and in Volume III, Chapter 4 on packaging, assembly, and self-assembly, where we use the same chip for electric field-driven assembly.

Surface Electrophoresis As a further extension of the open electrophoresis concept, consider Wickramasinghe's use of AFM tips for molecular manipulation.[62] Applying a field between an AFM tip and a conductive surface, DNA electrophoresis over the shaft of the AFM tip was demonstrated at unprecedented speeds (10^4 times faster than CE) (Figure 6.76). We saw earlier that in free solution electrophoresis, DNA behaves as a free-draining coil, so that both charge and hydrodynamic drag scale linearly with molecular size (Equation 6.117). This is why DNA sequencing is not normally possible in free solution. Traditional DNA sequencing is done in a gel or polymer matrix, which presents a series of obstacles or obstructions, providing friction that depends on molecular size. Small DNA molecules migrate more easily through the pores of the gel, and thus have greater electrophoretic mobility than large DNA molecules. In the AFM case, electrophoretic separation of molecules such as DNA takes place in the nanometer-thick water films covering the shaft of the AFM tip (Figure 6.76A). In this layer, the Debye screening length is greater than the water film thickness, and surface charges are not screened totally by the water molecules. This might explain these encouraging results.

Electrokinetic Effects—Alternating Currents (AC)

Introduction

The AC electrokinetic effects we discuss here are *dielectrophoresis* (DEP), *electrorotation* (ROT), and *traveling wave dielectrophoresis* (TWD). Pohl introduced the term dielectrophoresis in 1951 to describe the motion of particles caused by dielectric polarization effects in nonuniform electric fields.[63,64] The effect was first recorded more than 2500 years ago, when it was discovered that rubbed amber attracts bits of fluff and other matter. *Electrorotation* is the rotation of particles in rotating electric fields, and traveling wave dielectrophoresis describes the translational movement of particles exposed to traveling electric fields.[65]

In electrophoresis, motion of a particle is determined by a net intrinsic electrical charge carried by that particle. In dielectrophoresis, on the other hand, motion is determined by the magnitude and polarity of charges induced in the particle by an applied field. In electrophoresis, direct current (DC) or low-frequency electric fields, usually homogeneous, are applied; in dielectrophoresis, on the other hand, alternating current (AC) fields of a very wide range of frequencies (in principle, there is no upper limit) are used, and the field must be inhomogeneous.[66]

The dielectrophoresis effect can be understood with reference to Figure 6.77. A particle placed in an electric field becomes electrically polarized as a result of partial charge separation, which leads to

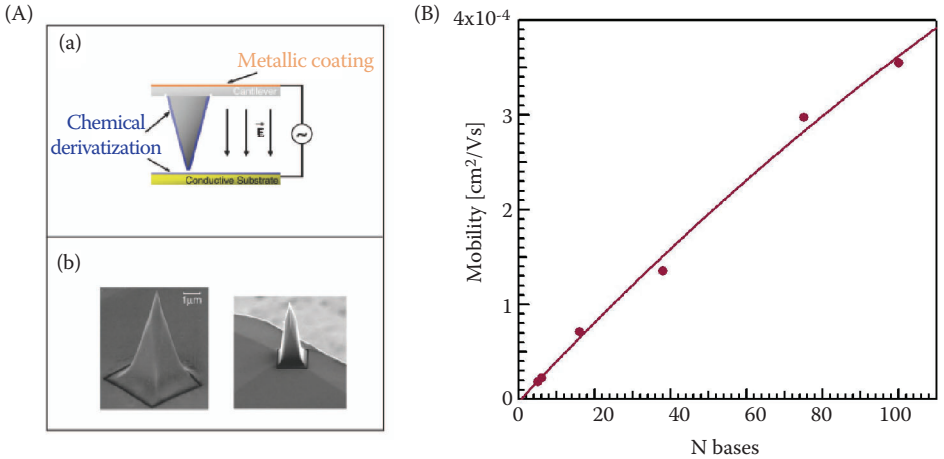


FIGURE 6.76 DNA electrophoresis. (A) (a) AFM setup and (b) AFM tips. (B) DNA mobility as a function of base pair number.

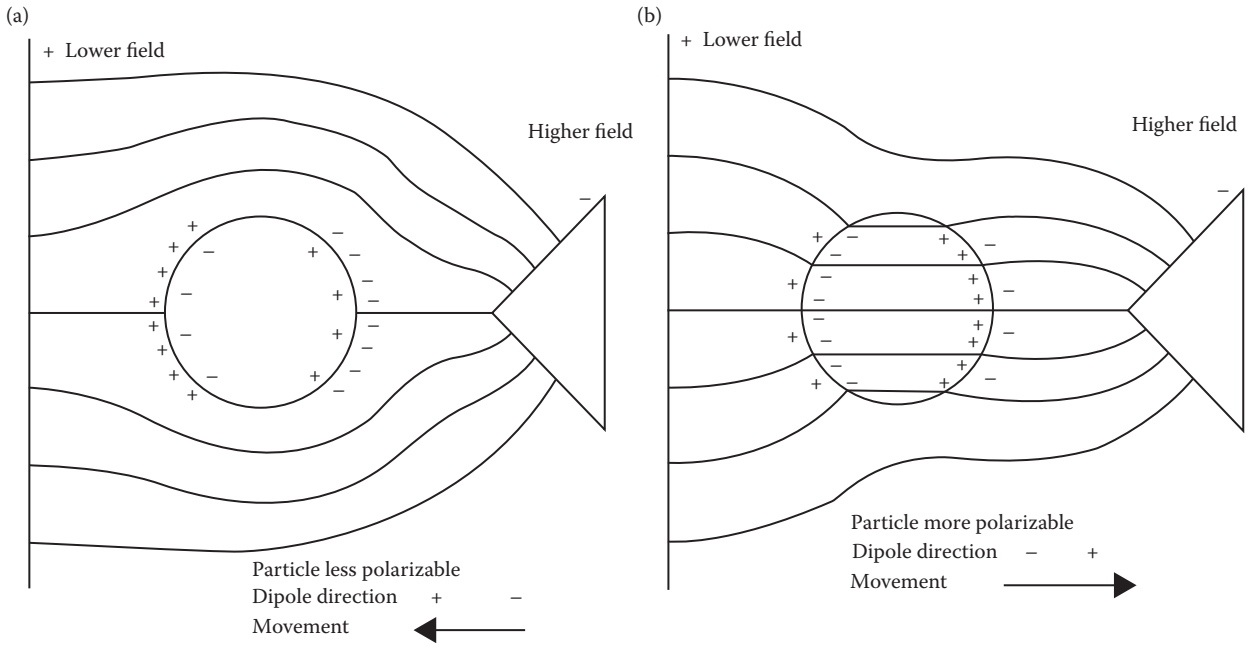


FIGURE 6.77 Negative (a) and positive (b) dielectrophoresis.

an induced dipole moment. The dipole moment is a consequence of the generation of equal and opposite charges ($+q$ and $-q$) at the boundary of the particle. This induced surface charge is only about 0.1% of the net surface charge normally carried by cells and microorganisms and is generated within about a microsecond. In a nonuniform electric field, the field on one side of the particle will be stronger than that on the other side, and a net dielectric force, F_{DEP}, is exerted. The magnitude of the induced dipole depends on the polarizability of the particle with respect to that of the medium. If a suspended particle has polarizability higher than the medium, the dielectrophoresis force will push the particle toward regions of higher electric field in so-called "positive dielectrophoresis." In negative dielectrophoresis, the medium has a higher polarizability than the suspended particles, and the particles are driven toward regions of low field strength. Dielectrophoresis causes a motion of neutral matter by polarization of the neutral matter in a nonuniform electric field. The method is particularly suited for the separation, aggregation, selective trapping, manipulation, and identification of cells and microorganisms.[66]

Dielectrophoresis Force

Spherical particles of radius r and complex permittivity ε_p, suspended in a fluid of absolute complex dielectric permittivity ε_m, are subject to a dielectric force given by:

$$\mathbf{F}_{DEP} = 2\pi r^3 \varepsilon_m \alpha_r \nabla \mathbf{E}^2 \qquad (6.128)$$

where $\nabla \mathbf{E}^2$ = the gradient of the square of the electric field (rms, or root mean square value) quantifying the nonuniformity of the electric field

∇ = the del vector operator

α_r = the real component of the Clausius-Mossotti factor, that is, the effective polarizability of the particle with respect to its suspending medium; the latter term may thus be written as:

$$\alpha_r = \mathrm{Re}\left(\frac{\varepsilon_p^* - \varepsilon_m^*}{\varepsilon_p^* + 2\varepsilon_m^*}\right) \qquad (6.129)$$

In the above expression:

Re = "the real part of"

ε_p^* and ε_m^* = the complex dielectric properties of the particle and its medium, with $\varepsilon^* = \varepsilon - j\sigma/\omega$, in which $j = \sqrt{-1}$

ε = permittivity

σ = conductivity

ω = angular frequency of the applied field ($\omega = 2\pi f$)

For a vacuum, which has no mobile charges, ε is 1, and for a conductor, it is infinite. Because $\varepsilon_p^* - \varepsilon_m^*$ can be positive or negative, depending on the relative magnitudes of ε_p^* and ε_m^*, controlled movement from and to areas of high electric field strengths is possible. In other words, a material with a higher dielectric constant will experience a force tending to move it into a stronger electric field, displacing a material with a lower dielectric constant in the process. A positive value for α_r in Equation 6.129 implies a positive DEP force; a negative value produces a negative DEP. Because the field $\boldsymbol{E}$ in Equation 6.128 appears as $\nabla \boldsymbol{E}^2$, reversing the bias does not reverse the DEP force. Ac voltages in a frequency range from 500 Hz to 50 MHz have been used. Based on Equation 6.129, one recognizes that the factor α_r is frequency-dependent. At low frequencies polarizability mainly is determined by the conductivity, and at high frequencies it is determined by the permittivity. Theoretically, the value of α_r may be between +1.0 and −0.5. Key in dielectrophoresis is the presence of a material of substantially different dielectric constant from its surroundings. In an aqueous conductive solution, ions shield the dielectric material from the external applied field. Varying the field in time reduces the shielding effect of the ions somewhat. If the mobile dielectric materials are embedded in a sufficiently low resistance medium, then the dielectrophoresis effect can be much more readily observed. For a rigorous derivation of Equation 6.128, we refer to Huang et al.[67] and Wang et al.[68]

Using DEP, individual biological cells have been moved and positioned by altering the ac frequency so as to switch from a "capture" mode using positive DEP to "release" mode using negative DEP.[69] Particles can also be separated using DEP. This is illustrated in Figures 6.78 and 6.79. Figure 6.78a reveals how cell type A switches from negative to positive DEP at a changeover frequency of about 10^5 Hz. From Figure 6.78b we learn that there is a frequency range above 10^5 Hz where cell type B displays negative dielectrophoresis, whereas A exhibits positive dielectrophoresis. This means that one cell type will move to higher electric field regions, while the other cell type will do the opposite, effectively separating the two types of cells. This effect is exploited in Figure 6.79, where we separate *Listeria* from whole blood. An array of 100-planar Pt electrodes [the Nanogen chip (see Volume II, Chapter 8, Example 8.2)] is used in this experiment. In the first panel, blood contaminated with *Listeria* is placed over the chip. In the second panel, a 10-kHz, 10-Vpp ac signal is applied, driving the red blood cells to low electric field regions (i.e., between the Pt electrodes) and the *Listeria* to high electric field regions (i.e., on the Pt electrodes).

In the third panel we wash off the red blood cells, leaving the *Listeria* highly enriched on the Pt electrodes where they stick better than the red blood cells on the insulating regions (Si_3N_4) between the Pt electrodes. In the fourth panel only *Listeria* is left on the chip. Finally, in the fifth panel, we release the *Listeria* by removing the bias and washing again. In Table 6.7 we summarize some pros and cons of using dielectrophoresis for particle separation.

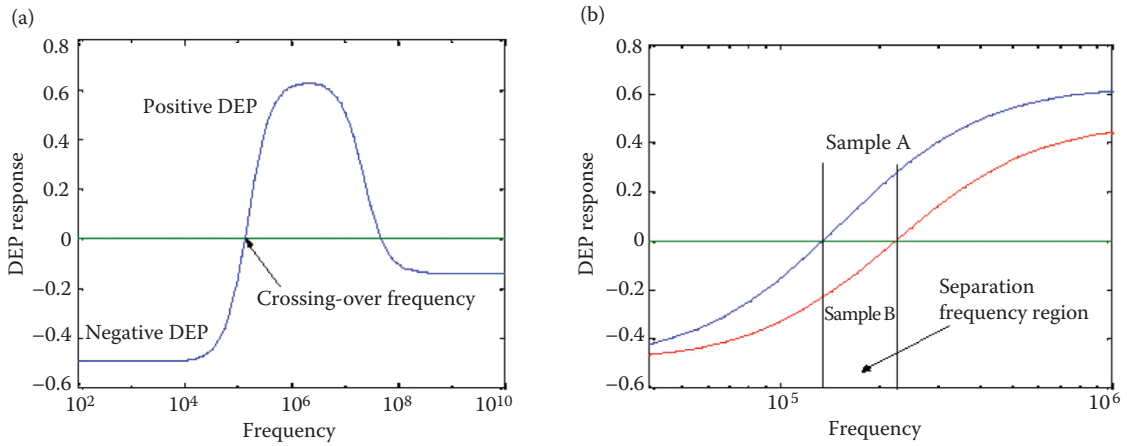


FIGURE 6.78 Dielectrophoresis: how to separate particle A from particle B. (a) Cell type A switches from negative to positive DEP at a change over frequency of about 10^5 Hz. (b) Above 10^5 Hz cell type B displays negative DEP, whereas A exhibits positive DEP.

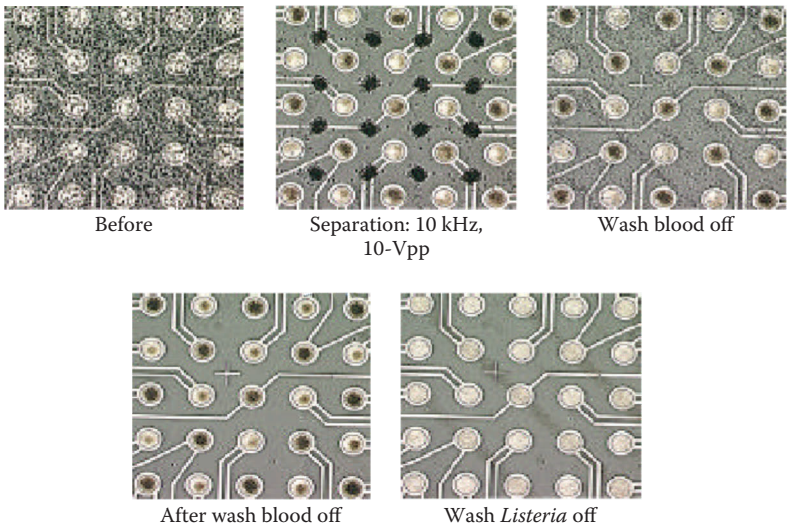


FIGURE 6.79 Dielectrophoretic separation of *Listeria* from whole blood.

One problem associated with traditional methods of DEP separation highlighted in Table 6.7 is that the DEP force, which is proportional to ∇E^2, rapidly decays as the distance from the planar electrodes increases. As a consequence, a common problem to most current dielectrophoresis devices using planar electrodes is the incapability to achieve high throughputs and high efficiencies. With a given flow channel cross-section, many targeted particles might flow over the planar electrodes without experiencing a force, its mean distance to the electrode surface being too long, making it necessary to reflow the same sample several times. This has prevented DEP from being used in high-volume applications. To correct this problem, the author's University of California, Irvine, team is using 3D electrodes to reduce the mean distance of any targeted particle contained in a channel or chamber to the closest electrode surface, as illustrated in Figure 6.80. Numerical simulations demonstrating such advantage were obtained by Park,[70] and the experimental verification was carried out by Martinez-Duarte.[71]

The 3D electrodes are implemented using the C-MEMS process detailed in Volume III, Chapter 5 (page 266 "Microfabrication of Carbon: C-MEMS Process"). In 3D C-MEMS dielectrophoresis, carbon 3D structures up to several hundreds of micrometers high are implemented as electrodes in the flow channel. An example of 65-μm-high posts is shown in Figure 6.81.

The difference in trapping efficiency of yeast cells, when only flowing the sample once, between 2D and 3D electrodes increases as flow rate increases, with the 2D efficiency always being lower as shown in Figure 6.82.

Microfabricated 3D electrodes also have been implemented by Wang[72] and Voldman,[73] who used electroplated gold electrodes for particle separation, applying the lateral field established by the electrodes that form the channel side walls and for particle trapping for cytometry purposes, respectively. More

TABLE 6.7 Pros and Cons for Particle Separation by Dielectrophoresis

Pros	Cons
Intrinsic dielectric properties are taken advantage of	Speed: clinical sample requires ml volume (1)
No immunolabeling involved	Throughput: can only separate two types of cells (2)
No narrow channels to clog	Selectivity/sensitivity
Simple and low cost	
Easy to scale down or scale up	
Can move even semiconductor components	

1) This problem may be solved using 3D volumetric DEP methods (described next).
2) This problem could be solved by using an array of signal generators.

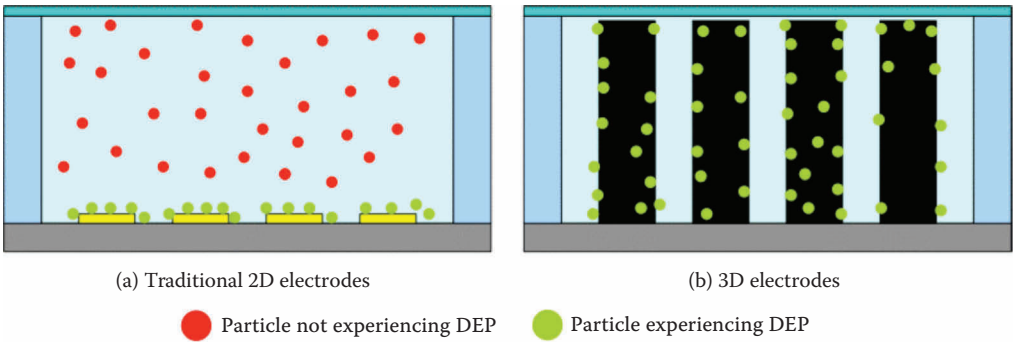


FIGURE 6.80 (a) In traditional 2D DEP, many particles might flow by without experiencing any DEP force. (b) In 3D DEP, the mean distance from the particle to the electrode surface decreases; thus, the number of particles experiencing DEP force greatly increases.

recently, Iliescu[74] has used complex microfabrication techniques to obtain 3D doped silicon electrodes in a multistep process. In a nontraditional way, Lapizco-Encinas et al.[75] have induced DEP on bacteria and proteins using insulator volumetric microelectrodes, in a technique that has become known as insulator DEP. The main concept behind this approach is the use of two metal electrodes, excited by thousands of volts, to generate a uniform electric field between them, which is then disrupted and turned nonuniform by the insulator structures in between. Another truly 3D DEP approach is that of Fatoyinbo et al.,[76] who use a drilled conductor-insulator-conductor laminate to obtain wells with electrodes all along the top and bottom surfaces of the laminate.

The final goal is to improve throughput and efficiency in DEP devices for a commercial product featuring high selectivity at high flow rates. Immediate applications might include cell enrichment for cell therapy, sample purification, drug screening, water analysis, and many other applications may become possible in the future. Throughput refers to inducing DEP on as many particles as possible without consideration of selectivity between targeted and nontargeted particles. This is also known as *maximal retention efficiency*.[77] Selectivity refers to how well one can induce DEP to a targeted particle without affecting nontargeted ones. It is also referred to as *maximal purification efficiency*.[77]

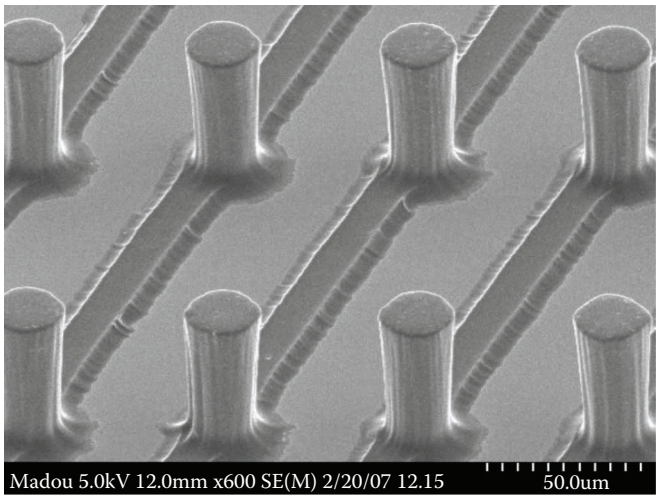


FIGURE 6.81 Carbon posts (65 μm high) for use as electrodes in C-MEMS dielectrophoresis (C-MEMS DEP).

Traveling Wave Dielectrophoresis

If, instead of using a stationary electric field, as is used in DEP, a particle is subjected to a moving electric field, the particle can be moved by an effect known as traveling wave dielectrophoresis, or TWD. To produce these traveling waves, electrode

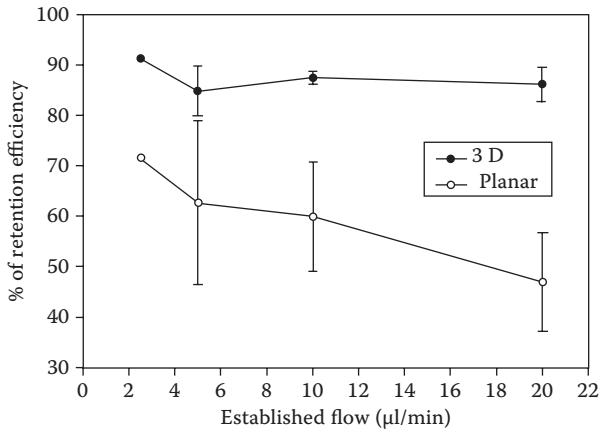


FIGURE 6.82 Higher retention efficiency at all flow rates are obtained with 3D DEP compared with 2D DEP. The difference between them sharply increases with flow rate. (Martinez-Duarte, R., H. A. Rouabah, N. G. Green, M. Madou, and H. Morgan. 2007. *Proceedings of the Eleventh International Conference in Miniaturized Systems for Chemistry and Life Sciences: microTAS 2007.* Paris.)

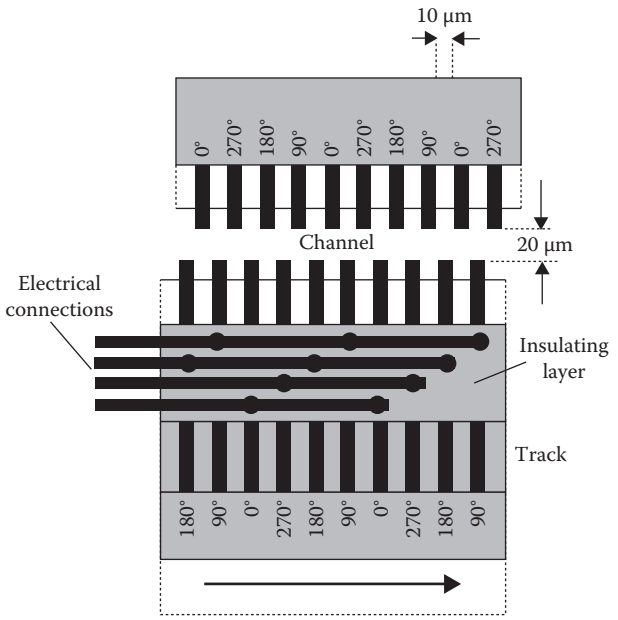


FIGURE 6.83 TWD setup. Electrode structures for producing traveling electric fields along channels or tracks. When addressed with sinusoidal voltages of the phase sequences shown, a traveling field is established by the propagation direction indicated by the arrow. The characteristic dimensions shown are suitable for manipulating yeast, erythrocytes, and parasites such as *Cryptosporidium*, for example. Correspondingly larger or smaller dimensions are required for larger bioparticles (e.g., white blood cells, plant cells, pollen) or smaller ones (e.g., bacteria, viruses), respectively. (Burt, J. P. H., R. Pethig, and M. S. Talary. 1998. Microelectrode devices for manipulating and analysing bioparticles. *Trans Inst MC* 19:1–9. With permission.)

geometries with tracks of electrodes in a channel, as shown in Figure 6.83, may be used. In this arrangement, electrodes are fabricated on a bottom-insulating substrate, and a second insulator layer is used to define a channel along the length of the electrode track. Another set of four electrodes is then deposited on this second insulator layer, and vias are used to link every fourth electrode together and to provide external connections. The four electrodes are addressed with sinusoidal voltages of 90° phase separation (0°, 90°, 180°, and 270°). The used phase quadrature voltages are equal in amplitude (around 1 to 5 V peak-peak) and of a frequency range from about 1 kHz to 10 MHz. TWD provides conveyor tracks over which particles, suspended by negative dielectrophoresis, are moved. By changing the magnitude and frequency of the energizing voltages, the speed and direction of particles in the channel can be controlled.[65]

TWD, in combination with DEP, enables selective and nonselective trapping of particles. The selective trap only holds desirable subpopulations, whereas a nonselective trap retains all particles from a fluid flow. Particles moving in a channel by TWD may be captured nonselectively and held immobile by suddenly changing the frequency and phases of the applied voltages so as to encourage positive dielectrophoresis. By returning to a TWD-type of field, all the particles, or a subset of them, can be made to resume their movement along the channel. In the case of selective trapping, the TWD remains in effect, and trapping is accomplished by the choice of the electric field frequency. The frequency may be chosen such that some particles are levitated by negative dielectrophoresis and thus keep on moving under the influence of the TWD, whereas others are captured by positive dielectrophoresis and are immobilized at the electrode surfaces.[65]

Particle Levitation

The gravitational settling force pulls down a spherical particle with radius r and mass density γ_1, suspended in a fluid with mass density γ_2, and is given by:

$$\mathbf{F}_g = \frac{4}{3}\pi r^3(\gamma_2 - \gamma_1)\mathbf{g} \qquad (6.130)$$

where $\mathbf{g}$ is the gravitational acceleration vector. A particle will hover above the electrode surface when the opposing gravitational and negative dielectrophoretic forces balance:

$$\mathbf{F}_{DEP} + \mathbf{F}_g = 0 \qquad (6.131)$$

Or, from Equations 6.128 and 6.130:

$$2(\gamma_2 - \gamma_1)\mathbf{g} = -3\,\varepsilon_m \alpha_r\,\nabla \mathbf{E}^2 \qquad (6.132)$$

From Equation 6.132, levitation is independent of particle size, as both DEP and gravitational force scale as l^3. Levitation height only depends on the intrinsic properties of the particle, such as permittivity, conductivity, and mass density. By working at very high frequencies, where the polarizability factor primarily depends on the permittivity rather than the conductivity, levitation may also be made independent of the conductivity of the solution. A quantitative

estimate of the levitation height requires knowledge of how the factor ∇E^2 varies as a function of height h above the electrode plane. Rousselet et al.[79] derived the following relationship between ∇E^2 and h:

$$\nabla E^2 = AV^2 \exp(-hk) \qquad (6.133)$$

where V = applied voltage
A = constant
k = another constant inversely proportional to the width of the electrodes[79]

Combining this expression with Equation 6.132, the expression for stable levitation height is derived as:

$$h = \frac{1}{k}\ln\left[\frac{-3\,\varepsilon_m \alpha_r\, AV^2}{2(\gamma_2 - \gamma_1)g}\right] \qquad (6.134)$$

This type of dielectrophoretic levitation was demonstrated, for example, with latex beads and interdigitated electrodes of widths and separations ranging from 5 to 40 μm, with ac signals of 0–10 V (rms) in the frequency range 1 kHz to 10 MHz. Maximum levitation was observed at 1 MHz. At this frequency it was confirmed experimentally that levitation height was independent of particle size and solution conductivity.[79] It also was shown that red blood cells could be separated from latex beads easily. Whereas the erythrocytes were attracted to the electrodes by positive dielectrophoresis, the latex particles levitated and could be separated by flowing liquid over the surface. By de-energizing the electric field, the red blood cells also could subsequently be removed by flowing more liquid.[79]

Particles possessing different dielectric properties levitate at different heights in a flat field-flow-fractionation (FFF) chamber through the balance of DEP and gravitational forces.[68] The different particles are levitated into streamlines with different velocities, as shown in the parabolic flow profile illustrated in Figure 6.84, and can be separated this way. Even continuous separation of particles is possible.[80]

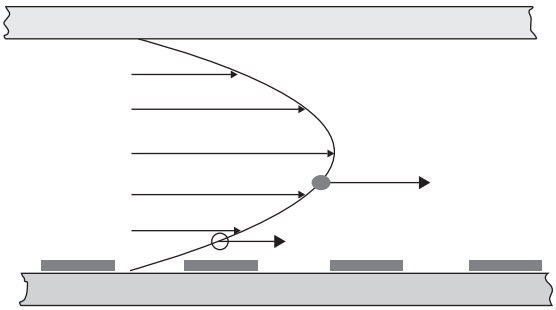


FIGURE 6.84 Field-flow fractionation (FFF). Outline of the combined DEP-FFF separation procedure. Particles with different dielectric properties or density are levitated by negative dielectrophoresis to different planes of the parabolic velocity profile of the liquid flowing through the chamber. (Rousselet, J., G. H. Markx, and R. Pethig. 1998. Separation of erythrocytes and latex beads by dielectrophoretic levitation and hyperlayer field-flow fractionation. *Colloids Surfaces* A 140:209–16. With permission.)

Electrorotation

Electrorotation (ROT) is the imaginary component of the ac electrokinetic effect, whereas its real component is the above-introduced dielectrophoresis (DEP). The rotating electric field usually is generated using four electrodes positioned at right angles to one another and energized with phase-quadrature signals of frequencies between 50 Hz and 100 MHz. The ROT torque acting on a particle positioned in the center of this electrode assembly is given by:

$$T_{ROT} = -4\pi\, r^3 E_m \alpha_i\, E^2 \qquad (6.135)$$

where α_i is the "imaginary part" of the Clausius-Mossotti factor. As with impedance spectroscopy, there is a direct relationship between the real and imaginary components of the ac electrokinetic effect, so that the electrorotation frequency spectrum can be calculated from the measured dielectrophoresis response, and vice versa. In Figure 6.85 the frequency variations of α_r and α_i for viable and nonviable yeast cells are shown. The polarity and peak values of α_i are determined by the rate of change of α_r. The frequency dependence of magnitude and polarity of α_r and α_i is the basis of the three ac electrokinetic applications discussed here.

Rotation analysis chambers in a biofactory-on-a-chip are used to monitor dielectric properties of particles.[65] Electrorotation has been shown to be a sensitive method for monitoring the physiological viability of cells.[82]

Scaling Considerations

Depending on particle size, different forces compete with the DEP force. For example, a typical biological cell with a radius of 1 μm and a density of

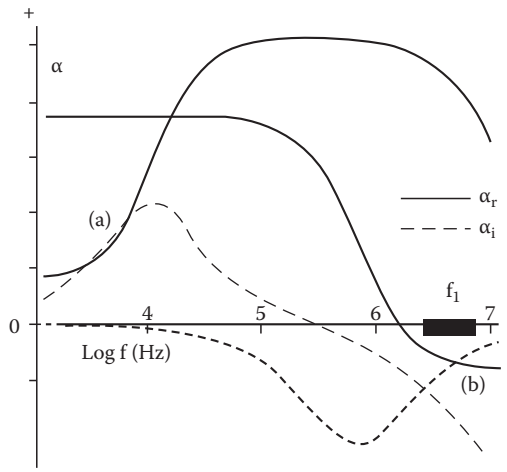


FIGURE 6.85 Alpha plots. The frequency variations of the real and imaginary polarizability parameters α_r and α_i of Equations 6.128 and 6.135 for viable (a) and nonviable (b) yeast cells in a suspending medium of conductivity 0.8 mS/m. Only the nonviable cells will exhibit traveling field dielectrophoresis (in the frequency range f_1) and thus the viable cells will remain immobilized at the electrodes. (Burt, J. P. H., R. Pethig, and M. S. Talary. 1998. Microelectrode devices for manipulating and analysing bioparticles. *Trans Inst MC* 19:1–9; and Talary, M. S., J. P. H. Burt, J. A. Tame, and R. Pethig. 1996. Electromanipulation and separation of cells using travelling electric fields. *J Phys D Appl Phys* 29:2198–203. With permission).

1.05 kg m^{-3} suspended in an aqueous solution at room temperature feels both the gravitational and the Brownian force of around 2×10^{-15} N each. For the DEP force to dominate, it should be at least 10 times larger than these competing forces; to affect this, ∇E^2 must be around 9×10^{12} V^2m^{-3} (with an α of 0.5).[66] Both the DEP and the gravitational effect scale as l^3, and, with increasing particle size, no scaling advantage of one over the other emerges. With submicrometer particles, on the other hand, the randomizing Brownian effect, being proportional to l^{-1}, becomes the dominating force competing with DEP. With particles below 0.1 μm, for example, values for ∇E^2 greater than 10^{17} V^2m^{-3} are required to make the DEP force 10 times larger than the Brownian. This requires field strengths in excess of 10^6 Vm^{-1}, and at a suspending medium conductivity higher than 0.1 Sm^{-1}, these high fields cause local heating, leading to hydrodynamic effects. The latter interfere with the DEP effect of submicron particles significantly.[66] Moreover, at these fields cell damage might become an issue. Müller et al.[83] used submicron-scale electrodes to achieve DEP levitation and trapping of particles as small as 14 nm, but, as predicted, hydrodynamic effects interfered significantly.

Downscaling of the electrodes is very favorable in dielectrophoresis. By using thin film metal electrodes rather than wire or pin electrodes, one approaches the size of the typical particles to be separated. In Figure 6.86, the expression for ∇E^2 for the specific represented electrode configuration is given. Maintaining the applied voltage at 1 V rms and reducing the radius of curvature 100-fold produces a 1000-fold increase in ∇E^2. Lower voltages can thus be used to produce the same DEP force on a particle in the same relative position. Despite the small cross-sectional area of the electrodes, the volume of liquid they energize is very large by comparison, so that heat produced by the electrical current is efficiently dissipated, and, moreover, surface electrochemical processes are reduced.[66]

Applications

In a 1983 patent, Batchelder[84] describes the use of dielectric forces to selectively position, transport,

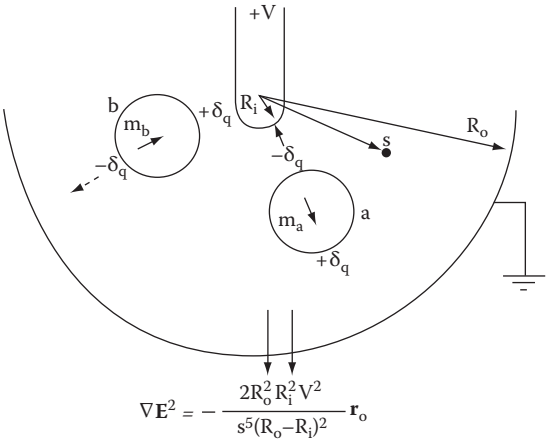


FIGURE 6.86 Two particles located in a nonuniform electric field, generated in this example by electrodes of spherical geometry and radii of curvature R_i and R_o. Different dipole moments *m* (depicted as separated microcharges *q*) are induced in each particle. Particle *a* is more polarizable than the surrounding medium and is directed by positive dielectrophoresis toward the high field at the inner electrode, whereas particle *b* is weakly polarizable and moves toward the low-field region at the outer electrode under the influence of negative dielectrophoresis. The force causing this motion is proportional to ∇E^2, described by the equation in the figure, where *s* is the distance of the particle from the inner electrode, *V* is the applied (rms) voltage, and r_0 is a unit-radius vector. (Pethig, R., and G. H. Markx. 1997. Applications of dielectrophoresis in biotechnology. *TIBTECH* 15:426–32. With permission.)

mix, and react one or more chemicals within a reaction chamber. The Batchelder patent was ahead of its time in terms of vision for microfluidics and constituted a laboratory-on-a-chip (LOC) long before academics introduced the buzzword. The emphasis in this patent is on the dielectrophoresis mechanism, and little is said about micromachining. Although metal electrodes are patterned with lithography techniques, the reaction chambers are made with more traditional manufacturing methods.

Once microelectrode technology and MEMS made their entry, it became apparent that dielectrophoresis would be a promising tool for the selective manipulation and separation of cells, bacteria, viruses, and perhaps even biomacromolecules such as DNA and proteins. A limitation may be cell damage, electroporation, and electrofusion, which start at field strengths of $1–3 \times 10^5$ Vm^{-1}. However, in dielectrophoresis, the field is rarely higher than 10^5 Vm^{-1}. In practice, no irreversible cell damage has been reported in DEP experiments.[66] Another possible limitation is that of particles being smaller than 10 nm; these might be too difficult to manipulate because of the hydrodynamic effects interfering with the DEP force.

Batch and continuous separation of particles has been achieved through positive and negative DEP in FFF chambers. A typical DEP-FFF separation chamber is shown in Figure 6.87 and consists of two glass plates spaced a small distance (typically <300 μm) apart. The inner surfaces of the glass plates are provided with interdigitated electrodes.[80,85] Working in the frequency range marked f_1 in Figure 6.85 separates viable and nonviable yeast cells. The nonviable cells levitate in this frequency range and are transported with the fluid flow, whereas the viable ones remain immobilized at the electrodes.[81] As just one more example of the use of this type of separation, we refer to the removal of human breast cancer cells from hematopoietic $CD34^+$ stem cells. The $CD34^+$ stem cells are levitated higher and carried faster by the fluid flow and exit the separation chamber earlier than the cancer cells. Using on-line flow cytometry, efficient separation of the cell mixture is observed in less than 12 min, and $CD34^+$ stem cell fractions with a purity >99.2% are obtained.[85]

Electrorotation is used, for example, to determine the dielectric properties of human leukocyte subpopulations.[86] Pethig et al.[65] are experimenting with combinations of DEP, ROT, and TWD elements on a single substrate to make a so-called biofactory-on-a-chip.

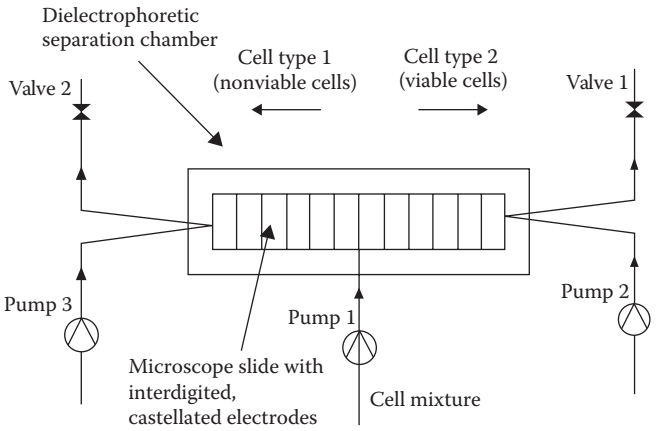


FIGURE 6.87 Outline of a continuous dielectrophoretic separation system. Valves, frequency generators, and pumps are all under computer control. (Markx, G. H., and R. Pethig. 1995. Dielectrophoretic separation of cells: continuous separation. *Biotech Bioeng* 45:337–43. With permission.)

Electrowetting

Electrical control of interfacial tension between a liquid and a solid provides another means of fluid pumping without moving mechanical parts. Altering the apparent surface tension at the solid/liquid interface becomes possible by applying a potential over the interface. This phenomenon is known as electrowetting and is described by the Lippmann equation:

$$\gamma_{SL} = \gamma_{SL}^{max} - \frac{\varepsilon_r \varepsilon_0}{2\delta}(V - V_{pzc})^2 \qquad (6.136)$$

where V_{pzc} represents the potential of zero charge for the solid/liquid (electrolyte) interface, and γ_{SL}^{max} is the surface tension corresponding to $V = V_{pzc}$. The thickness of the diffusion layer in the solution is given by δ, and ε_r and ε_0 are the relative permittivity of the solution and the permittivity of vacuum, respectively. A voltage applied across the interface will alter the surface tension and cause the liquid to move within the capillaries. The flow velocities achieved with electrocapillary pressure are several centimeters per second or nearly two orders of magnitude higher than the velocities achieved by electro-osmotic pumping.[87]

Surface tension becomes important only when dimensions drop well below 1 mm. As in the case of electro-osmotic pumping, subtle uncontrollable changes at the solid/electrolyte interface make this actuation principle difficult to control and electrolyte specific.

Light has been used to control the wettability of the surface of certain photoresponsive materials. The exact mechanisms of the surface wettability changes of these materials on illumination still are being debated; however, it is generally accepted that the effect is caused by a photoelectrochemical reaction at the surface.[88] With ZnO, UV irradiation is responsible for generating electron-hole pairs on its surface. These holes react with the lattice oxygen to form surface oxygen vacancies in a process described by Zhang et al.[89] in the following equations:

$$ZnO + 2h\nu \rightarrow 2h^+ + 2e^-$$

$$O^{2-} + h^+ \rightarrow O^-$$

$$O^- + h^- \rightarrow \frac{1}{2}O_2 + V_0 \text{ (oxygen vacancy)} \quad \text{Reaction 6.1}$$

In ambient air, water and oxygen from the surroundings compete with lattice oxygen to dissociatively absorb on those vacancies. In solution the defect sites, however, kinetically favor the absorption of hydroxyl groups over oxygen,[90] and with the surface attracting more hydroxyl (–OH) groups, the surface becomes more hydrophilic.

This switch from a hydrophobic to a hydrophilic surface has been extensively documented. Sun et al.,[91] for example, showed changes of contact angle from 109° before to 5° after UV exposure for a droplet of water on a ZnO surface.

Other wide bandgap semiconductor oxides, such as Ta_2O_5, TiO_2, indium–tin oxide, $InTaO_4$, and In_2O_3, also exhibit optowetting. These materials reverse their characteristic and tend to go back to their hydrophobic state if the UV-exposed sample is placed in a dark room; over time the hydroxyl groups absorbed on the defect sites are replaced by oxygen atoms. This is because the surface is thermodynamically unstable after the hydroxyl absorption, and oxygen is more strongly bonded on the defect site than a hydroxyl group.

Because there is some degree of reversibility, these semiconductors show promise as a wettability switch for future microfluidic applications.

Centrifugal Fluidic Platform—CD Fluidics

Introduction

An attractive approach to fluid propulsion is centrifugation in fluidic channels and reservoirs crafted in a CD-like plastic substrate as shown in Figure 6.88. The centrifugal force caused by the rotation of the CD results in the release and flow of reagents and analytes. No external pumps and valves are needed—fluidic flow can be controlled by the spinning rate. The centrifugal force generated in such an instrument overcomes (depending on the rotation speed) the capillary forces in the fluidic network and moves fluid in a valveless manner from reservoir chamber

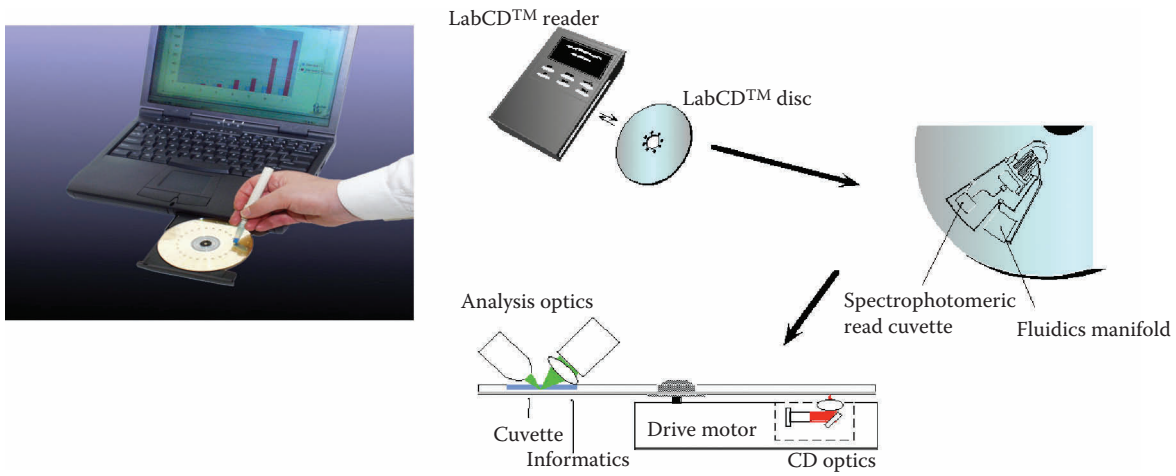


FIGURE 6.88 Centrifugal platform on a CD. The name LabCD™ was introduced by Gamera, the first fluidic CD company (now defunct).

to reaction chamber and eventually to a waste site. A whole range of fluidic functions, including valving, mixing, metering, sample splitting, and sample separation, has been implemented this way.[92] Such analytical functions make for a laboratory-on-a-disc with applications in diagnostics and drug discovery. Analytical measurements may be fluorescent or absorption based, and informatics embedded on the same disc could provide test-specific information (merging of fluidics with informatics).

Prototyping methods of this microfluidic platform are covered in Volume III, Chapter 5 on selected new materials and processes for MEMS and NEMS. In Volume II, Chapter 10 on replication techniques and LIGA, we cover manufacturing of the fluidic CD. The aspect of the CD platform we are interested in here is the physics behind its operation and the comparison of CD fluidics to mechanical, osmotic, and acoustic pumping.

Centrifugal Platform Operation

In a centrifugal platform, the centrifugal force provides the pumping force, whereas hydrophilic and hydrophobic valves are used to inhibit flow. In Figure 6.89a, we show a schematic of a centrifugal platform together with relevant parameters; Figure 6.89b and c demonstrates two types of hydrophobic valves, and Figure 6.89d is a hydrophilic or capillary valve. For a hydrophobic CD surface, to make a valve, a flow restriction can be made in the flow channel by reducing its radius from r_1 to r_2. A positive pressure is then needed to push the fluid across this restriction. From Equation 6.96, the pressure p is determined by r, the radius of the capillary, and the required pressure Δp to overcome a sudden narrowing in capillary tubing is calculated from Equation 6.96 as:

$$\Delta p = 2\sigma_L \cos\theta_c \left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{6.137}$$

Equation 6.137 is independent of the length of the restriction and flow rate. Once the capillary downstream is wetted, the pressure needed to continue flow is given by Equation 6.54 (Hagen-Poiseuille's Law). These hydrophobic restrictions can be used as passive valves in fluidic networks: by setting the pumping pressure below the pressure associated with the restriction, no flow will take place unless the pumping pressure is increased. Mixers, diluters, sample splitters, and sample integrators have been built using passive hydrophobic valving in combination with an external mechanical pump.[93] Of course, the same principles can be applied in combination with centrifugal, electrokinetic, or acoustic pumping. As shown in Figure 6.89c, one can also create a hydrophobic valve by locally changing the hydrophobicity of the surface, i.e., by locally increasing θ_c. In the case of a contact angle less than 90° (a hydrophilic surface), to make a valve the capillary is widened, which leads to the need of increased pressure as more energy is required now to wet a larger surface area (Figure 6.89d).

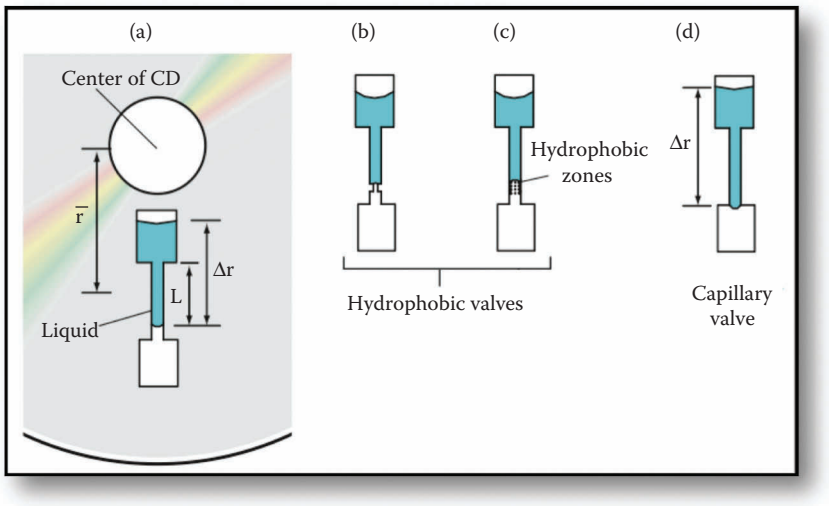


FIGURE 6.89 Centrifugal platform on a CD. (a) CD platform, (b and c) hydrophobic valves, and (d) hydrophilic capillary valve.

Assuming a capillary valve, when spinning the CD, the two opposing forces at work can be described as follows (see Figures 6.89a and 6.90).

The pumping force (P_c) caused by spinning of the disc is given by:

$$\frac{dP_c}{dr} = \rho\omega^2 r \qquad (6.138)$$

where ρ = the density of the liquid
ω = the angular velocity of the CD platform
r = the distance between a liquid element and the center of the CD

Integration of Equation 6.138 from $r = R_1$ to $r = R_2$ gives:

$$\Delta P_c = \rho\omega^2 (R_2 - R_1)\frac{R_2 + R_1}{2} = \rho\omega^2 \Delta R \cdot \bar{R} \qquad (6.139)$$

The capillary force (P_s) caused by interface tension is given by:

$$\Delta P_s = \frac{\gamma \cos\theta_c \cdot C}{A} \qquad (6.140)$$

where γ = the interface tension
θ_c = the contact angle (in a practical example θ_c may be 61.5°)
A = the cross-section area of the capillary
C = the associated contact line length

The burst frequency is the frequency at which the fluidics are released from their reservoirs. The liquid will be released when ΔP_c is greater than ΔP_s; consequently, the burst frequency calculated from Equations 6.139 and 6.140 is given by:

$$f_b \geq \left(\frac{\gamma \cos\theta_c \cdot C}{\pi^2 \rho \cdot \bar{R} \cdot \Delta R \cdot 4A}\right)^{\frac{1}{2}} = \left(\frac{\gamma \cos\theta_c}{\pi^2 \rho \cdot \bar{R} \cdot \Delta R \cdot d_H}\right)^{\frac{1}{2}} \qquad (6.141)$$

Here d_H is the so-called hydrodynamic diameter of the channel.

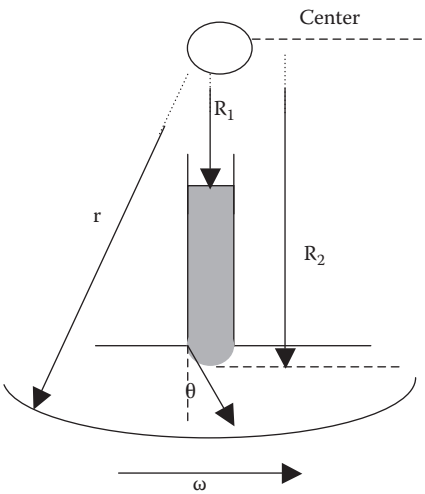


FIGURE 6.90 Schematic illustration of fluid propulsion in centrifugal microfluidics.

Example 6.8: Two-Point Calibration Disc

As an early first example of the use for the CD fluidic platform, the author and his team, by appropriately choosing the channel diameter and reservoir dimensions and locations, designed an automated two-point optrode calibration system, capable of five sequential fluid movement steps as shown in Figure 6.91a and b. The liquid flows in the order of calibrant 1, wash 1, calibrant 2, wash 2, and sample to an optrode chamber by increasing the rotating speed (see Figure 6.91a for the placement of these fluidic chambers, top figure is an actual photograph of the CD and the bottom is a schematic). Figure 6.91b shows the calculated burst frequencies of the reservoirs in the two-point calibration platform. The same optrode chamber is used for measuring calibrants and sample to eliminate artificial system errors common with devices using separate chambers for measuring sample and calibrants.

To demonstrate the functionality of this two-point calibration system, we deposited a potassium ion selective membrane in the optrode chamber using a microdrop delivery setup. The membrane was composed of poly(vinyl chloride) (PVC), valinomycin (ionophore), dioctyl sebacate (plasticizer), potassium tetrakis(4-chlorophenyl) borate, and chromoionophore for detection at a 640-nm wavelength. The slope (sensitivity) of a dose-response curve (absorption of the 640-nm wavelength vs. concentration) was determined by the measurements of calibrant 1 and calibrant 2.[94] The ion concentration of unknown samples was then deduced from this calibration curve.

Example 6.9: The CD Platform as a Microscope

Horacio Kido and Jim Zoval from the author's research team recognized that the optical

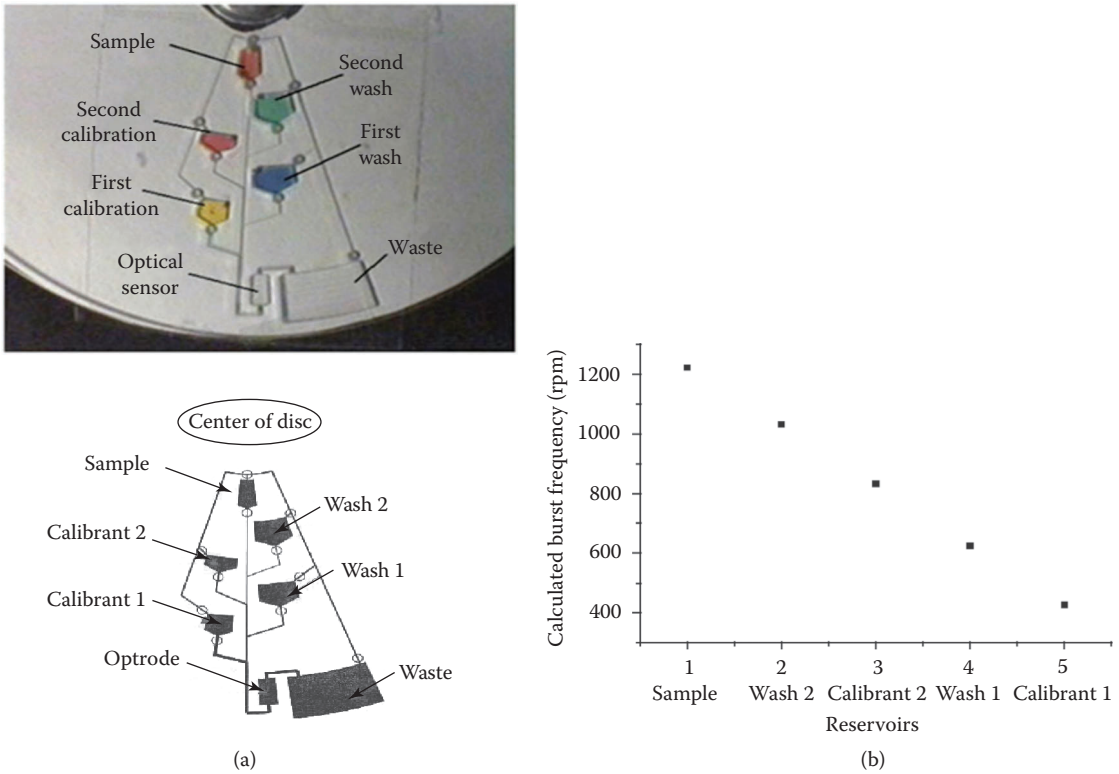


FIGURE 6.91 Two-point calibration of an optrode on a CD. (a) Functional diagram and photo of a two-point calibration unit on a CD. (b) A plot of calculated burst frequencies.

disc drive is a sophisticated laser scanning microscope designed to characterize and identify micrometer-sized features at a rate of about a megahertz (Figure 6.92). By taking the voltages from the photodetector and sending them to a computer using a fast A/D converter, an image of what is on the disc can then be reconstituted using simple graphics software. It does not matter whether the image is a pit in the CD (see Figure 6.92) or objects placed on the disc (see Figure 6.93).

Example 6.10: Using the Coriolis Force on the CD

On a spinning CD, the Coriolis force can be used to switch fluid flow between two reservoirs depending on the direction of the CD's rotation. The Coriolis force, we remember, is a fictitious force exerted on a body when it moves in a rotating reference frame. It is called a fictitious force because it is a by-product of a coordinate transformation. To understand the Coriolis force

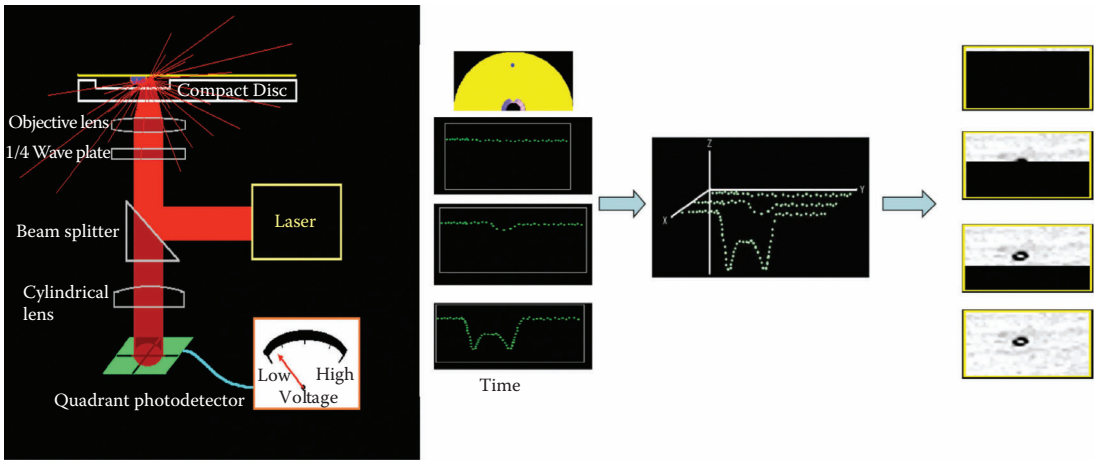


FIGURE 6.92 Using the CD as a microscope.

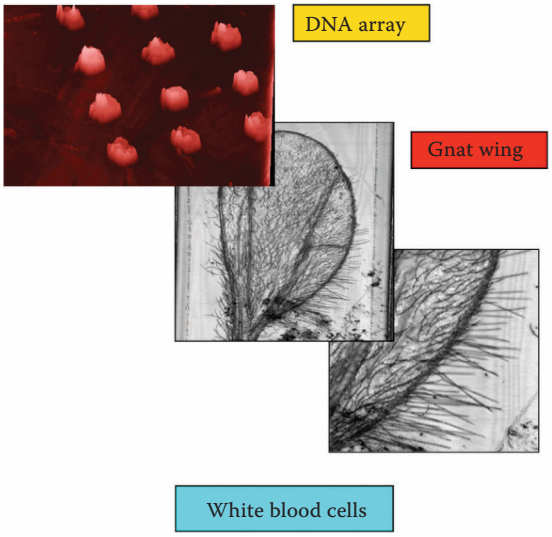


FIGURE 6.93 Images from the CD microscope.

on the CD, it is convenient to rewrite the Navier-Stokes equation (Equation 6.8) in a rotating reference frame. This results in an expression for the Coriolis force given as:

$$F_{Coriolis} = -2\rho \cdot \mathbf{u} \times \mathbf{u} \qquad (6.142)$$

The formula implies that the Coriolis force is perpendicular both to the direction of the velocity of the moving mass and to the frame's rotation axis. In this expression, ρ is the fluid density, ω is the angular velocity vector, and $\mathbf{u}$ is the flow velocity vector. The Coriolis force is transverse to the CD radius and acts to transport the fluid stream left or right depending on the direction of the CD rotation (see Figure 6.94a). The Coriolis force thus can be used to switch fluid flow between a left reservoir and a right reservoir depending on the direction of the CD rotation (Figure 6.94b).

Some other applications of the CD, including cultivation of *C. elegans* worms, can be found in the References.[95,96] A recent review article on CD fluidics can be found in Madou et al.[97]

In Table 6.8 we compare four popular fluidic platforms: mechanical, osmotic, centrifugal, and acoustic. This table should be read together with the more detailed Table 1.8 from Chapter 1 in Volume III. For more background on the physics behind acoustic fluidics see Volume III, Chapter 8 on actuators.

Scaling in Analytical Separation Equipment

Introduction

When incorporating sensors in microfluidics one finds that most sensing techniques scale poorly in the microdomain. As we analyze in Chapter 7, different detection schemes exhibit different signal scaling factors. For example, in an amperometric technique, in which an analytical current is the signal, the output exhibits a (l^2) scaling (i.e., proportional to the area of the electrode); an optical absorption technique scales as (l^3), and potentiometric techniques are, in principle, size-independent. In general we find that most detection methods are less sensitive

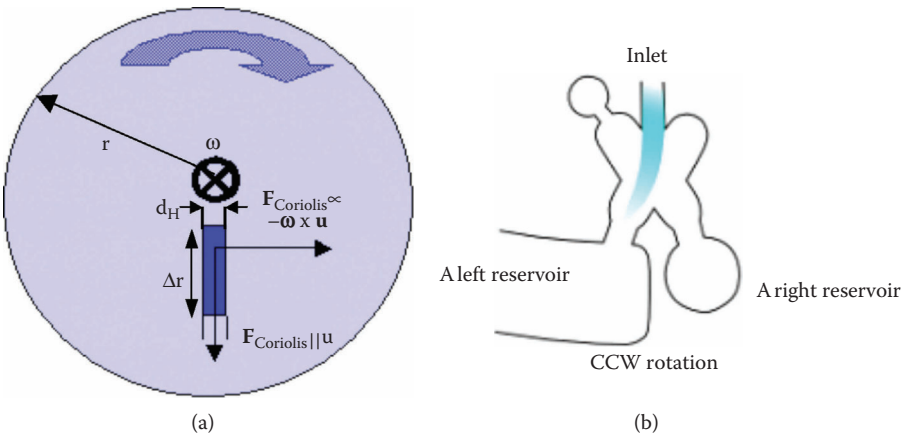


FIGURE 6.94 Coriolis force for flow switching. (a) Forces shown on a disc rotating at the angular velocity vector ω pointing into the paper plane. When this disc spins in the counterclockwise direction, the Coriolis force switches the fluid in the channel to the left, and when this disc spins in the clockwise direction, the fluid moves to the right. (b) When the CD rotates counterclockwise (CCW), the Coriolis switching valve diverts fluid to the waste reservoir (left reservoir), and when the CD rotates clockwise, fluid is diverted to the extraction chamber (right reservoir).

Table 6.8 Comparison of CD Fluidics with Mechanical, Osmotic, and Acoustic Pumping (see also Table 1.8 , Chapter 1 in Volume III)

Mechanical (e.g., with Johnson and Johnson's blister pouch HIV test; see Figure 1.12 in Volume III, Chapter 1)
– Scales as l^3
+ No contact of sample with electrodes
+ Generic (all types of sample can be used)
+ Innovation in the blister pouch
+ Solves liquid and vapor valving
– Difficult to further miniaturize
– Difficult to multiplex
Electrokinetic (electro-osmosis) (e.g., with the microchip in Figure 6.71 used for high-speed electrophoretic separations)
+ Scales as l^2
– High-voltage source is not convenient
– Many parameters influence propulsion force
– Not generic (only limited types of samples can be used)
– Mixing difficult to implement
– Sample contact with capillary walls modifies the electro-osmotic force (EOF)
+ First products on the market
/ Solves liquid valving but not for vapors!
+ Better for high-throughput screening and smaller samples
Centrifugal (e.g., with the two-point calibration CD in Figure 6.91)
+ Scales a bit better than l^3
+ Compatible with a wide range of samples
+ Mixing easy to implement
+ Sample preparation easier than with an electrokinetic set-up
+ Simple and inexpensive CD player for drive
+ No fluid contact
+ Established
+ Generic
/ Solves liquid valving elegantly but not vapor valving
+ Most functions demonstrated
+ Cell work easier
+ Better for diagnostics
Acoustic (e.g., with White's flexural plate wave device; see Volume III, Chapter 8)
+ Scales as l^2
+ No fluidic contact
– Research and development phase
+ Generic
– Does not solve valving yet
– ZnO technology to implement the acoustic transducer still difficult to reproduce
– Easy to further miniaturize

when used in microchambers in microfluidic platforms. In this chapter, we saw already that all flow in small tubes is laminar (making mixing a challenge), that evaporation from small droplets is very fast, and that surface tension in the microdomain becomes very appreciable. These represent mostly negative aspects associated with scaling down any analytical equipment involving fluid handling. Positive attributes of miniaturizing such equipment are the savings on expensive reagents, short analysis times, and

fast heating/cooling (e.g., for PCR), which all dictate small devices.

We now look into one class of analytical equipment involving fluid handling, namely, separation equipment, with some more detail. Electrophoresis, as described earlier in this chapter, is only one of many separation techniques used in analytical chemistry; other methods include gas chromatography (GC) and many varieties of liquid chromatography. Miniaturization in electrophoresis and chromatography has been a study topic for many years. The goal is not only to separate smaller amounts of chemical compounds, but theory also predicts that a reduction in dimensions of the separation column should result in an enhancement of analytical performance, such as a shorter separation time and more efficient heat dissipation.[98,99] To put recent advances in miniaturization of separation devices in context, we need to reconsider the fundamentals of separation methodology. The following theoretical analysis will shed light on scaling laws applicable to analytical separation systems. We closely follow the derivations presented by Manz et al.[98] One principal goal of separation techniques is to create the narrowest possible separation bands, w, to increase the separation efficiency. Several mechanisms contribute to band broadening, each with its own standard deviation σ, and we need to understand each of the contributing mechanisms to appreciate the effect of miniaturization on separation performance.

Commonly Used Terms in Separation Chemistry

Analytical separation of compounds often is affected by forcing the sample compounds, suspended in a carrier medium or mobile phase, through a selectively absorbing medium or stationary phase immobilized in some sort of flow channel. The mobile phase in chromatography may be a liquid or a gas. The stationary phase (the one that stays in place inside the flow channel or column) is most commonly a viscous liquid coated on the inner side of a capillary tube or on the surface of solid particles packed in the column. Alternatively, the solid particles in the column may form the stationary phase themselves. To force the sample and carrier phase (i.e., the eluent), through the column, either a pressure or electrical gradient, respectively, defining chromatography techniques and electrophoresis techniques, can be employed. The fluid emerging from the column is called the *eluate*, and the process of passing liquid or gas through a chromatography column is called *elution*. The partitioning of solutes between the mobile and stationary phases gives rise to separation of the various components in the sample. Columns are either packed or open tubular. A packed column is filled with particles as the stationary phase or the particles are coated with the stationary phase. An open columnar column is a narrow capillary in which the inside wall itself forms the stationary phase or a stationary phase is coated on that wall. The center of an open tubular column is hollow. Based on the mechanisms of interactions of the solute with the stationary phase, different types of chromatography are distinguished: adsorption, partition, ion exchange, molecular exclusion (also called *gel filtration*), and affinity chromatography.[100]

An electrochemical or optical detector, in the case of liquid chromatography, or a thermal conductivity detector (TCD) or flame ionization detector (FID), in the case of gas chromatography, may detect solutes eluted from a column. A chromatograph is a graph showing the response of such a detector as a function of elution time (Figure 6.95). Ideally, a solute applied as an infinitely narrow band at the inlet of a separation column emerges with a Gaussian shape at the outlet. In less ideal circumstances, the band becomes asymmetric. Different compounds are retained for different amounts of time, called *retention times* (t_r), onto the immobilized medium through which they are forced. Retention time t_r is the time needed after injection of the mixture for each component to reach the detector. The retention time for a solute often is corrected for the time it takes the unretained mobile phase (e.g., air) to travel through the column (t_m). This so-called adjusted retention time t'_r may be written as:

$$t'_r = t_r - t_m \tag{6.142}$$

For any two components 1 and 2, the relative retention, α, is given as:

$$\alpha = \frac{t'_{r2}}{t'_{r1}} \tag{6.143}$$

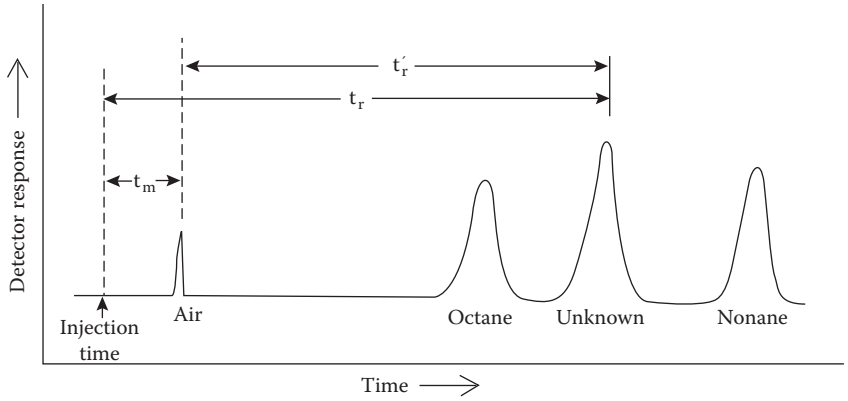


FIGURE 6.95 Schematic gas chromatogram showing measurement of retention times.

in which $t'_{r2} > t'_{r1}$ so that α is always larger than 1. The larger the α, the better the separation between compounds 1 and 2. The relative retention is independent of flow rate and is used to identify compounds, even when the flow rate has changed between experiments. For each compound in the column, the capacity factor, k', is defined as:

$$k' = \frac{t_r - t_m}{t_m} \tag{6.144}$$

A larger capacity factor signifies a compound better retained onto the column. The capacity factor in Equation 6.144 corresponds to the ratio of time the solute spends in the stationary phase over the time it spends in the mobile phase. The latter ratio in turn is based on the partition coefficient K, an equilibrium constant defined by the ratio of the concentration C_s of solute in the stationary phase over C_m, the concentration of the solute in the mobile phase:

$$k' = \frac{C_s V_s}{C_m V_m} = K\frac{V_s}{V_m} \tag{6.145}$$

where V_s is the volume of the stationary phase and V_m the volume of the mobile phase. From Equations 6.143–6.145, we can conclude that:

$$\alpha = \frac{t'_{r2}}{t'_{r1}} = \frac{k'_2}{k'_1} = \frac{K_2}{K_1} \tag{6.146}$$

The relative retention of two solutes is proportional to the ratio of their partition coefficients.

Two factors contribute to the efficiency of separation of components on a column. One is the difference in elution times between peaks; the other is the width of the peaks. Different compounds in an applied sample on a column separate into bands, and the further apart they are, the better the separation performance of the setup. The resolution of two peaks is defined as:

$$\text{resolution} = \frac{\Delta t_r}{w_{av}} = \frac{\Delta V_r}{w_{av}} \tag{6.147}$$

in which Δt_r or Δw_r is the separation between peaks in units of time (t) or volume (V), and w_{av} is the average width of the two peaks in corresponding units. A band of solute in a column invariably spreads as it travels through a separation medium and emerges at the detector with a standard deviation σ; the narrower the bands (w), the better the resolution.

Band Broadening in the Column

Individual solute bands in a column generally assume symmetric concentration profiles that can be described in terms of a Gaussian distribution curve and its standard deviation, σ_x. The Gaussian curve is given by:

$$C = \frac{m}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}} \tag{6.148}$$

where D = diffusion coefficient (m^2/s) (10^4 faster in gases than in liquids)
m = moles per unit cross-sectional area
C = concentration (mol m^{-3})
t = time
x = distance along the column (total length L_t) from the center of the band (at the band center, x = 0)

The standard deviation of this distribution curve, σ_x, is used to describe the bandwidth length and is given by:

$$\sigma_x = \sqrt{2Dt} \tag{6.149}$$

i.e., the width at time t of a band progressing in the column.

If a solute moves a distance x at the linear flow rate u_x (m s^{-1}), the time it has been on the column is $t = x\,u_x^{-1}$ (when $x = L_t$, then $t = t_r$). Or we can write that:

$$\sigma_x^2 = 2Dt = 2D\frac{x}{u_x} = \left(\frac{2D}{u_x}\right)x = Hx \tag{6.150}$$

and:

$$H = \frac{\sigma_x^2}{x} \tag{6.151}$$

In this equation, plate height H (also called the *height equivalent to a theoretical late*) is the proportionality constant between the variance (σ^2) of the band and the distance it has traveled (x). Thus, plate height is simply a quantity relating the width of a band to the distance traveled through the column. The number of theoretical plates, N, in the entire column is the total length L_t divided by the plate height:

$$N = \frac{L_t}{H} = \frac{Lx}{\sigma_x^2} = \frac{L_t^2}{\sigma_x^2} = \frac{16L_t^2}{w^2} \tag{6.152}$$

because $x = L$ and $\sigma = w/4$. The above analysis can be made more generic; the bandwidth parameter, σ, for a flow channel can be expressed not only in terms of length in m (σ_x, see above) but also in terms of time in s (σ_t), and volume in m^3 (σ_v). Rewriting Equation 6.152 with L and w (or σ) in units of time instead of length leads to:

$$N = \frac{16t_r^2}{w^2} = \left(\frac{t_r^2}{\sigma^2}\right) \tag{6.153}$$

where t_r is the retention time of the peak, and w is the width at the base of the peak in units of time.

The smaller the H and thus the larger the N, the narrower the bands and the better the separation. Plate heights are ~0.1 to 1 mm in gas chromatography (GC), ~10 µm in high-performance liquid chromatography (HPLC), and <1 µm in capillary electrophoresis (CE).[101] In HPLC the grains size of the matrix is reduced: this leads to an increased number of plates N, but at the cost of a much increased pressure drop, Δp, across the length of the column.

The resolution between two compounds in a chromatography technique (Equation 6.147) now can be expressed in terms of number of theoretical plates in the column, N:

$$\text{resolution} = \frac{\sqrt{N}}{4}\left(\frac{\alpha-1}{\alpha}\right)\left(\frac{k_2'}{1+k_{av}'}\right) \tag{6.154}$$

In this equation, α is the relative retention, k'_2 is the capacity factor for the strongest retained component of two compounds, and k'_{av} is the average capacity factor for the two components. Resolution, like band spreading, scales as $N^{1/2}$. Because the number of theoretical plates is proportional to column length, doubling the column length increases the resolution by $\sqrt{2}$. Resolution also increases with increasing α and k'_2. Increasing the capacity factor only works up to a certain level, as retention times might become too long and the peaks too broad.

Golay[102] calculated the height equivalent of a theoretical plate, H, as a function of flow rate (u) as:

$$H = \underbrace{\frac{2D_m}{u_x}}_{A} + \underbrace{\frac{1+6k'+11k'^2}{96\left(1+k'\right)^2}\left(\frac{d^2}{D_m}\right)u_x}_{B} + \underbrace{\frac{2k'}{3\left(1+k'\right)^2}\left(\frac{d_f^2}{D_s}\right)u_x}_{C} \tag{6.155}$$

In this equation, D_m and D_s (m^2s^{-1}) correspond, respectively, to the diffusion coefficients of the sample molecules in the mobile and stationary phase; u_x (ms^{-1}) is the linear flow rate of the mobile phase, k' is the capacity factor of the specific compound, d is the diameter of the capillary; and d_f is the thickness of the stationary phase layer. The A term in Equation 6.155 ($2D_m/u_x$) is the result of longitudinal diffusion (spreading out along the length of the column) and contributes an amount H_D to the expression for H. This type of diffusion is called Taylor-Aris dispersion. This diffusional term can be derived from Equations 6.150 and 6.151 as:

$$H_D = \frac{\sigma^2}{L_t} = \frac{2D_m}{u_x} \tag{6.156}$$

An expression given by Taylor[103] approximates D_m in Equation 6.156 as:

$$D_m \sim Pe^2 D \tag{6.157}$$

where the Pe number is given by:

$$Pe = \frac{ud}{D_m} \tag{6.158}$$

Formula 6.157 works at high Péclet numbers; for microsystems Pe is typically between 0.1 and 100. For Péclet numbers in the upper range, diffusion along the flow can be increased by several orders of magnitude compared with molecular diffusion, a feature very useful in micromixers.

The faster the linear flow, the less time is spent in the column, and the less diffusional broadening occurs. Whereas the *A* term in the Golay equation is associated with longitudinal diffusion, *B* is the result of mass transfer in the mobile phase and contributes H_m to the expression of *H*. The *C* term is associated with mass transfer to the stationary phase, and it contributes a term H_s to *H* as a result of the finite time it takes for a solute to equilibrate between the mobile phase and the stationary one. Decreasing the thickness of the stationary phase d_f and increasing the temperature both reduce H_s. The thickness uniformity of coatings in capillaries is one of the most difficult parameters to control in micromachined columns. As a result, the current commercial micro gas chromatographs have miniaturized detectors and injectors but no micromachined columns (see Figure 1.22). Glass capillary columns with inner diameters as small as 100 µm continue to outperform micromachined columns because of the difficulty in maintaining coating uniformity in micromachined structures. Equation 6.155 was derived for round fluid conduits; Wong et al.[104] have derived a modified Golay equation for rectangular channels with different aspect ratios. These authors calculated that the difference between slip and no-slip flow in those channels becomes appreciable (as high as 12%) if the column size is reduced to 10 µm.[104]

For capillary electrophoresis in an open tube, the expression for the number of theoretical plates simplifies as both the *B* term and the mass transfer *C* term disappear, reducing plate height to submicron values. Moreover, the high surface-to-volume ratio improves heat dissipation, enabling the application of higher voltages and thus faster separations. The method provides unprecedented resolution as compared to a packed wider column. In the case of capillary electrophoresis, miniaturized planar columns can easily outcompete traditional glass capillaries.

Other Band-Broadening Effects

In practice, bands broaden not only because of diffusional terms (σ_{diff}), which cause broadening in the column itself, but also because of injector, connector, and detector finite volumes. Total band broadening is calculated from a summing of the squares of the standard deviations, that is, the variances (variance is additive but standard deviation is not):

$$\sigma^2 = \sigma^2_{diff} + \sigma^2_{inj} + \sigma^2_{det} \tag{6.159}$$

These are all important parameters to calculate when designing a separation system. Injection and detection variances in traditional separation systems usually can be made smaller than a quarter of the diffusional variance. To contribute less than 10% of the total band broadening, injection and detection volume must be less than $\sigma_v/2$ of the column, and the detector response time must be faster than $\sigma_t/2$. If in injection the band is applied as a sharp-edged zone of width Δt (in units of time), the contribution to the variance of the final bandwidth is:

$$\sigma^2_{inj} = \frac{(\Delta t)^2}{12} \tag{6.160}$$

A similar equation holds for broadening in the detector that requires a time *t* to pass through. Systems must be designed such that the dead volume can be kept minimal and the capillary heated and cooled very quickly and uniformly with a minimal power budget; that is, the heat capacity of the system must be minimized.

To retain a small footprint while making a separation channel longer, the channel can be made to turn back and forth or made into a spiral. Unfortunately, distance and field strength vary laterally across the width of the channel in those turns; that is, a "racetrack effect" is generated. Racetrack effects cancel

the benefit of additional length. Molho et al.,[105] using Memcad 4.6, designed and fabricated "compensated" turns in PMMA, greatly reducing dispersion compared with an uncompensated corner. Using turns with large radii of curvature and/or very narrow channel widths, Culbertson et al.[106] fabricated glass microchips with less dispersion in the turns. These authors also built a separation channel in glass in the shape of a spiral 25 cm long and fitting on a 5 × 5-cm piece of real estate. Negligible band broadening from the turns and separation efficiencies of more than 1,000,000 theoretical plates in less than 46 seconds with an applied potential of 1100 V/cm for dichlorofluorescein dye were observed.[106]

Scaling

Miniaturization of the detector in a miniaturized analysis system, we alluded to earlier, is quite a challenge, especially with detector chamber sizes in the order of picoliters. By fixing the number of theoretical plates, N; the retention time, t; and the heating power per length (for capillary electrophoresis), the resulting flow channel dimensions and operating conditions for different separation techniques are compared. In Table 6.9 we reproduce calculations by Manz et al.[98,99] comparing capillary electrophoresis (CE), liquid chromatography (LC), and supercritical fluid chromatography (SFC). Listed are flow channel lengths, diameter, operating voltage or pressure, minimum sample bandwidths (expressed in σ_x, σ_t, and σ_v), peak capacity, length/diameter ratio of eluting peaks (σ_x/d), and the required detector and injector specifications.

Still following Manz et al.'s analysis,[99] we now introduce a set of dimensionless or reduced parameters to make comparisons of separation systems easier. By introducing reduced parameters, including the Reynolds number, Péclet number (flow rate), Fourier number (elution time), and Bodenstein

Table 6.9 Calculated Parameter Sets for a Given Separation Performance

Parameter		Electro-Osmotic Chromatography			Liquid Chromatography (LC)		Supercritical Fluid Chromatography (SFC)	
Number of theoretical plates	N	100k	1M	10M	100k	1M	100k	1M
Analysis time	1 (k′ = 5) (min)	1	1	1	1	1	1	1
Heating power	P/L (W/m)	1.1	1.1	1.1				
Capillary inner diameter	d (μm)	24	7.6	2.4	2.8	0.9	6.9	2.2
Capillary length	L (cm)	6.5	21	65	8.1	26	20	64
Pressure drop	p (atm)				26	2600	1.4	140
Voltage	U (kV)	5.8	58	580				
Peak capacity	n	180	570	>2000	220	700	220	700
Signal bandwidth	σ_x (mm)	0.21	0.21	0.21	0.56	0.56	1.4	1.4
	σ_t (ms)	42	13	4.2	70	22	70	22
	σ_v (pl)	94	9.4	0.94	3.3	0.33	52	5.2
Detection volume	V (pl)	47	4.7	0.47	0.8	0.08	1.2	12
Response time	t (ms)	21	6.5	2.1	16	5	16	5
Injection pulse	p × t (*s* × atm)				1.5	49	0.075	2.4
	U × t (*s* × kV)	0.41	1.3	4.1				
Stop time	t (*s*)	3.3	3.3	3.3	5.1	5.1	5.1	5.1

Note: Obtained with capillary electro-osmotic (EC), liquid (LC), and supercritical fluid chromatography (SFC). Assumed constants are diffusion coefficients 1.6×10^{-9} m²/s (LC, EC) and 10^{-8} m²/s (SFC); viscosities of the mobile phase 10^{-3} Ns/m² (LC, EC) and 5×10^{-5} Ns/m² (SFC); electrical conductivity of the mobile phase 0.3 S/m (EC); electrical permittivity × zeta potential, 5.6×10^{-11} N/V (EC); and heating power 1.1 W/m (EC).

Source: After Manz, A., E. Verpoorte, C. S. Effenhauser, N. Burggraf, D. E. Raymond, D. J. Harrison, and H. M. Widmer. 1993. Miniaturization of separation techniques using planar chip technology. HRC *J High Resol Chromatogr* 16:433–36; and Manz, A., N. Graber, and H. M. Widmer. 1990. Miniaturized total chemical analysis systems: a novel concept for chemical sensing. *Sensors Actuators B* B1:244–48.

number (pressure drop), systems are easier to compare because all reduced parameters remain constant regardless of the size of the system.[107] First, we define some of these reduced parameters in terms of those variables that can be assumed constant over the entire range of interest, including inner capillary diameter (d), mobile phase viscosity (η_m), average diffusion coefficient (D_m) of the sample in the mobile phase, and a Poiseuille number, C, of 32 for a circular cross-section. The diameter, d, is also the characteristic length here. With those constants, other quantities can be grouped into dimensionless forms—for example, volume, V; column length, L; linear flow rate, u; retention time, t; and pressure drop, Δp. A reduced volume parameter, for example, is obtained by division through d^3. Thus, one obtains:

$$w = \frac{V}{d^3} \text{ and } s_v = \frac{\sigma_v}{d^3} \tag{6.161}$$

Similarly, time-related parameters such as the migration time, t_0; the retention, t, for a compound with a capacity factor k'; and the time bandwidth, σ_t, can be reduced to their dimensionless terms, known as Fourier numbers, in the following fashion:

$$\tau_0 = \frac{t_0 D_m}{d^2}, \tau = \frac{tD_m}{d^2}\left(k'+1\right), \text{ and } s_t = \frac{\sigma_t D_m}{d^2} \tag{6.162}$$

The first expression describes the ratio of the time required for a molecule to migrate in a field-driven flow through the capillary from end to end to the time required to diffuse from wall to wall. The second term defines the analogous expression for chromatography. Terms with a length dimension are reduced by dividing by d:

$$\lambda = \frac{1}{d}, \; h = \frac{H}{d}, \text{ and } s_x = \frac{\sigma_x}{d} \tag{6.163}$$

where λ is the so-called reduced length. The linear flow rate of a nonretained component or the mobile phase is reduced to the Péclet number, Pe, given earlier as:

$$Pe = \frac{ud}{D_m} \tag{6.158}$$

which represents the average linear flow rate divided by the absolute value of the average rate of diffusion orthogonal to the direction of flow, D_m/d. The applied pressure can be reduced to the so-called Bodenstein number as:

$$\Pi = \frac{\Delta p d^2}{\eta_m D_m \Phi} = \frac{uL}{D_m} \tag{6.164}$$

The Bodenstein number thus represents the average linear flow rate divided by the absolute value of the average rate of longitudinal diffusion, D_m/L. In this expression, Φ is the Poiseuille number (32 for a circular cross-section). The applied voltage can be reduced to the "electrical" Bodenstein number, Ψ:

$$\Psi = \frac{U\varepsilon\xi}{\eta_m D_m} \tag{6.165}$$

with ζ the zeta potential. Both Bodenstein numbers, describing the reduced pressure and voltage, can be rewritten as the so-called flow equations:

$$\Psi = \lambda Pe \tag{6.166}$$

in electro-osmotic and electrophoretic flow, and:

$$\Pi = \lambda Pe \tag{6.167}$$

for pressure-driven chromatography. Also, the Fourier numbers can be further simplified as:

$$\tau_0 = \frac{\lambda}{Pe} \text{ and } \tau = \frac{\lambda}{Pe}\left(k'+1\right) \tag{6.168}$$

The equations for reduced plate height, h, and the Péclet number, Pe, lead to the following simplified Golay equation for band broadening in pressure-driven flow, neglecting the third term (C) in Equation 6.155:

$$h = \frac{2}{Pe} + \frac{1+6k'-11k'^2}{96\left(1+k'\right)^2} Pe \tag{6.169}$$

for a field-driven flow we simply obtain:

$$h = \frac{2}{Pe} \tag{6.170}$$

The reduced bandwidth is then given by:

$$s_x^2 = \lambda h \tag{6.171}$$

and the plate number by:

$$N = \frac{\lambda}{h} \text{ and } N = \lambda \frac{Pe}{2} \tag{6.172}$$

Table 6.10 Reduced Parameter Set for the Example Separation Systems in Table 6.9

Parameter	CE	Capillary LC	Capillary SFC
Number of theoretical plates, N	100,000	100,000	100,000
Analysis time, t (k′ = 5) (min)	1	1	1
Capillary inner diameter, d_c (μm)	24	2.8	6.9
Capillary length, L (cm)	6.5	8.1	20
Reduced length, λ	2700	29,000	29,000
Péclet number (reduced flow rate), Pe	100	14	14
Fourier number (reduced retention time), τ	28	2100	2100
Bodenstein number (reduced pressure decrease), ∏	—	400,000	400,000
Electric Bodenstein number (reduced voltage decrease), ψ	260,000	—	—

CE = capillary electrophoresis; LC = liquid chromatography; SFC = supercritical flow chromatography.
Source: Manz, A., E. Verpoorte, C. S. Effenhauser, N. Burggraf, D. E. Raymond, D. J. Harrison, and H. M. Widmer. 1993. Miniaturization of separation techniques using planar chip technology. HRC *J High Resol Chromatogr* 16:433–36; and Manz, A., N. Graber, and H. M. Widmer. 1990. Miniaturized total chemical analysis systems: a novel concept for chemical sensing. *Sensors Actuators B* B1:244–48.

It can be recognized here how the number of theoretical plates is in a way representative of the Péclet number of the system. From Equations 6.172 and 6.163, the smaller D_m, the higher the number of plates, and the larger the velocity and the length of the column, the more the separation is "efficient."

With these reduced parameters, the examples of Table 6.9 now are compared again in Table 6.10. We conclude that the values for the reduced variables for LC and SFC are identical, regardless of differences in capillary diameter, lengths, diffusion coefficients, and viscosities. This can be used conveniently to deduce the influence of changing the capillary diameter on retention time, pressure, and signal bandwidth for a given number of theoretical plates and a single set of reduced parameters, as illustrated in Table 6.11.

The scaling for diffusion-controlled separations, in which the time scale is proportional to d_2, is summarized in Table 6.12. All reduced parameters remain constant regardless of the size of the system; that is, hydrodynamic diffusion, heat diffusion, and molecular diffusion effects behave in the miniaturized system as in the original large system. Reducing the characteristic length d (i.e., the tube diameter) by a factor of 10 makes for a 100-times faster analysis. The pressure required is 100 times higher, and, more importantly, the voltage requirements remain unchanged in electrophoresis/electro-osmosis systems. Miniaturization in diffusion-controlled

Table 6.11 Calculated Parameter Set for an Open Tubular Column LC System

	Diameter d (μm)					Reduced Parameter
	1	2	5	10	20	
Length, L (m)	0.45	0.9	2.3	4.5	9	λ = 450,000
Time, t (min)	0.12	0.5	3	12	50	t = 11,800
Pressure, Δp (atm)	8700	2200	350	87	22	∏ = 17,000,000
Peak, σ_x (μm)	450	890	2200	4500	8900	s_x = 447
Peak, σ_t (ms)	7.4	30	190	740	3000	s_t = 12
Peak, σ_v (pl)	0.35	2.8	44	350	2800	s_v = 354

Note: One million theoretical plates at zero retention (Péclet number Pe = 38). Assume diffusion coefficient is $D_m = 10^{-3} m^2 s$.
Source: Manz, A., E. Verpoorte, C. S. Effenhauser, N. Burggraf, D. E. Raymond, D. J. Harrison, and H. M. Widmer. 1993. Miniaturization of separation techniques using planar chip technology. HRC *J High Resol Chromatogr* 16:433–36; and Manz, A., N. Graber, and H. M. Widmer. 1990. Miniaturized total chemical analysis systems: a novel concept for chemical sensing. Sensors Actuators B B1:244–48.

Table 6.12 Proportionality Factors for Some Mechanical Parameters in Relation to the Characteristic Length, d, in a Diffusion-Controlled System

	Diffusion-Controlled System
Space, d	d
Time, t	d^2
Linear flow rate, u	L/d
Volume flow rate, F	d
Pressure drop (laminar flow) Δp	L/d^2
Voltage (electro-osmotic flow), U	Constant
Electric field, U/L	L/d
Reynolds number, Re	Constant
Péclet number, reduced flow rate, Pe	Constant
Fourier number, reduced elution time, τ	Constant
Bodenstein number, reduced pressure, Π	Constant
Reduced voltage, ψ	Constant

Source: After Manz, A., E. Verpoorte, C. S. Effenhauser, N. Burggraf, D. E. Raymond, D. J. Harrison, and H. M. Widmer. 1993. Miniaturization of separation techniques using planar chip technology. HRC J High Resol Chromatogr 16:433–36; and Manz, A., N. Graber, and H. M. Widmer. 1990. Miniaturized total chemical analysis systems: a novel concept for chemical sensing. Sensors Actuators B B1:244–48.

systems leads to a higher rate of separation while maintaining separation efficiency.

Separation efficiency in capillary electrophoresis, in terms of theoretical plates per second, can be estimated from Equations 6.170 and 6.172:

$$N = \frac{\Psi}{2} \propto U \tag{6.173}$$

where Ψ is the reduced voltage so that N is proportional to U, the applied voltage; the higher the applied voltage, the higher the number of theoretical plates. The voltage cannot be increased too much though, because heat evolution quickly becomes the limiting factor. In standard capillary electrophoresis, the maximum allowable heat generation is about 1 W/m. Because of the higher surface-to-volume ratio in micromachined channels, heat dissipation is faster, which permits higher electric fields than in standard capillary electrophoresis. For now, we will admit 1 W/m as an upper limit even for micromachined capillaries. Keeping the power per unit length constant means:

$$\frac{P}{L} = \frac{UI}{L} = \text{const} \tag{6.174}$$

where I is the current through the capillary. The resulting upper limit for the voltage is then determined by the geometry of the capillary in terms of the reduced characteristic length ($\lambda = 1/d$):

$$U_{max} \propto \lambda \tag{6.175}$$

The plate number N can thus reach values up to N_{max}, where:

$$N_{max} \propto \lambda \tag{6.176}$$

The minimum migration time is given by:

$$t_0 \propto \frac{L^2}{U} \tag{6.177}$$

and:

$$t_{min} \propto L^2 d \tag{6.178}$$

The maximum number of plates obtainable per second is then given by:

$$\frac{N_{max}}{t_{min}} \propto \frac{1}{L^2 d^2} \tag{6.179}$$

clearly demonstrating the benefits of reducing the inner diameter of the capillary for rapid separation in capillary electrophoresis.

The reduced height, h, of a chromatographic system according to Equation 6.169 shows a minimum, h_{min}, at an optimum reduced flow rate, Pe_{opt} (Péclet number). To reduce losses in pressure drop and to minimize analysis time, one must operate close to this optimum. Analogous to Equation 6.172, we can write the maximum number of theoretical plates as:

$$N_{max} = \frac{\lambda}{h_{min}} \tag{6.180}$$

where h_{min} is constant for optimum Péclet number, Pe_{opt}, at a fixed capacity factor k'. This makes Equation 6.180 for pressure-driven chromatography equivalent to Equation 6.176 for electrophoresis. Equations 6.178 and 6.179 are also equally valid for capillary electrophoresis and capillary chromatography.

The reduced parameter analysis shows that a downscale of 1/10 of the original size (diameter of a tube) reduces the related time variables (analysis time, required response time of a detector) to 1/100. The pressure requirements increase by a factor of 100, but the voltage requirements (for electrophoresis/electro-osmosis) remain constant. The main advantage of capillary electrophoresis is thus higher speed of separation with a comparable efficiency.[108]

Example 6.11: Droplet Microfluidics

In a new branch of microfluidics, a rich variety of droplets are made by emulsification in microfluidic systems. Droplet-based microfluidics allows for independent control of each droplet, thus generating microreactors that can be individually transported, mixed, and analyzed.[109] The generation of droplets involves the disruption of interfacial forces between two streams of immiscible fluids either via flow-induced shear forces or other surface forces (e.g., electrowetting, dielectrophoresis). An often-used breakup method of streams of immiscible fluids is in the confined geometry of a microfluidic T-junction (see Figure 6.56). For a detailed theoretical treatment of droplet formation in a T-junction, we refer to the work by De Menech et al.[110] The droplets thus formed can be filled with reagents to perform bioassays in discretized femtoliter to nanoliter volumes separated by immiscible fluids as opposed to more conventional microfluidics, where reactions are not confined either spatially or temporally within the microchannels. Rapid mixing is possible because there is relative fluid motion between the immiscible fluids at the droplet interfaces that create recirculation of flows inside the droplets. The sizes of the droplets are extremely monodisperse (within 2%) and can be controlled over a wide range (typically 1–100 μm in diameter).[111–115] Because of high surface area to volume ratios at the microscale, heat and mass transfer times and diffusion distances are shorter, facilitating faster reaction times. Because multiple identical microreactor units can be formed in a short time, parallel processing and experimentation can easily be achieved, allowing large data sets to be taken efficiently and enabling high throughput screening and syntheses. Droplet microfluidics also offers greater potential for increased throughput and scalability than continuous single-phase flow microfluidics. In the past 5 years, several groups have used digital microfluidics to form irregular particles,[116] double emulsions,[117] hollow microcapsules,[118] and microbubbles.[74,119–123]

As a curious return to fluidic logic (see the fluidic logic chips from 1965 in Volume III, Figure 1.2), MIT's Neil Gershenfeld et al.[124] are building rudimentary bubble-based logic devices. They constructed an entire family of chips modeled on the architecture of existing digital circuits that mimic the functions of electronic components: logic gates (AND/OR/NOT), memory, and even a microfluidic ring oscillator constructed from three AND gates (see Figure 6.96). These digital microfluidic devices show the nonlinearity, gain, bistability, synchronization, cascadability, feedback, and programmability required for scalable universal computation. Instead of using high and low voltages to represent a bit of information, the presence or absence of a bubble is used. This means that a bubble can not only carry some kind of payload—say a therapeutic drug or chemical reagents (see above)—but

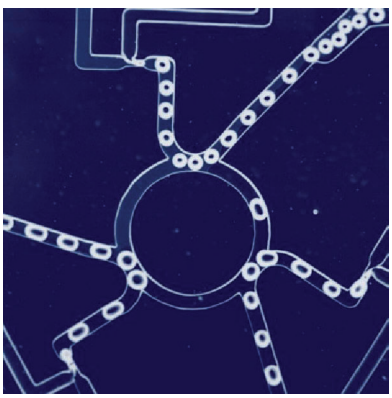


FIGURE 6.96 A bubble microfluidic ring oscillator constructed from three AND gates.

it also can carry information at the same time. The bubble chips are 100 times slower than an average microprocessor, but among other uses they could enhance the memory capabilities of microfluidic chips: thousands of reagents could be stored on a single chip—much like conventional data storage technology—using counters to dispense exact amounts and built-in logic circuits to deliver the reagents to targeted spots. They could also be used to sort biological cells in assays and to make programmable print heads for inkjet printers, for example. With increasing complexity in large-scale microfluidic processors, bubble logic provides an on-chip process control mechanism integrating chemistry and computation.

Acknowledgments

Special thanks to Dr. Suman Chakraborty (Indian Institute of Technology Kharagpur), Dr. Nahui Kim (University of California, Irvine), and Marleen Madou.

Questions

Questions by Omid Rohani, UC Irvine

6.1: We partially fill an open vessel with a liquid and rotate it at a constant angular velocity ω about a vertical axis. After some time the liquid also rotates at the same angular velocity as the vessel and the surface of the liquid develops into a steady shape. What is the equation that describes the free surface of the liquid?

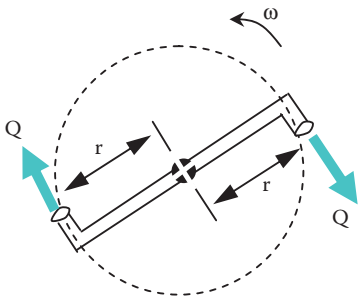
6.2: Consider the flow of blood through a micro or nano channel. List 4 assumptions that introduce potential errors when modelling blood flows through such a micro channel using the Navier-Stokes equations.

6.3: What conclusions can one draw regarding liquid flows in micro and nano channels based on the currently available data in the literature? What do you know about turbulence in the macro scale and in the micro and nano scale? How can one enhance good mixing of two chemicals in a microreactor?

6.4: A lawn sprinkler consists of two arms equipped with nozzles at right angles which rotate about a pivot in a horizontal plane. Assuming the flow rate through each nozzle equals Q and the nozzle's exit area is A, find the velocity of the water leaving the nozzle, in two cases:

(a) There is no friction at the pivot.

(b) There is a constant frictional torque T_f at the pivot.



6.5: Derive an equation for the flow rate Q as a function of U, b, $\frac{dp}{dx}$, η for fully developed Couette shear flow, where U is the velocity of a moving plate and b is the height of the channel between this moving plate and a fixed bottom plate. We will assume that the flow is in the x-direction and b is in the y-direction. The viscosity is represented by η and $\frac{dp}{dx}$ is the pressure gradient in the x-direction.

6.6: What are the advantages of microfluidics? Give an example application.

6.7: Compare the breakdown of continuum theory for gasses and liquids.

6.8: What techniques can be implemented for mixing two reagents in the microscale (in low Reynolds number liquids)?

6.9: Given is a mixture of two proteins of similar molecular mass. Protein A comes with a MW = 12,400 and a pI = 10.7 and protein B has a MW = 12,800 and a pI = 7.5 (pI is the isoelectric point). You are told that capillary zone electrophoresis, CZE, is the best way to separate these two proteins.

(a) Do you agree with the suggestion of using CZE or would you employ another method? Briefly support your decision and your choice of separation method.
(b) Explain the principle behind capillary electrophoresis and draw a schematic illustrating its operation.
(c) What will be the elution order if you perform the separation using a buffer with a pH of 5.0? Why?

6.10: When do the macroscopic laws of gas flow begin to fail and when does the no-slip condition cease to apply?

6.11: Explain some of the changes in the behavior of a liquid when the channel diameter decreases to below ten molecular diameters (the channel diameter of the order of a few nanometers).

6.12: What is electrophoresis? Explain why DNA strands of varying size cannot be separated by electrophoresis in a free solution.

6.13: Describe how electro-osmotic phenomena can cause a pumping effect in a fused silica micro capillary.

6.14: Using the Hagen-Poiseuille law for flow rate, explain why conventional pumping is not useful for the delivery of small samples? What pumping method do you suggest for delivery of small samples? How does this method change the flow profile?

6.15: In a centrifugal fluidic platform (CD-fluidics), what constitutes the pumping force, and how are valves for liquids implemented?

6.16: A simple and accurate viscometer can be made from a length of capillary tubing. If the flow rate and pressure drop are measured, and the tube geometry is known, the viscosity can be computed. A test of a certain liquid in a capillary viscometer gave the following data: flow rate: 880 mm^3/s; tube diameter: 500 μm; tube length: 1 m; pressure drop: 1.0 MPa. Determine the viscosity of the liquid.

6.17: The accepted transition Reynolds number for flow in a circular pipe is $Re \approx 2300$. For flow through a 6 cm diameter pipe, at what velocity will this occur at 20°C for (a) airflow and (b) water flow?

6.18: List five electrokinetic effects taken advantage of in the MEMS field and list one application for each approach.

6.19: Why are more theoretical plates achievable in capillary electrophoresis than in the following separation techniques?
(a) Slab electrophoresis
(b) Liquid chromatography

6.20: Compare/contrast how separation is achieved in chromatography and electrophoresis.

6.21: List 6 different methods to propel fluids through micro channels.

6.22: Why is there a lower limit to the size of an uncharged particle one can move with dielectrophoresis (~14 nm) and not on the size of a charged particle one can move in electrophoresis? What types of particles (size and charge) can one move with electro-osmosis?

6.23: If a 100 bp long DNA is to be sequenced in a microfabricated channel, calculate the length of a channel needed to get single base pair separation. What voltage (E field) would you pick to perform the separation in a polyacrylamide gel. Make the necessary assumptions! The width and height of the channel is 20 μm × 20 μm.
Thanks to Dr. Rashid Bashir, Purdue University.

6.24: Demonstrate that $\delta/\delta_T \sim Pr^{1/2}$.

Further Reading

Abraham, S., E. H. Jeong, T. Arakawa, S. Shoji, K. C. Kim, I. Kim, and J. S. Go. 2006. Microfluidics assisted synthesis of well-defined spherical polymeric microcapsules and their utilization as potential encapsulants. *Lab Chip* 6:752–56.

Anna, S. L., N. Bontoux, and H. A. Stone, 2003. Formation of dispersions using "flow focusing" in microchannels. *Appl Phys Lett* 82:364–66.

Belder, D. 2005. Microfluidics with droplets. *Angewandte Chemie* 44:3521–22.

Cubaud, T., M. Tatineni, X. Zhong, and C. M. Ho. 2005. Bubble dispenser in microfluidic devices. *Phys Rev E* 72:4.

Cubaud, T., and C. M. Ho. 2004. Transport of bubbles in square microchannels. *Phys Fluids* 16:4575–85.

Fair, R.B., 2007. Digital microfluidics: is a true lab-on-a-chip possible? *Microfluidics Nanofluidics* 3:245–81.

Forsberg, F., J. B. Liu, D. A. Merton, N. M. Rawool, and B. B. Goldberg. 1995. Parenchymal enhancement and tumor visualization using a new sonographic contrast agent. *J Ultrasound Med* 14:949–57.

Garstecki, P., I. Gitlin, W. DiNuzio, and G. M. Whitesides. 2004. Formation of monodisperse bubbles in a microfluidic flow-focusing device. *Appl Phys Lett* 85:2649–51.

Gerdts, C. J., V. Tereshko, M. K. Yadav, I. Dementieva, F. Collart, A. Joachimiak, R. C. Stevens, P. Kuhn, A. Kossiakoff, and R. F. Ismagilov. Time-controlled microfluidic seeding in nL-volume droplets to separate nucleation and growth stages of protein crystallization. *Angewandte Chemie* 45:8156–60.

Goldberg, B. B., J. S. Raichlen, and F. Forsberg. 2001. *Ultrasound contrast agents*. London: Dunitz Ltd.

Gordillo, J. M., Z. Cheng, A. M. Ganan-Calvo, M. Márquez, and D. A. Weitz. 2004. A new device for the generation of microbubbles. *Phys Fluids* 16:2828–34.

Gunther A., and K. F. Jensen. 2007. Multiphase microfluidics: from flow characteristics to chemical and materials synthesis. *Lab Chip* 7:935.

Hettiarachchi, K., E. Talu, M. L. Longo, P. A. Dayton, and A. P. Lee. 2007. On-chip generation of microbubbles as a practical technology for manufacturing contrast agents for ultrasonic imaging. *Lab Chip* 7:463–68.

Jensen, K., and A. P. Lee. 2004. The science and applications of droplets in microfluidic devices. *Lab Chip* 4:31N–32N.

Kobayashi, I., K. Uemura, and M. Nakajima. 2007. Formulation of monodisperse emulsions using submicron-channel arrays. *Colloids Surfaces A Physicochem Eng Aspects* 296:285–89.

Liu, K., H. J. Ding, J. Liu, Y. Chen, and X. Z. Zhao. 2006. Shape-controlled production of biodegradable calcium alginate gel microparticles using a novel microfluidic device. *Langmuir* 22:9453–57.

Nisisako, T., S. Okushima, and T. Torii. 2005. Controlled formulation of monodisperse double emulsions in a multiple-phase microfluidic system. *Soft Matter* 1:23–27.

Nisisako, T., and T. Torii. 2007. Formation of biphasic Janus droplets in a microfabricated channel for the synthesis of shape-controlled polymer microparticles. *Adv Mater* 19:1489–93.

Sakai, S., I. Hashimoto, and K. Kawakami. 2007. Agarose-gelatin conjugate for adherent cell-enclosing capsules. *Biotech Lett* 29:731–35.

Sakai, S., I. Hashimoto, and K. Kawakami. 2006. Usefulness of flow focusing technology for producing subsieve-size cell enclosing capsules: application for agarose capsules production. *Biochem Eng J* 30:218–21.

Song, H., D. L. Chen, and R. F. Ismagilov. 2006. Reactions in droplets in microfluidic channels. *Angewandte Chemie* 45:7336–56.

Tan, Y. C., V. Cristini, and A. P. Lee. 2006. Monodispersed microfluidic droplet generation by shear focusing microfluidic device. *Sensors Actuators B Chem* 114:350–56.

Teh, S.-Y., R. Lin, L. Hung, and A. Lee. 2008. Droplet microfluidics. *Lab Chip* 8:198–220.

Utada, A. S., E. Lorenceau, D. R. Link, P. D. Kaplan, H. A. Stone, and D. A. Weitz. 2005. Monodisperse double emulsions generated from a microcapillary device. *Science* 308:537–41.

Wei, K., and S. Kaul. 1997. Recent advances in myocardial contrast echocardiography. *Curr Opin Cardiol* 12:539–46.

Woodward, A., T. Cosgrove, J. Espidel, P. Jenkins, and N. Shaw. 2007. Monodisperse emulsions from a microfluidic device, characterised by diffusion NMR. *Soft Matter* 3:627–33.

Xu, J. H., S. W. Li, Y. J. Wang, and G. S. Luo. 2006. Controllable gas-liquid phase flow patterns and monodisperse microbubbles in a microfluidic T-junction device. *Appl Phys Lett* 88:3.

Yobas, L., S. Martens, W. L. Ong, and N. Ranganathan. 2006. High-performance flow-focusing geometry for spontaneous generation of monodispersed droplets. *Lab Chip* 6:1073–79.

Zheng, B., C. J. Gerdts, and R. F. Ismagilov. 2005. Using nanoliter plugs in microfluidics to facilitate and understand protein crystallization. *Curr Opin Struct Biol* 15:548–55.

Zheng, B., L. S. Roach, and R. F. Ismagilov. 2003. Screening of protein crystallization conditions on a microfluidic chip using nanoliter-size droplets. *J Am Chem Soc* 125: 11170–71.

Zhou, C. F., P. T. Yue, and J. J. Feng. 2006. Formation of simple and compound drops in microfluidic devices. *Phys Fluids* 18:14.

References

1. Denn, M. M. 1980. *Process fluid mechanics.* Englewood Cliffs, NJ: Prentice Hall.
2. Potter, M. C., and J. F. Foss. 1982. *Fluid mechanics.* Okemos, MI: Great Lakes Press.
3. Barron, A. E., and Blanch, H. 1995. DNA separations by slab gel and capillary electrophoresis: Theory and practice. *Separation and Purification Methods* 24:1–118.
4. White, F. M. 1994. *Fluid mechanics.* New York: McGraw-Hill.
5. Allen, T. J., and R. L. Ditsworth. 1972. *Fluid mechanics.* New York: McGraw-Hill.
6. Granger, R. A. 1995. *Fluid mechanics.* New York: Dover Publications.
7. Ozisik, M. N. 1977. *Basic heat transfer.* New York: McGraw-Hill.
8. Corcoran, P., H. V. Shurmer, and J. W. Gardner. 1993. Integrated tin oxide sensors of low power consumption for use in gas and odour sensing. *Sensors Actuators B* B15–16:32–37.
9. Coulson, J. M., and J. F. Richardson. 1999. *Chemical engineering.* New York: Pergamon.
10. Stokes, G. G. S. 1851. On the effect of the internal friction of fluids on the motion of pendulums. *Trans Cambridge Phil Soc* 9:8–106.
11. Poiseuille, J. L. M. 1840. Experimental investigations upon the flow of liquids in tubes of very small diameter. *Rheological Memoirs* 1.
12. Pfahler, J., J. Harley, and H. Bau. 1991. *ASME 1991: micromechanical sensors, actuators, and systems.* Atlanta: ASME, 49–60.
13. Gravesen, P., J. Branebjerg, and O. S. Jensen. 1993. Microfluidics: a review. *J Micromech Microeng* 3:168–82.
14. Liu, J., and Y.-C. Tai. 1995. *Proceedings: IEEE micro electro mechanical systems (MEMS '95).* Amsterdam: IEEE, 209–15.
15. Arkilic, E. B., K. S. Breuer, and M. A. Schmit. 1994. *Applications of microfabrication to fluid mechanics.* Chicago: ASME, 57–66.
16. Wong, C. C., D. R. Adkins, G. C. Frye-Mason, M. L. Hudson, R. Kottenstette, C. M. Matzke, J. N. Shadid, and A. G. Salinger. 1999. *Proceedings: microfluidic devices and systems II.* Santa Clara, CA: SPIE, 120–29.
17. Flik, M. I., B. I. Choi, and K. E. Goodson. 1991. *ASME 1991: micromechanical sensors, actuators, and systems.* Atlanta: ASME, 31–46.

18. Harley, J., H. Bau, J. N. Zemel, and V. Dominko. 1989. *Proceedings: IEEE micro electro mechanical systems (MEMS '89)*. Salt Lake City, UT: IEEE, 25–28.
19. Israelachvili, J. N. 1986. Measurement of the viscosity of liquids in very thin films. *Colloid Interface Sci* 110:263–71.
20. Tretheway, D. C., and C. D. Meinhart. 2002. Apparent fluid slip at hydrophobic microchannel walls. *Phys Fluids* 14:L9.
21. Pit, R., H. Hervet, and L. Léger. 2000. Direct experimental evidence of slip in hexadecane: solid interfaces. *Phys Rev Lett* 85:980–83.
22. Zhu, Y., and S. Granick. 2001. Rate-dependent slip of Newtonian liquid at smooth surfaces. *Phys Rev Lett* 87:096105.
23. Craig, V. S., J. Neto, and C. Williams. 2001. Shear-dependent boundary slip in an aqueous Newtonian liquid. *Phys Rev Lett* 87:054504.
24. Israelachvili, J. N. 1992. *Intermolecular and surface forces: with applications to colloidal and biological systems (colloid science)*. New York: Academic Press.
25. Gee, M. L., P. M. McGuiggan, J. N. Israelachvili, and A. M. Homola. 1990. Liquid to solidlike transitions of molecularly thin films under shear. *J Chem Phys* 93:1895–906.
26. Ho, C. M. 2001. 14th IEEE International Conference MEMS, Interlaken, Switzerland, 375.
27. Barthlott, W., and C. Neinhuis. 1997. The purity of sacred lotus or escape from contamination in biological surfaces. *Planta* 202:1–8.
28. Colgate, E., and H. Matsumoto. 1990. An investigation of electro-wetting-based microactuation. *J Vac Sci Technol* A8:3625–33.
29. Kenis, P. J. A., R. F. Ismagilov, and G. M. Whitesides. 1999. Microfabrication inside capillaries using multiphase laminar flow patterning. *Science* 285:83–85.
30. Aref, H. 1984. Stirring by chaotic advection. *J Fluid Mech* 143:1–21.
31. Whitesides, G. M. 2003. The "right" size in nanobiotechnology. *Nat Biotech* 21:1161–65.
32. Ismagilov, R. F., A. D. Stroock, P. J. A. Kenis, and G. Whitesides. 2000. Experimental and theoretical scaling laws for transverse diffusive broadening in two-phase laminar flows in microchannels. *Appl Phys Lett* 76:2376–78.
33. Brody, J. P., P. Yager, R. E. Goldstein, and R. H. Austin. 1996. Biotechnology at low Reynolds numbers. *Biophys J* 71:3430–41.
34. Stroock, A. D., S. K. W. Dertinger, A. Ajdari, I. Mezi, H. A. Stone, and G. M. Whitesides. 2002. Chaotic mixer for microchannels. *Science* 295:647–651.
35. Knight, J. B., A. Vishwanath, J. P. Brody, and R. H. Austin. 1998. Hydrodynamic focusing on a silicon chip: mixing nanoliters in microseconds. *Phys Rev Lett* 80:3863–66.
36. Purcell, E. M. 1977. Life at low Reynolds number. *Am J Phys* 45:3–11.
37. Nguyen, N. T., and W. Zhigang. 2005. Micromixers—a review. *J Micromech Microeng* 15:R1–R16.
38. Virtanen, R. 1974. Zone electrophoresis in a narrow-bore tube employing potentiometric detection. *Acta Polytech Scand* 123:1–67.
39. Jorgenson, J. W., and K. D. Lukacs. 1981. High resolution separations based on electrophoresis and electroosmosis. *J Chromatogr* 218:209–14.
40. Barron, A. E., and H. Blanch. 1995. DNA separations by slab gel and capillary electrophoresis: theory and practice. *Separation Purification Methods* 24:1–118.
41. Cobb, K. A., V. Dolnik, and M. Novotny. 1990. Electrophoretic separations of proteins in capillaries with hydrolytically stable surface structures. *Anal Chem* 62:2478–83.
42. Jorgenson, J. W., and E. J. Guthrie. 1983. Liquid chromatography in open-tubular columns. *J. Chromatog* 255: 335–48.
43. Harrison, D. J., Z. Fan, K. Fluri, and K. Seiler. 1994. *Technical digest: 1994 solid-state sensor and actuator workshop*. Hilton Head Island, SC.
44. Jacobson, S. C., R. Hergenröder, A. W. Moore, and J. M. Ramsey. 1994. *Technical digest: 1994 solid-state sensor and actuator workshop*. Hilton Head Island, SC.
45. Jacobson, S. C., J. Kutter, C. Culbertson, and J. Ramsey. 1998. *Micro total analysis systems '98*. Banff, Canada, 315.
46. Jacobson, S. C., and J. M. Ramsey. 1998. Microchip structures for submillisecond electrophoresis. Microchip structures for submillisecond electrophoresis. *Anal Chem* 70: 3476–80.
47. Jacobson, S. C., C. T. Culbertson, and J. M. Ramsey. 1998. *c-state sensor and actuator workshop*. Hilton Head Island, SC.
48. Pace, S. 1990. E. I. Du Pont De Nemours & Co. US Patent 4,908,112.
49. Soane, D. S., and Z. Soane. 1998. Soane Biosciences . US Patent 5,750,015.
50. Manz, A., J. C. Fettinger, E. Verpoorte, H. Ludi, H. M. Widmer, and D. J. Harrison. 1991. Micromachining of monocrystalline silicon and glass for chemical analysis systems: a look into next century's technology or just a fashionable crazy? *Trends Anal Chem* 10:144.
51. Manz, A., D. J. Harrison, E. M. J. Verpoort, J. C. Fettinger, A. Paulus, H. Ludi, and H. M. Widmer. 1992. Planar chips technology for miniaturization and integration of separation techniques into monitoring systems: capillary electrophoresis on a chip. *J Chromatogr* 593:253–58.
52. Ramsey, J. M. 1999. Lockheed Martin Energy Systems. US Patent 6,001,229.
53. Ramsey, J. M. 2000. Lockheed Martin Energy Research Corporation. US Patent 6,010,607.
54. Effenhauser, C. S., A. Manz, and H. M. Widmer. 1995. Manipulation of sample fractions on a capillary electrophoresis chip. *Anal Chem* 67:2284–87.
55. Burggraf, N., A. Manz, E. Verpoorte, C. S. Effenhauser, H. M. Widmer, and N. F. de Rooij. 1994. A novel approach to ion separation in solution: synchronization cyclic capillary electrophoresis (SCCE). *Sensors Actuators B* 20:103.
56. Bousse, L., A. R. Kopf-Sill, and J. W. Parce. 1998. *Systems and technologies for clinical diagnostics and drug discovery*. Ed. G. E. Cohn. San Jose, CA: SPIE.
57. Haab, B. B., and R. A. Mathies. 1998. *Systems and technologies for clinical diagnostics and drug discovery*. Ed. G. E. Cohn. San Jose, CA: SPIE.
58. Wahl, M., R. Erdmann, K. Lauritsen, and H.-J. Rahn. 1998. *Systems and technologies for clinical diagnostics and drug discovery*. Ed. G. E. Cohn. San Jose, CA: SPIE.
59. Righetti, P. G. 1990. *Immobilized pH gradients: theory and modelling*. Amsterdam: Elsevier.
60. Herr A. E., J. I. Molho, J. G. Santiago, T. W. Kenny, D. A. Borkholder, G. J. Kintz, P. Belgrader, and M. A. Northrup. 2000. *Technical digest: 2000 solid-state sensor and actuator workshop*. Hilton Head Island.
61. Heller, M. J., A. H. Forster, and E. Tu. 2000. Active microelectronic chip devices which utilize controlled electrophoretic fields for multiplex DNA hybridization and other genomic applications. *Electrophoresis* 21: 157–64.

62. Martin, Y., and H. K. Wickramasinghe. 1987. Magnetic imaging by "force microscopy" with 1000 Å resolution. *Appl Phys Lett* 50:1455–57.
63. Pohl, H. A. 1951. The motion and precipitation of suspensoids in divergent electric fields. *J Appl Phys* 22:869–71.
64. Pohl, H. A. 1978. *Dielectrophoresis*. Cambridge, UK: Cambridge University Press.
65. Pethig, R., J. P. H. Burt, A. Parton, N. Rizvi, M. S. Talary, and J. A. Tame. 1998. Development of biofactory-on-a-chip technology using excimer laser micromachining. *J Micromech Microeng* 8:57–63.
66. Pethig, R., and G. H. Markx. 1997. Applications of dielectrophoresis in biotechnology. *TIBTECH* 15:426–32.
67. Huang, Y., X. B. Wang, F. F. Becker, and P. R. C. Gascoyne. 1997. Introducing dielectrophoresis as a new force field for field-flow-fractionation. *Biophys J* 73:1118–29.
68. Wang, X.-B., J. Vykoukal, F. F. Becker, and P. R. C. Gascoyne. 1998. Separation of polystyrene microbeads using dielectrophoretic/gravitational field-flow-fractionation. *Biophys J* 74:2689–701.
69. Suehiro, J., and R. Pethig. 1998. The dielectrophoretic movement and positioning of a biological cell using a three-dimensional grid electrode system. *J Phys D Appl Phys* 31:3298–305.
70. Park, B. Y., and M. J. Madou. 2005. 3-D electrode designs for flow-through dielectrophoretic systems. *Electrophoresis* 26:3745–57.
71. Martinez-Duarte, R., H. A. Rouabah, N. G. Green, M. Madou, and H. Morgan. 2007. *Proceedings of the Eleventh International Conference in Miniaturized Systems for Chemistry and Life Sciences: microTAS 2007*. Paris.
72. Wang, L., L. Flanagan, and A. P. Lee. 2007. Side-wall vertical electrodes for lateral field microfluidic applications. *J Microelectromech Systems* 16:454–61.
73. Voldman, J., M. L. Gray, M. Toner, and M. A. Schmidt. 2002. A microfabrication-based dynamic array cytometer. *Anal Chem* 74:3984–90.
74. Iliescu, C., L. Yu, G. Xu, and F. E. H. Tay. 2006. A dielectrophoretic chip with a 3-D electric field gradient. *J Microelectromech Systems* 15:1506–13.
75. Lapizco-Encinas, B. H., B. A. Simmons, E. B. Cummings, and Y. Fintschenko. 2004. Dielectrophoretic concentration and separation of live and dead bacteria in an array of insulators. *Anal Chem* 76:1571–79.
76. Fatoyinbo, H. O., D. Kamchis, R. Whattingham, S. L. Ogin, and M. P. Hughes. 2005. A high-throughput 3-D composite dielectrophoretic separator. *IEEE Trans Biomed Eng* 52:1347–49.
77. Becker, F. F., X. B. Wang, Y. Huang, R. Pethig, J. Vykoukal, and P. R. C. Gascoyne. 1994. The removal of human leukaemia cells from blood using interdigitated microelectrodes. *J Phys D Appl Phys* 27:2659–62.
78. Burt, J. P. H., R. Pethig, and M. S. Talary. 1998. Microelectrode devices for manipulating and analysing bioparticles. *Trans Inst MC* 19:1–9.
79. Rousselet, J., G. H. Markx, and R. Pethig. 1998. Separation of erythrocytes and latex beads by dielectrophoretic levitation and hyperlayer field-flow fractionation. *Colloids Surfaces A* 140:209–16.
80. Markx, G. H., and R. Pethig. 1995. Dielectrophoretic separation of cells: continuous separation. *Biotech Bioeng* 45:337–43.
81. Talary, M. S., J. P. H. Burt, J. A. Tame, and R. Pethig. 1996. Electromanipulation and separation of cells using travelling electric fields. *J Phys D Appl Phys* 29:2198–203.
82. Goater, A. D., J. P. H. Burt, and R. Pethig. 1997. A combined travelling wave dielectrophoresis and electrorotation device applied to the concentration and viability determination of cryptosporidium. *J Phys D Appl Phys* 30:L65–69.
83. Müller, T., S. Fiedler, T. Schnelle, K. Ludwig, H. Jung, and G. Fuhr. 1996. *Biotech Tech* 10:221–26.
84. Batchelder, J. S. 1983. Method and apparatus for dielectrophoretic manipulation of chemical species. US Patent 4,390,403.
85. Huang, Y., J. Yang, X.-B. Wang, F. F. Becker, and P. R. C. Gascoyne. 1999. The removal of human breast cancer cells from hematopoietic CD34+ stem cells by dielectrophoretic field-flow-fractionation. *J Hematother Stem Cell Res* 8:481–90.
86. Yang, J., Y. Huang, X. Wang, X.-B. Wang, F. F. Becker, and P. R. C. Gascoyne. 1999. Dielectric properties of human leukocyte subpopulations determined by electrorotation as a cell separation criterion. *Biophys J* 76:3307–14.
87. Prins, M. W. J., W. J. J. Welters, and J. W. Weekamp. 2001. Fluid control in multichannel structures by electrocapillary pressure. *Science* 291:277–80.
88. Asakuma, N., T. Fukui, M. Toki, K. Awazu, and H. Imai. 2003. Photoinduced hydroxylation at ZnO surface. *Thin Solids Films* 445:284–87.
89. Zhang, Z., H. Chen, J. Zhong, G. Saraf, and Y. Lu. 2007. Fast and reversible wettability transitions on ZnO nanostructures. *J Electr Mater* 36:895–99.
90. Feng, X., L. Feng, M. Jin, J. Zhai, L. Jiang, and D. Zhu. 2004. Reversible super-hydrophobicity to super-hydrophilicity transition of aligned ZnO nanorod films. *J Am Chem Soc* 126:62–63.
91. Sun, R. D., A. Nakajima, A. Fujishima, T. Watanabe, and K. Hashimoto. 2001. Photoinduced surface wettability conversion of ZnO and TiO_2. Thin films. *J Phys Chem B* 105:1984–90.
92. Madou, M. J., and G. J. Kellogg. 1998. *Systems and technologies for clinical diagnostics and drug discovery*. Eds. G. E. Cohn and A. Katzir. San Jose, CA: SPIE.
93. McNeely, M. R., M. K. Spute, N. A. Tusneem, and A. R. Oliphant. 1999. *Microfluidic devices and systems II*. Eds. C. H. Ahn and A. B. Frazier. Santa Clara, CA: SPIE.
94. Madou, M. J., Y. Lu, S. Lai, Y.-J. Juang, L. J. Lee, and S. Daunert. 2000. *Technical digest: 2000 solid sensor and actuator workshop*. Hilton Head Island, SC.
95. Kim, N., C. M. Dempsey, C. J. Kuan, J. V. Zoval, E. O'Rourke, G. Ruvkun, M. J. Madou, and J. Y. Sze. 2007. Gravity force transduced by the MEC-4/MEC-10 DEG/ENaC channel modulates DAF-16/FoxO activity in Caenorhabditis elegans. *Genetics* 177:835–45.
96. Kim, N., C. M. Dempsey, J. V. Zoval, J. Y. Sze, and M. J. Madou. 2007. Automated microfluidic compact disc (CD) cultivation system of Caenorhabditis elegans. *Sensors Actuators B Chemical* 122:511–18.
97. Madou, M. J., J. Zoval, G. Jia, H. Kido, J. Kim, and N. Kim. 2006. Lab on a CD. *Ann Rev Biomed Eng* 8:601–28.
98. Manz, A., E. Verpoorte, C. S. Effenhauser, N. Burggraf, D. E. Raymond, D. J. Harrison, and H. M. Widmer. 1993. Miniaturization of separation techniques using planar chip technology. HRC *J High Resol Chromatogr* 16:433–36.
99. Manz, A., N. Graber, and H. M. Widmer. 1990. Miniaturized total chemical analysis systems: a novel concept for chemical sensing. *Sensors Actuators B* B1:244–48.
100. Nölting, B. 2005. *Methods in modern biophysics*. Berlin: Springer.

101. Harris, D. C. 1999. *Quantitative chemical analysis*. New York: W. H. Freeman.
102. Golay, M. J. E. 1958. *Gas chromatography*. New York: Academic Press.
103. Taylor, G. I. 1953. Dispersion of soluble matter in solvent flowing slowly through a tube. *Proc Roy Soc A* 219:186–203.
104. Wong, C. C., D. R. Adkins, G. C. Frye-Mason, M. L. Hudson, R. Kottenstette, C. M. Matzke, J. N. Shadid, and A. G. Salinger. 1999. *Microfluidic devices and systems II*. Santa Clara, CA: SPIE.
105. Molho, J. I., A. E. Herr, B. P. Mosier, J. G. Santiago, T. W. Kenny, R. A. Brennen, and G. Gordon. 2000. A low dispersion turn for miniaturized electrophoresis. *Technical digest: 2000 solid-state sensor and actuator workshop*, Hilton Head, SC.
106. Culbertson, C. T., S. C. Jacobson, and J. M. Ramsey. 2000. *Technical digest: 2000 solid-state sensor and actuator workshop*. Hilton Head, SC.
107. Manz, A., and W. Simon. 1987. Potentiometric detector for fast high-performance open-tubular column liquid chromatography. *Anal Chem* 59:74–79.
108. Manz, A., J. C. Fettinger, E. Verpoorte, and H. Ludi. 1991. Micromachining of monocrystalline silicon and glass for chemical analysis systems: a look into the next century's technology or just a fashionable craze. *Trends Anal Chem* 10:144–49.
109. Fair, R. B. 2007. Digital microfluidics: is a true lab-on-a-chip possible? *Microfluid Nanofluid* 3:245–81.
110. De Menech, M., P. Garstecki, F. Jousse, and H. A. Stone. 2008. Transition from squeezing to dripping in a microfluidic T-shaped junction. *J Fluid Mech* 595:141–61.
111. Anna, S. L., N. Bontoux, and H. A. Stone. 2003. Formation of dispersions using "flow focusing" in microchannels. *Appl Phys* Lett 82:364–66.
112. Zhou, C. F., P. T. Yue, and J. J. Feng. 2006. Formation of simple and compound drops in microfluidic devices. *Physi Fluids* 18:14.
113. Tan, Y. C., V. Cristini, and A. P. Lee. 2006. Monodispersed microfluidic droplet generation by shear focusing microfluidic device. *Sensors Actuators B Chem* 114:350–56.
114. Woodward, A., T. Cosgrove, J. Espidel, P. Jenkins, and N. Shaw. 2007. Monodisperse emulsions from a microfluidic device, characterised by diffusion NMR. *Soft Matter* 3:627–33.
115. Yobas, L., S. Martens, W. L. Ong, and N. Ranganathan. 2006. High-performance flow-focusing geometry for spontaneous generation of monodispersed droplets. *Lab Chip* 6:1073–79.
116. Nisisako, T., and T. Torii. 2007. Formation of biphasic Janus droplets in a microfabricated channel for the synthesis of shape-controlled polymer microparticles. *Adv Mater* 19:1489–93.
117. Nisisako, T., S. Okushima, and T. Torii. 2005. Controlled formulation of monodisperse double emulsions in a multiple-phase microfluidic system. *Soft Matter* 1:23–27.
118. Utada, A. S., E. Lorenceau, D. R. Link, P. D. Kaplan, H. A. Stone, and D. A. Weitz. 2005. Monodisperse double emulsions generated from a microcapillary device. *Science* 308:537–41.
119. Gordillo, J. M., Z. D. Cheng, A. M. Ganan-Calvo, M. Marquez, and D. A. Weitz. 2004. A new device for the generation of microbubbles. *Phys Fluids* 16:2828–34.
120. Garstecki, P., I. Gitlin, W. DiLuzio, G. M. Whitesides, E. Kumacheva, and H. A. Stone. 2004. Formation of monodisperse bubbles in a microfluidic flow-focusing device. *Appl Physi Lett* 85:2649–51.
121. Cubaud, T., and C. M. Ho. 2004. Transport of bubbles in square microchannels. *Phys Fluids* 16:4575–85.
122. Cubaud, T., M. Tatineni, X. L. Zhong, and C. M. Ho. 2005. Bubble dispenser in microfluidic devices. *Phys Rev* E72:4.
123. Hettiarachchi, K., E. Talu, M. L. Longo, P. A. Dayton, and A. P. Lee. 2007. On-chip generation of microbubbles as a practical technology for manufacturing contrast agents for ultrasonic imaging. *Lab Chip* 7:463–68.
124. Prakash, M., and N. Gershenfeld. 2007. Microfluidic bubble logic. *Science* 315:832–35.
125. Hagen, G. H. L. 1839. III Über die Bewgung des Wassers in engen cylindrischen Röhren, *Poggendorfs Annalen der Physik und Chemie* 2(46):423–42.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

7

Electrochemical and Optical Analytical Techniques

Outline

Introduction

Intermolecular Forces

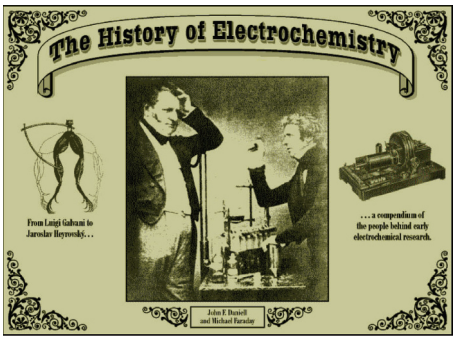
Electrochemistry

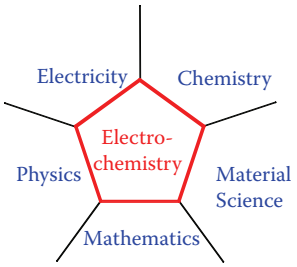
Optical Spectroscopy

Comparison of Optical versus Electrochemical Sensors

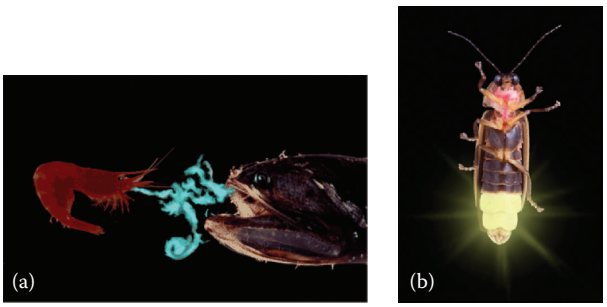
Questions

References





The History of Electrochemistry (calendar from http://www.basinc.com).



(a) Shrimp vomiting a bioluminescent substance toward a predator. (b) Firefly exhibiting bioluminescence.

Introduction

In all electrochemical and many optical sensors, molecules interact with a solid surface and generate an analytical signal that can be used to identify and sometimes quantify an unknown analyte. This chapter begins with a detailed study of the nature of the attractive and repulsive forces

at an interface between a liquid and a solid (mostly in the form of small particles and bulk electrodes). Forces at the liquid/solid interface are the same as those that control the interaction between liquid molecules but are modulated by the presence of the solid. Intermolecular nonbonding forces between molecules, ions, or molecules and ions are much weaker than intramolecular bonding forces, such as covalent bonds within a molecule. Nevertheless, intermolecular forces sum to an impressive total binding energy, often several hundreds of kilojoules per mole, when many are present within a macromolecule or in a solid. The intermolecular forces we will study include the classical (coulomb) electrostatic forces, van der Waals interactions, and the quantum mechanical Casimir force. These forces, when exerted between individual molecules, extend typically over a few nanometers only, but when molecules come together into condensed phases, the force range at the surface is extended to several tens of nanometers. Intermolecular forces explain the electrical double layer (EDL) at the interface of liquid/solid surfaces, surface tension, viscosity, vapor pressure, boiling points, and many other important phenomena, including DNA hybridization and self-assembly. Hydrophilic and hydrophobic interactions, also based on intermolecular forces, are treated in Chapter 6 on fluidics.

Electrochemistry is the study of reactions taking place at the interface of a solid conductor and an electrolyte. Using sensor examples, we introduce some of the most important concepts in electrochemistry, i.e., the EDL, potentiometry, voltammetry, two- and three-electrode systems, Marcus theory of electron transfer, reaction rate- and diffusion rate-controlled electrochemical reactions, and ultramicroelectrodes (UMEs). We also learn how different materials, such as metals, semiconductors, solid electrolytes, and mixed conductors, exposed to a solution offer the possibility for building different types of electrochemical sensors. Besides insights in electrochemical sensor construction, this section prepares the reader for an appreciation of the role of electrochemistry in the manufacture of microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS), i.e., in electrochemical machining as covered in Volume II, Chapter 4 (electrochemical etching, a subtractive technique) and Volume II, Chapter 8 (electrochemical deposition, a forming technique). We conclude the electrochemistry section with more electrochemical sensor examples, including ion-sensitive field-effect transistors (ISFETs), potentiometric and amperometric immunosensors, and glucose sensors.

Many researchers use MEMS and NEMS to miniaturize optical components or whole instruments for absorption, luminescence, or phosphorescence measurements. The latter techniques, like the electrochemical methods reviewed, constitute a branch of analytical chemistry, a branch that is concerned with the production, measurement, and interpretation of electromagnetic data arising from either emission or absorption of radiant energy by matter. We compare the sensitivity of the various optical sensing techniques and analyze how amenable they are to miniaturization (scaling laws).

We end this chapter by comparing the merits and problems associated with electrochemical and optical measuring techniques.

Intermolecular Forces

Introduction

For the nonchemist, a few simple chemistry concepts are recalled here. A covalent bond, holding a molecule together, is a bonding or intramolecular force, whereas the attraction between molecules is a nonbonding or intermolecular force. Ionic chemical bonding results from the maximization of the electrostatic forces between large numbers of cations and anions in ionic solids. Intermolecular forces between molecules, between ions in solution, or between molecules and ions are much weaker than intramolecular forces (e.g., leading to a binding energy of only 16 kJ/mol vs. 431 kJ/mol for HCl) (see Figure 7.1).

Covalent bonds (300–400 kJ/mol) determine molecular architecture, whereas noncovalent bonds, typically 10–100 times weaker, determine molecular conformations. Being individually weak, noncovalent interactions nevertheless sum to an impressive total, often several hundreds of kilojoules per mole, when many are present within a given

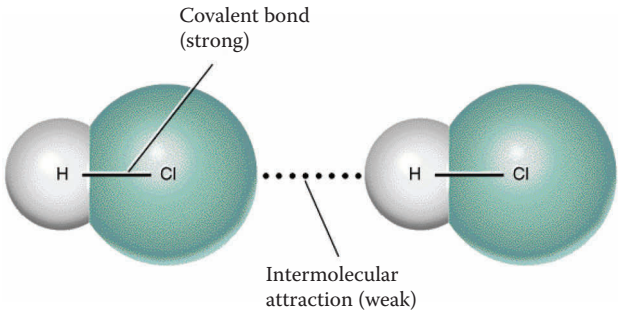


FIGURE 7.1 In a HCl molecule, there is a bonding or intramolecular bond that is covalent and strong and a nonbonding or intermolecular hydrogen bond that is weak.

macromolecule, between macromolecules in a solution, or between the building blocks making up a solid. The intermolecular energy term (used in the analytical calculations involving energy) usually includes the classical (coulomb) electrostatic and van der Waals interactions. The nonbonding forces we will analyze below are:

- Charge-charge interactions (Coulomb's law in a dielectric medium)
- Permanent and induced dipole interactions (van der Waals interactions):
  - Charge-dipole interactions
  - Dipole-dipole interactions
  - Charge-induced dipole interactions
  - Dipole-induced dipole interactions
  - Dispersion (induced dipole-induced dipole) interactions
- Hydrogen bonds: donor group-acceptor group interactions (a very important case of dipole-dipole interactions)
- Casimir force: a quantum mechanical effect (a special type of van der Waals interaction, i.e., the retarded van der Waals force also known as the *Casimir-Polder force*)

Hydrophilic versus hydrophobic interactions, a very important group of van der Waals interactions, were treated in Chapter 6 on fluidics.

In Table 7.1 we compare bonding and nonbonding forces.

Properties controlled by intermolecular forces include vapor pressure, boiling point, molecular crystals, protein folding, colloidal stability, rheological behavior, self-assembly, surface tension, viscosity, and so on. To understand how important intermolecular forces are, the two following quotes help:

> The very existence of condensed phases of matter is conclusive evidence of attractive forces between molecules, for in the absence of attractive forces, the molecules in a glass of water would have no reason to stay confined to the glass…
>
> **Anthony J. Stone**
> *The Theory of Intermolecular Forces* (1996)

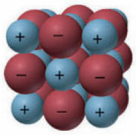
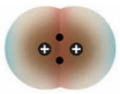
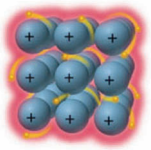
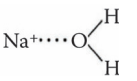
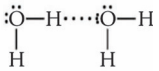
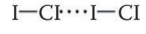
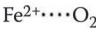
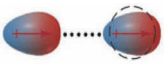
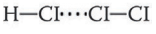
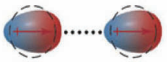
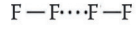
> The most important forces dictating the properties of biomolecules are the non-bonded [intermolecular] interactions…
>
> **Alexander D. MacKerell, Jr.**
> *Journal of Computational Chemistry* (2004)

In other words, in the absence of ionic bonding and interatomic or intermolecular forces, all matter would exist in the gas form, even at 0 K. Thus, intermolecular forces play a crucial role in many of the properties of liquids, gases, and solids. When a substance melts or boils, the intermolecular forces are broken (not the covalent bonds), and when a substance condenses, intermolecular forces are formed.

Intermolecular forces at a solid/liquid interface are the same forces that control the interaction between the constituent molecules of the solution modulated by the presence of foreign molecules at the solid/liquid interface. In this section, we establish the operational range of these forces and find that forces between individual molecules typically only extend over a few nanometers. However, when molecules come together into condensed phases, the force range at the surface is extended to several tens of nanometers.

A couple of examples of intermolecular forces at work follow. When intermolecular forces are weak, little energy is required to overcome them. Consequently, boiling points are low for such compounds. The vapor pressure of a liquid depends on intermolecular forces such that, when the intermolecular forces in a liquid are strong, the vapor pressure is low. The boiling point is related to vapor pressure and is lowest for liquids with the weakest intermolecular forces. Surface tension increases with increasing intermolecular forces, and because surface tension

Table 7.1 Comparison of Bonding (Intramolecular), Ionic, and Nonbonding (Intermolecular) Interactions

Force	Model	Basis of Attraction	Energy (KJ/mol)	Example
Bonding				
Ionic		Cation–anion	400–4000	NaCl
Covalent		Nuclei-shared e^- pair	150–1100	H—H
Metallic		Cations–delocalized electrons	75–1000	Fe
Nonbonding (intermolecular)				
Ion-dipole		Ion charge–dipole charge	40–600	
H-bond		Polar bond to H–dipole charge (high EN of N,O,F)	10–40	
Diploe-bond		Dipole charges	5–25	
Ion-induced dipole		Ion charge-polarizable e^- cloud	3–15	
Dipole-induced dipole		Dipole charge-polarizable e^- cloud	2–10	
Dipersion (London)		Polarizable e^- cloud	0.05–40	

is the energy needed to reduce the surface area of a liquid, it is necessary to pull molecules apart against the intermolecular forces of attraction (cohesion) to increase surface area. Viscosity increases with increasing intermolecular forces because increasing these forces increases the resistance to flow. Other factors, such as the possibility of molecules tangling together, can also affect the viscosity. Liquids with long molecules that tangle together are expected to have higher viscosities. Intermolecular forces explain all the data of the liquids listed in Table 7.2, and we will learn in the pages that follow and elsewhere in the book that they explain many other important phenomena, from DNA hybridization to self-assembly.

Many observed phenomena in nature have their basis in intermolecular forces! For example, in Volume III, Chapter 7 on scaling laws, we will learn that the gecko's ability to walk on ceilings is due to van der Waals attractive forces between the spatulae (1 billion of them) on the gecko's feet and the wall. This example shows how short-range forces can have long-range (macroscopic) effects. Intermolecular forces exert their influence especially in the micro- and nanoworld. In the nanoworld, gravitational forces are negligible, and forces such as van der Waals and the quantum mechanical Casimir force take over. At distances less than a micrometer, the Casimir force (a special type of van der Waals force) becomes the dominant force between two neutral objects in a vacuum. In Volume III, Chapter 8 on actuators, we calculate that two plates with an area of 1 cm^2 separated by a distance of 1 μm have an attractive Casimir force of about 10^{-7} N. At separations of 10 nm, the Casimir effect produces the equivalent of 1 atm. Obviously,

TABLE 7.2 Properties of Some Liquids at 20°C

Substance	Molecular Weight (amu)	Vapor Pressure (mmHg)	Surface Tension (J/m²)	Viscosity (kg/m · s)
Water, H_2O	18	1.8×10^{1}	7.3×10^{-2}	1.0×10^{-3}
Carbon dioxide, CO_2	44	4.3×10^{4}	1.2×10^{-3}	7.1×10^{-5}
Pentane, C_5H_{12}	72	4.4×10^{2}	1.6×10^{-2}	2.4×10^{-4}
Glycerol, $C_3H_8O_3$	92	1.6×10^{-4}	6.3×10^{-2}	1.5×10^{0}
Chloroform, $CHCl_3$	119	1.7×10^{2}	2.7×10^{-2}	5.8×10^{-4}
Carbon tetrachloride, CCl_4	154	8.7×10^{1}	2.7×10^{-2}	9.7×10^{-4}
Bromoform, $CHBr_3$	253	3.9×10^{0}	4.2×10^{-2}	2.0×10^{-3}

Intermolecular forces explain the data of the liquids in this table.

this is a force to reckon with in MEMS and NEMS devices.

Interparticle Interaction Energy (U) and Interaction Forces (F)

The interaction energy between two particles (atoms or molecules) is defined as the pair potential (energy), $U(r)$, which depends on the intermolecular separation, r. The force, $F(r)$, between the two particles is obtained as the negative of the gradient of the pair potential:

$$F(r) = -\frac{dU(r)}{dr} \tag{7.1}$$

At very large separations (as $r \rightarrow \infty$), the interparticle interaction is negligible ($U \rightarrow 0$, $F \rightarrow 0$). At shorter separations, the particles attract ($F < 0$, particles are pulled together). A dimer is more stable than the two isolated particles, hence $U(r) < 0$. At very short separations ($r \rightarrow 0$), the particles repel each other ($F > 0$, particles are pushed apart). At $r = r_e$, $U(r)$ is at its minimum (U_{min}), and the interparticle force, $F = 0$. This is summarized in Figure 7.2.

The diagram in Figure 7.2 shows the total potential energy between two particles. A power law with r raised to the exponent $-n$ generally models the intermolecular potential energy function:

$$U(r) = Cr^{-n} = \frac{C}{r^{n}} \tag{7.2}$$

where C is a constant. The general expression for $F(r)$ is then:

$$F(r) = -\frac{nC}{r^{n+1}} \tag{7.3}$$

For the case of attractive interactions, $U(r) < 0$, whereas for repulsive interactions, $U(r) > 0$. To understand the exact shape of the curve in Figure 7.2, we must introduce the exact equations for repulsive and attractive potentials first. The most important repulsive term corresponds to the limit of compressibility of matter. At short internuclear separations, there are both electrostatic repulsions between the atomic nuclei and between the electrons (both valence and core) on neighboring atoms. The latter is based on the Pauli exclusion principle. The Pauli exclusion principle is quantum mechanical in origin and only allows for two electrons with opposite spin per energy level. The importance of the Pauli principle was exemplified in Chapter 3, where we saw that without the Pauli principle all material would collapse! The total short-range repulsive interaction is generally modeled by steep $1/r^{12}$ dependence on interatomic distance (Figure 7.2, left

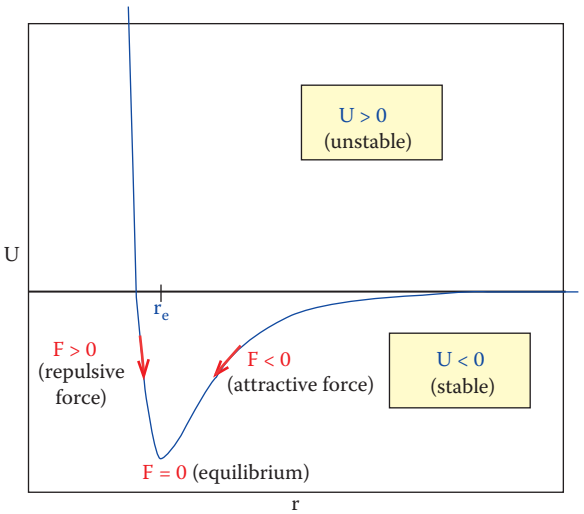


FIGURE 7.2 Interparticle interaction energy (U) and forces (F) as a function of interparticle distance (r).

side of the curve). Forces that scale with an exponent of n ≥ 3 involve short-ranged forces, whereas values of $n < 3$ mark long-range forces. To appreciate where this distinction comes from, we need to briefly introduce the scaling of forces between sets of molecules.

Consider a region of space (solid, liquid, or gas) with a molecular density ρ. The summation of all interaction energies, U_{Total}, of a molecule with all other molecules in a spherical region of size L in that space is then given as:

$$U_{Total} = \int_{\sigma}^{L} U(r)\rho 4\pi r^2 dr = \int_{\sigma}^{L} \frac{\rho C 4\pi r^2}{r^n} dr \qquad (7.4)$$

or:

$$U_{Total} = \frac{-4\pi C\rho}{(n-3)\sigma^{n-3}}\left[1-\left(\frac{\sigma}{L}\right)^{n-3}\right]$$

where σ is the molecular diameter. This important equation reveals that molecular forces contribute to interaction energies for n ≥ 3. At distances $L >> \sigma$, intermolecular forces become inefficiently weak and do not contribute to U_{Total}; however, with $n < 3$, the size of the system L becomes important. The latter is the case, for example, for gravity (distant planets and stars) and Coulomb interactions where n = 1, i.e., both are very long-ranged forces that fall off very slowly. For n ≥ 3, the intermolecular potential becomes important and is the reason why bulk properties of a material are size-independent (water boils at the same temperature in a cup as it does in a bucket) unless we go to submicrometer size distances, where properties of material change as intermolecular forces start taking over. The Pauli-based repulsive interaction, for example, is very short ranged (n = 12).

Types of Interparticle Interaction Forces

Coulomb Interactions

The Coulomb electrostatic interaction potential between two charges (Figure 7.3) is represented by the following equation:

$$U(r) = \frac{Q_1 Q_2}{4\pi\varepsilon_0 \varepsilon r} \qquad (7.5)$$

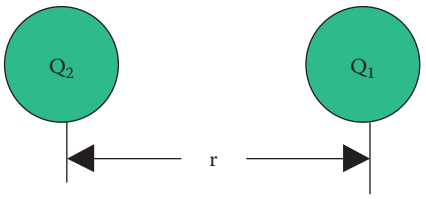


FIGURE 7.3 Coulomb interaction between two point charges or monopoles.

(Equation 7.2 with n = 1 and $C = \frac{Q_1 Q_2}{4\pi\varepsilon_0\varepsilon}$). In Equation 7.5, $Q_1 = z_1 e$, where e is the magnitude of the electron charge (1.602×10^{-19} C) and z_1 is an integer value. In the denominator, ε_0 is the permittivity of free space (8.854×10^{-12} J^{-1} C^2 m^{-1}), and ε is the relative permittivity of the medium between the ions (ε = 1 in a vacuum or ε > 1 in a gas or liquid). In the case of repelling like charges $U(r)$ is positive, and in the case of attracting opposite charges $U(r)$ is negative. This Coulomb interaction potential is additive when charges are grouped in crystals. With ions in solution, the Coulomb interaction is modulated by the dielectric constant of the medium. Highly polar solvents have high dielectric constants. Water, for example, has a dielectric constant ε = 78 (at interfaces this number is much lower as movement of the water dipoles is more restricted). This leads to a large reduction in the Coulombic potential between ions in polar solvents and can lead to the solvation or hydration of a salt in a solution (Figure 7.4). The hydration of cations and anions is responsible for the high solubility of many ionic compounds in water.

When molecules are in close contact, $U(r)$ is typically ~10^{-18} J, corresponding to about 200–300 kT (where k is the Boltzmann constant and T the temperature in Kelvin) at room temperature. To contrast

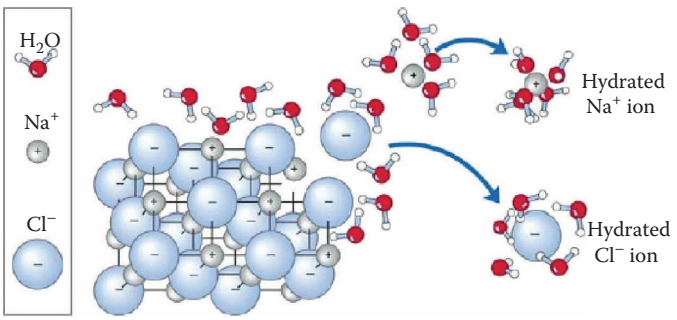


FIGURE 7.4 Solvation power of water. NaCl dissolution.

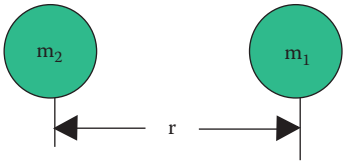


FIGURE 7.5 Gravitational interaction between two molecules.

this Coulomb interaction with the gravity interaction of molecules (Figure 7.5), consider that the gravitational interaction energy is given by:

$$U(r) = \frac{gm_1m_2}{r} \tag{7.6}$$

This is equivalent to Equation 7.2 with n = 1 and $C = gm_1m_2$. In this expression the gravitational constant $g = 6.67 \times 10^{-11}$ Nm^2kg^{-1}, and for molecules in contact this leads to a negligible gravitational interaction energy $U(r)$ of about 10^{-52} J compared with 10^{-18} J for the Coulomb interaction energy.

The Coulomb interaction is an electrostatic interaction, i.e., such an interaction may occur between charged atomic or molecular species (ions = monopoles) or between asymmetric charge distributions in neutral molecules (e.g., dipoles, quadrupoles, etc.). A dipole is an asymmetric charge distribution in a molecule, where there is no net charge; however, one end of the molecule is negative (partial charge = −q) relative to the other (partial charge = +q). Molecules may possess higher-order electric multipoles, arising from their nonspherical charge distributions (Figure 7.6). Each type of multipole has an associated multipole moment: the monopole moment is the charge of the atom/molecule; the dipole moment is a vector whose magnitude is the product of the charge (q) and distance (λ) between the charge centers ($\mu = q\lambda$). Higher-order multipole moments have tensor properties. A polar molecule is one that possesses a permanent dipole moment: there is an asymmetric charge distribution, with one end of the molecule relatively negative ($-q$) with respect to the other ($+q$).

The interaction energy between a stationary point charge Q_2 and a permanent fixed dipole μ_1 separated by a distance r is given as (Figure 7.7):

$$U(r) = -\frac{Q_2\mu_1\cos\theta}{4\pi\varepsilon_0\varepsilon r^2} \tag{7.7}$$

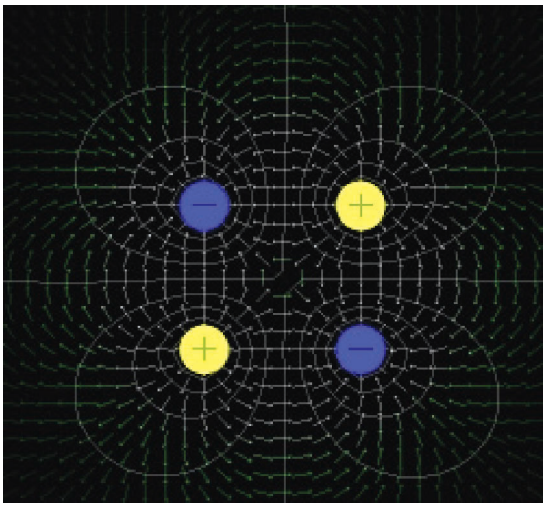


FIGURE 7.6 Electric quadrupole. Four charges, two positive and two negative. The arrows depict the electric fields, and the lines are the equipotential surfaces.

with $Q_2 = z_2e$ in units of C and μ_1 ($q_1\lambda$) in units of Cm [often expressed in Debye (D), where D = 3.336 10^{-30} Cm]. The interaction energy now depends on the orientation of the dipole relative to the charge (θ) and the sign of the charge. In the case of a permanent dipole that can freely rotate, one obtains a new expression:

$$U(r) = -\frac{Q_2^2\mu_1^2}{6(4\pi\varepsilon_0\varepsilon)^2 kTr^4} \tag{7.8}$$

with k being the Boltzmann constant. Notice that temperature dependence now appears in the denominator. One example of this type of interaction is the condensation of water (μ = 1.65 D) caused by the presence of ions in water. During a thunderstorm, ions are created that nucleate raindrops in thunderclouds (ionic nucleation).

When an electric charge (i.e., an ion = electric monopole) or asymmetric charge distribution (e.g.,

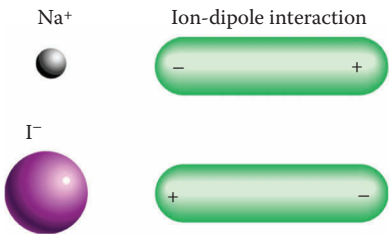


FIGURE 7.7 Ion-dipole interaction.

dipole, quadrupole, etc.) gives rise to an electric field (E) that causes the polarization of neighboring nonpolar atoms or molecules, an induced dipole is created. The magnitude of the induced dipole is:

$$\mu = \alpha E \tag{7.9}$$

where α is the polarizability of the neutral atom or molecule. Polarizability is the ease with which the electron distribution in the atom or molecule can be distorted. Polarizability increases as the number of electrons increases and as the electron cloud becomes more diffuse (larger molecules). An ion-induced dipole interaction energy is given as:

$$U(r) = -\frac{Q_1^2 \alpha_2}{2(4\pi\varepsilon_0\varepsilon)^2 r^4} \tag{7.10}$$

Van der Waals Forces

To explain deviations from ideal gas behavior (equation of state by Boyle and Gay-Lussac: $PV = RT$), van der Waals, in 1873, proposed a modification to the equation of state of a gas as follows:

$$\left(P + \frac{a}{V^2}\right)(V - b) = RT \tag{7.11}$$

This new expression accounted for the fact that molecules have a finite volume (b) and that there is an attractive force between these molecules (a). This relation correctly predicts gas behavior over a wider pressure range.

The attractive intermolecular forces between gas molecules are now known as *van der Waals forces*. The three types of interactions that constitute van der Waals forces are summarized in Table 7.3. The exponent $n = 6$ indicates that these types of van der Waals forces are short range. The attractive interactions between dipoles, quadrupoles, and between permanent multipoles are sometimes named after Keesom. Another source of attraction is induction (i.e., polarization), which is the interaction between a permanent dipole or multipole on one molecule with an induced dipole or multipole. This interaction is named after Debye. The third type of attraction, London forces, is named after the German-American physicist Fritz London, who himself called it *dispersion forces*. This is the only attraction experienced by noble gas atoms, but it is operative between any pair of molecules, regardless of their symmetry. London dispersion forces are weak intermolecular forces that arise from the attractive forces between transient dipoles (or multipoles) in molecules without permanent dipole (or multipole) moments. We will see that the London component is often the most dominant interaction from the three interactive forces.

Dipole-Dipole Interactions—Keesom The expression for the interaction energy of two fixed dipoles, μ_1 and μ_2, that are not coplanar (i.e., there is an angle ϕ between them) is calculated as:

$$U(r) = -\frac{\mu_1^2 \mu_2^2 f(\theta_1, \theta_2, \phi)}{4\pi\varepsilon_0\varepsilon r^3} \tag{7.12}$$

where r is the distance between the centers of mass of the two molecules. The expression for two freely rotating dipoles, the so-called *Keesom equation*, is given as:

$$U(r) = -\frac{\mu_1^2 \mu_2^2}{3(4\pi\varepsilon_0\varepsilon)^2 kTr^6} \tag{7.13}$$

Table 7.3 Types of van der Waals Forces

Interaction Component	Origin of Interactions	Equation
Keesom	Dipole-dipole	$U(r) = -\frac{\mu_1^2\mu_2^2}{3(4\pi\varepsilon_0\varepsilon)^2 kTr^6}$
Debye (induction)	Dipole-induced dipole	$U(r) = -\frac{\mu_1^2\alpha_2}{(4\pi\varepsilon_0\varepsilon)^2 r^6}$
London (dispersion)	Induced dipole-induced dipole	$U(r) = -\frac{3\alpha^2 h\nu}{4(4\pi\varepsilon_0\varepsilon)^2 r^6}$

Here a temperature dependence of the interaction energy appears. Dipole-dipole interactions are generally weak. These dipole-dipole interactions become significant when two interacting dipoles approach each other closely, e.g., in $O^{\delta-}$—$H^{\delta+}$, $N^{\delta-}$—$H^{\delta+}$, $F^{\delta-}$—$H^{\delta+}$, etc. Such strong dipole-dipole interactions are the underlying reason for the well-known hydrogen bonds (see below). The electrostatic interaction between higher-order multipoles can be treated in an analogous fashion to that described here for dipole-dipole interactions. The distance dependence of the interaction energy between an *n*-pole and an *m*-pole (assuming no molecular rotation) is given as:

$$U(r) \propto \pm \frac{1}{r^{(n+m-1)}} \tag{7.14}$$

where *n* and *m* are the orders of the multipoles: n, m = 1 (monopole), 2 (dipole), 3 (quadrupole), 4 (octopole), etc. (Table 7.4). It is noted that the higher the order of the multipole, the shorter ranged the interaction is, i.e., the more rapidly it falls off with distance.

Dipole-Induced Dipole Interactions—Debye An induced dipole interacts in an attractive fashion with a fixed, permanent dipole (μ_1). The inductive interaction energy between the fixed permanent dipole (μ_1) associated with molecule 1 and the induced dipole on molecule 2 (which has no fixed dipole but has certain polarizability α_2 and is a distance *r* away from molecule 1) is given by:

$$U(r) = -\frac{\mu_1^2 \alpha_2 \left(1 + 3\cos^2\theta\right)}{2(4\pi\varepsilon_0\varepsilon)^2 r^6} \tag{7.15}$$

Table 7.4 Interaction Energies between an n-Pole and an m-Pole

Interaction Energies for…	
Monopole-monopole (coulomb)	$U(r) \propto r^{-1}$
Monopole-dipole	$U(r) \propto r^{-2}$
Dipole-dipole	$U(r) \propto r^{-3}$
Dipole-quadrupole	$U(r) \propto r^{-4}$
Quadrupole-quadrupole	$U(r) \propto r^{-5}$

In the case where the permanent dipole is free to rotate, the so-called *Debye interaction* energy results:

$$U(r) = -\frac{\mu_1^2 \alpha_2}{(4\pi\varepsilon_0\varepsilon)^2 r^6} \tag{7.16}$$

As the direction of the induced dipole follows that of the permanent dipole, the effect survives even if the permanent dipole is subjected to thermal reorientations. Debye interactions are thus independent of temperature. Dipoles (and higher multipoles) may also be induced by neighboring molecules featuring permanent charges (monopoles) or by high-order multipoles. Similar expressions can be derived for the interactions between these permanent and induced multipoles.

Induced Dipole-Induced Dipole—Dispersive or London Forces The intermolecular forces between nonpolar molecules and closed shell atoms (e.g., rare gas atoms such as He, Ne, and Ar) are the "London" or dispersion forces, attractive forces that arise as a result of temporary dipoles, which are induced in atoms or molecules (Figure 7.8). Keesom and Debye interactions require the presence of at least one permanent dipole, but London interactions do not. Hence, London interactions exist between all molecules. It is also generally the dominant contribution to van der Waals forces. The London interaction component was known for a long time, but the theory only evolved after the development of quantum mechanics. Nonpolar molecules exhibit London forces because electron density moves about a molecule probabilistically. There is a high chance that the electron density will not be evenly distributed throughout a nonpolar molecule. When an uneven distribution occurs, a temporary multipole is created, which may interact with other nearby multipoles. London forces are also present in polar molecules, but they are usually only a small part of the total interaction force. This phenomenon is the only attractive intermolecular force between neutral atoms (e.g., helium) and is the major attractive force between nonpolar molecules (e.g., nitrogen or methane). Without London forces, there would be no attractive force between noble gas atoms, and they could not then be obtained in a liquid form.

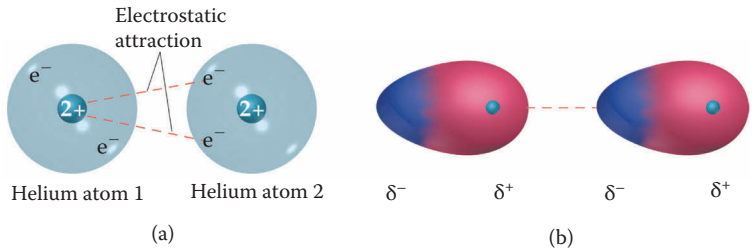


FIGURE 7.8 Induced dipole-induced dipole interaction. Dispersion or "London" forces in helium atoms (a) before the dipole and (b) with the dipole.

A general expansion for the attractive energy resulting from these dispersion forces can be written as:

$$U(r) = -\left\{\frac{C_6}{r^6} + \frac{C_8}{r^8} + \frac{C_{10}}{r^{10}} + \cdots + \frac{C_n}{r^n}\right\} \qquad (7.17)$$

where the C_n terms are constants and r is the interatomic distance. The first term represents the instantaneous dipole-dipole interaction and is the dominant variable; therefore, the higher terms (which contribute less than 10% of the total dispersion energy) are often omitted when calculating dispersion energies. A reasonable approximation to the dispersion energy is found in the London formula (derived by London in 1930, based on a model by Drude), in which the dispersion energy between two identical atoms/molecules is given as:

$$U(r) = -\frac{C_6}{r^6} = -\frac{3\alpha^2 h\nu}{4(4\pi\varepsilon_0\varepsilon)^2 r^6} \qquad (7.18)$$

where α is the polarizability and $h\nu$ is the ionization energy of the atom/molecule. The polarizability α is squared in Equation 7.18 because the magnitude of the instantaneous dipole depends on the polarizability of the first atom/molecule. The strength of the induced dipole (in the second atom/molecule) also depends on α. When molecules are in close contact, $U(r)$ is typically ~10^{-21}–10^{-20} J/mol, corresponding to about 0.2–2 kT at room temperature, which is a comparable magnitude to the thermal energy. London forces become stronger as the atom (or molecule) in question becomes larger. This is because of the increased polarizability of molecules with larger, more dispersed electron clouds. London forces provide molecular level reasoning that at room temperature small molecules such as Ar, He, and CH_4 are gases, whereas larger molecules, like hexane and decane, are liquids, and yet even larger molecules are solids. Interactions involving charged species (e.g., ions) tend to be stronger than polar-polar interactions.

All intermolecular forces are anisotropic (except those between two noble gas atoms), which means that they depend on the relative orientation of the molecules. The induction and dispersion interactions are always attractive, but the electrostatic interaction changes sign on rotation of the molecules (can be attractive or repulsive, depending on the mutual orientation of the multipoles). When molecules are in thermal motion, as they are in the gas and liquid phase, the electrostatic forces are diminished to a large extent because the molecules thermally rotate and thus probe both repulsive and attractive parts of the potential. The thermal averaging effect is much less pronounced for induction and dispersion forces.

Further below we will see that the London-van der Waals forces are related to the quantum mechanical Casimir effect introduced in Chapter 3 (Figure 3.69). The latter is sometimes referred to as the *retarded* van der Waals force.

Example of Dipole-Dipole Interaction: Hydrogen Bonding Earlier we saw that weak dipole-dipole interactions can become significant when two interacting dipoles approach each other closely, e.g., in $O^{\delta-}$—$H^{\delta+}$, $N^{\delta-}$—$H^{\delta+}$, $F^{\delta-}$—$H^{\delta+}$, and so on. Such strong dipole-dipole interactions are the underlying reason for hydrogen bonds. Hydrogen bonding is an intermolecular linkage of the form X—H—Y, where a hydrogen atom is bound to one electronegative atom (e.g., X=N, O, F) and interacts with a second electronegative atom (Y) that has an available pair

of free electrons. For example, X—H is the hydrogen bond donor (e.g., O—H, peptide N—H), and Y is the hydrogen bond acceptor (e.g., :O, :N, peptide C=O). The donor is, in this case, an acid (proton donor), whereas the acceptor is a base (proton acceptor). A hydrogen bond is stronger than most other nonbonded interactions. Typical hydrogen bond strength is 10–40 kJ mol^{-1} (about 10% of a covalent bond). A rare strong H-bond may be 60 kJ mol^{-1} when the donor and acceptor are very close for significant orbital overlap.

Hydrogen bonding is especially strong when H is attached to F, O, or N because these three elements all have high electronegativity and are small enough to get close to. Most biological H-bonds are ~2.8 Å (e.g., O—H…O is 2.7 Å and O—H…N is 2.9 Å). Water forms hydrogen bonds: the O—H bonds act as H-bond donors, and the O lone pairs act as H-bond acceptors (Figure 7.9). The proton in the water molecule is unshielded and makes an electropositive end to the bond, which explains the partial ionic character of the hydrogen bond. Most of the unique properties of water are the result of this hydrogen bonding and its polar nature. Water, we saw above, is a good solvent for ionic species because it has a high dielectric constant. Each molecule of water in the solid state (ice, see Figure 7.10) is engaged in 4.0 hydrogen bonds (accepting two and donating two), and each molecule of water in the liquid state is engaged in approximately 3.4

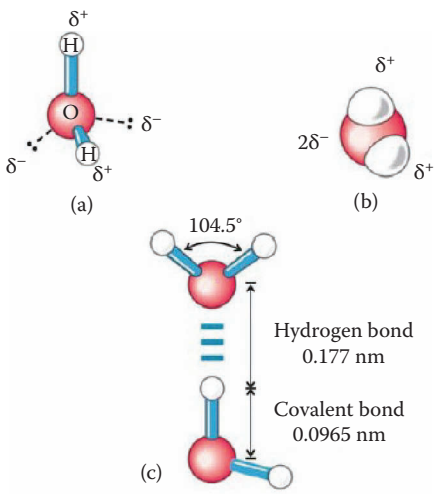


FIGURE 7.9 Hydrogen bonding in water. The structure of a water molecule (a) makes it a perfect hydrogen bond donor (b) and acceptor (c).

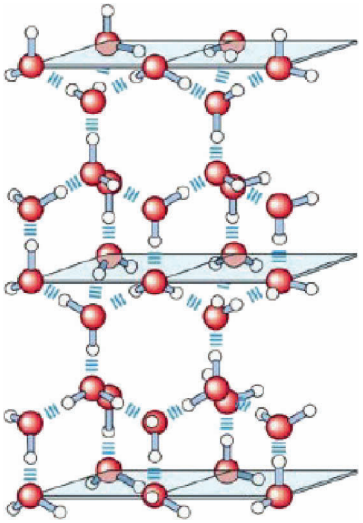


FIGURE 7.10 In ice, water molecules form an H-bonded network that fully satisfies the H-bonding potential (4 per H_2O).

hydrogen bonds that are transitory in nature. With fewer hydrogen bonds, the oxygen atoms may on average be closer to each other (denser). About 15% of H-bonds break when ice melts. As a consequence, water is denser than ice.

The hydrogen bond endows specificity to an interaction as particular atoms are required as donors and acceptors. It also imposes geometrical constraints; hydrogen bonds, like covalent bonds, are highly directional. The X—H—Y bond is optimally a straight line (see Figure 7.9c), where the donor H bond tends to point directly at the acceptor electron pair.

The major component (around 80%) of a hydrogen bond is electrostatic, whereas other contributions are inductive and dispersion forces. It is the latter forces that impose the directionality in the direction of the acceptor lone pair. Hydrogen bonding plays a very important role in biology, being responsible for the structural organization of proteins and other biomolecules. Actually, the existence of life on Earth is a direct consequence of the hydrogen bonding present in water and the wide temperature range over which water remains liquid (0–100°C). Also, the tetrahedral three-dimensional (3D) structure of ice and the local structure of liquid water is a consequence of hydrogen bonding. We will come back to hydrogen bonding in discussions on protein folding, DNA, and self-assembly

(Volume III, Chapters 2–4). In the gas phase, the optimal H-bond strength between water molecules is ~23 kJ mol^{-1}. The H-bonding in condensed phases of water is cooperative (nonadditive), and the strength of H-bonding increases with increasing number of water molecules, as this increases the polarization of the O—H bonds. This is observed as an increase in the average dipole moment (μ) per water molecule, which increases from 1.85 D (isolated H_2O) to 2.4–2.6 D (liquid H_2O at 0°C).

In Table 7.5, we summarize what we have learned about intermolecular forces thus far. We describe the type of interaction, give a model and an example, and list the distance dependence of the energy.

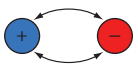
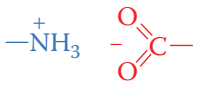
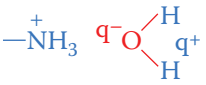
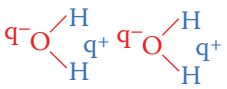
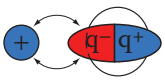
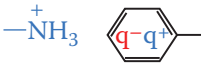
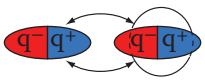
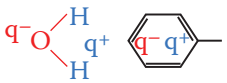
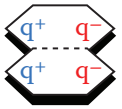
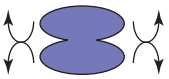
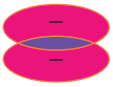
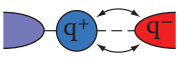
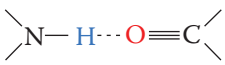
The Lennard-Jones Potential

Having introduced the equations for repulsive and attractive intermolecular forces we are now ready to combine them to describe the exact shape of the curve in Figure 7.2 for the total potential energy of interaction between two neutral atoms or polar molecules as a function of the distance between them. In combining forces, one typically groups the attractive forces between polar and nonpolar molecules in the so-called *van der Waals force*. This refers to all electrostatic induction and dispersion interactions between polar and nonpolar molecules, varying as r^{-6}. From Table 7.3, van der Waals forces include Keesom (dipole-dipole), Debye (dipole-induced dipole), and London dispersive interactions (induced dipole-induced dipole). These short-range forces (n = 6) can all be expressed as:

$$U(r) = -\frac{C}{r^6} \tag{7.19}$$

The most important component of the van der Waals force, as seen earlier, is the London term.

Table 7.5 Intermolecular Energy Dependence on Distance for Various Types of Interactions

Type of Interaction	Model	Example	Dependence of Energy on Distance
(a) Charge-charge Longest-range force; nondirectional			1/r
(b) Charge-dipole (fixed) Depends on orientation of dipole			$1/r^2$
(c) Dipole-dipole Depends on mutual orientation of dipole			$1/r^3$
(d) Charge-induced dipole Depends on polarizability of molecule in which dipole is induced			$1/r^4$
(e) Dipole-induced dipole Depends on polarizability of molecule in which dipole is induced			$1/r^6$
(f) Dispersion-London forces Involves mutual synchronization of fluctuating charges			$1/r^6$
(g) Fermi repulsion Occurs when outer electron orbitals overlap			$1/r^{12}$
(h) Hydrogen bond Charge attraction + partial covalent bond			Length of bond fixed

Regardless of other interactions found within a complex system of atoms and molecules, there will almost always be a contribution from van der Waals forces. Taking into account both repulsive and attractive interactions, a good approximation to the total potential energy of interaction between two neutral atoms or polar molecules is given by the Lennard-Jones potential:

$$U(r) = +\frac{B}{r^{12}} - \frac{C}{r^{6}} \qquad (7.20)$$

The C/r^6 term is the attractive van der Waals term, also known as the *dispersive term* (Equation 7.19). The term B/r^{12} describes the hard-core repulsion, also known as *Fermi repulsion*, and stems from the Pauli exclusion principle. Atoms are treated here as hard spheres with diameter σ [in the extreme; $U(r) = \left(\frac{\sigma}{r}\right)^{\infty}$]. This leads to a potential expression of the following form:

$$U(r) = 4\varepsilon\left\{\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right\} \qquad (7.21)$$

as illustrated in Figure 7.11. The force between the two molecules is derived from the Lennard-Jones potential as:

$$F(r) = -\frac{dU}{dr} = 24\varepsilon\left\{2\left(\frac{\sigma^{12}}{r^{13}}\right) - \left(\frac{\sigma^{6}}{r^{7}}\right)\right\} \qquad (7.22)$$

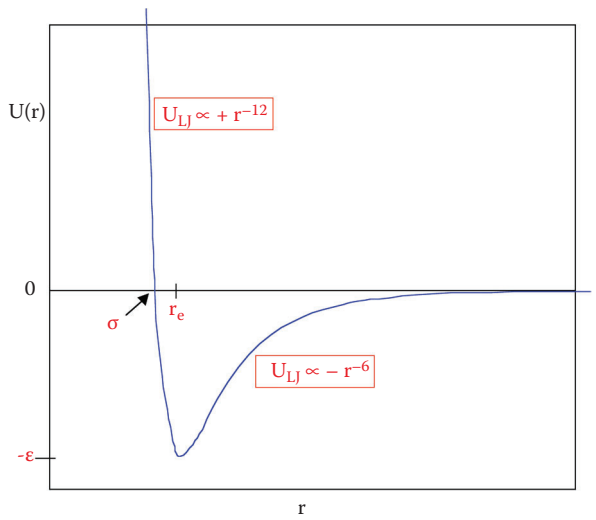


FIGURE 7.11 The Lennard-Jones potential.

In Equation 7.21, ε is the depth of the energy well and σ is the separation at which U(r) = 0. The equilibrium separation, i.e., the position of the m inimum in $U(r)$, occurs at $r_e = 2^{1/6} \times \sigma$ [$U(r) = -\varepsilon$]. The *van der Waals' radii* (σ) of atoms, which is the size of an atom or the preferred distance atoms will pack to if no other significant interactions occur, may be calculated from this equation. The increase in ε when going from lighter to heavier rare gases reflects the increase in the attractive dispersion forces because of the higher polarizability that occurs as a result of the lower effective nuclear charge experienced by the outer electrons in the heavier rare gas atoms. As molecules are larger and (usually) more polarizable than rare gas atoms, typical binding energies caused by dispersion forces are of the order ε ≈ 10 kJ mol^{-1} per intermolecular interaction. The interactive attraction forces are the derivative of the van der Waals term, and these contributions (polar or nonpolar) all decrease as $1/r^7$. In all cases, substituting numerical values in Equation 7.22, these forces do not exceed a range of a few nanometers.

Example 7.1:

The magnitude of the dispersion energy for two methane molecules separated by 0.3 nm is approximately −4.7 kJ mol^{-1}, i.e., ε ≈ 4.7 kJ mol^{-1}.

Calculating Interaction Forces between Surfaces and Particles

Introduction

Thus far, we have mainly discussed intermolecular forces at work between various types of molecules. Now we turn our attention to the detailed nature of the attractive and repulsive forces at an interface between a liquid and a solid. Based on this knowledge, we will be ready to introduce the electrode/liquid interface and the electrochemical principles that enable electrochemical sensors.

Van der Waals Forces on Solid Surfaces

Van der Waals forces for molecules, we saw, vary as $1/r^7$ (Equation 7.22), with r being the distance between them. In 1937, Hamaker calculated the

Table 7.6 Hamaker Constants H for Various Materials in 10^{-20} J

Material	Vacuum	Medium Water	n-Dodecane
6H-SiC	25	13	1
tetra-ZrO_2	20	8.8	6.8
β-Si_3N_4	18	7.0	5.0
α-Al_2O_3	15	5.2	3.6
Y_2O_3	13	4.0	2.6
MgO	12	3.2	1.9
$MgAl_2O_4$	13	3.5	2.1
SiO_2	6.5	0.83	0.15

Source: Bergström, L. 1997. *Adv Colloid Interface Sci*, 70:125.[64]

total van der Waals interaction force between two spherical particles (an agglomeration of molecules/atoms) by adding contributions for each atom of the interacting particles and found*:

$$F(r) = -\frac{H\sigma}{12L_0^2} \tag{7.23}$$

where σ = particle radius

L_0 = equilibrium separation distance of r, the distance where repulsive quantum forces equal attractive van der Waals forces (for liquids, 0.16 nm; for solids, 0.3 nm)

H = Hamaker constant

A typical value of the Hamaker constant *H* for SiO_2 in water is 0.83×10^{-20} J. For the Hamaker constant of some selected materials in different media, see Table 7.6.

The Hamaker constant *H* is defined as:

$$H = \pi^2 C\rho_1\rho_2 \tag{7.24}$$

where ρ_1 and ρ_2 are the atom densities in the two spherical particles and *C* is the coefficient in the interaction potential energy equation.

Based on Equation 7.23, the energy of adhesion between two spheres is:

$$U(r) = -\int F(r)dr = \int -\frac{H\sigma}{12L_0^2}dr = -\frac{H\sigma}{12L_0} \tag{7.25}$$

where $r = L_0$ is the equilibrium distance.

* This is a simplified equation in which we assume that the distance between particles is significantly smaller than the particle radius ($\sigma << L$).

Example 7.2:

Considering only van der Waals attraction, what is the force and energy of adhesion between two particles of radius 1 μm with $H = 10^{-19}$ J? ($L_0 = 3$ Å).

The force of adhesion is calculated as:

$$F(r) = -\frac{H\sigma}{12L_0^2} = -\left(\frac{10^{-19}\,J \times 10^{-6}\,m}{12 \times (3 \times 10^{-10})^2 m^2}\right)$$

$$= -9.26 \times 10^{-8} N$$

and the energy of adhesion is given by:

$$U(r) = -\int F(r)dr = \int -\frac{H\sigma}{12L_0^2}dr = -\frac{H\sigma}{12L_0}$$

$$= \left(\frac{10^{-19}\,J \times 10^{-6}\,m}{12 \times (3 \times 10^{-10})m}\right) = -28 \times 10^{-18} J = -6800\,kT$$

The theoretical energy of adhesion between two flat plates (van der Waals approach) is given as:

$$U(r) = -\frac{H}{12\pi L_0^2} = -2\gamma \quad \text{or}$$

$$\frac{H}{24\pi L_0^2} = \gamma \tag{7.26}$$

where γ is the surface energy (for solids) or the surface tension for liquids (when considering the interface between two liquids), and L_0 is the equilibrium distance between the two planes [the surface tension $\gamma = (\partial G/\partial A)$ is the change of Gibbs free energy per area change, i.e., a measure of the energy required to produce a unit area of new surface]. Energy of 2γ is necessary to separate two planes bound by van der Waals forces. The strength of the van der Waals interaction is essentially a function of the surface area of contact. The larger the surface area, the stronger the interaction will be. For solids, the surface energy depends on the exact crystal structure of the surface, and as a consequence the surface tension might be anisotropic. Surface energy of solids is not only a material property, it also is significantly influenced by the environment and may change dramatically based on the history

Table 7.7 Surface Energies of Different Materials

Material	Surface Energy (mJ/m²)
Teflon	20
Paraffin wax	26
Polypropylene	28
Polyethylene	36
Lead iodide	130
Silica	462
Iron	1360
Mica	4500

of the surface. Factors influencing surface energy include organic contamination, roughness/edges/high-energy sites, wetting agents, humidity, and the crystallography. The contact angle of water on silica changes from approximately 10° to 45° with heat treatment, and the contact angle of water on TiO_2-silicone thin film changes from 12° to ~3° on irradiation with UV. The surface energies of some different materials are listed in Table 7.7. The surface of an object determines its optical appearance, stickiness, wetting behavior, frictional behavior, and chemical and electrochemical reactivity.

From Equation 7.26, the van der Waals force between two planes is:

$$F(r) = \frac{H}{6\pi L_0^3} \tag{7.27}$$

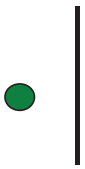
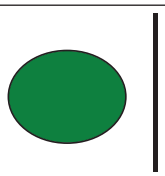
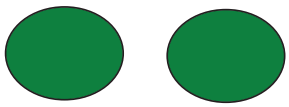
The van der Waals interactions potentials between surfaces, molecules, and particles are summarized in Table 7.8.

Perhaps it is somewhat troubling for the observant reader that the potentials for sphere and plate interactions all have n values less than 3(n = 1); did not we say earlier that intermolecular forces exist only for n ≥ 3? To explain this, we consider the sphere-plane interaction sketched in Figure 7.12. Integrating the interaction of every element on the sphere with the plate, we obtain:

$$U(r) = -\int_0^\infty \frac{\pi\rho C}{6(L+z)^3} 2\pi\sigma\rho z dz \quad \text{or}$$
$$U(r) = -\frac{\pi^2\rho^2 C\sigma}{6} = -\frac{H\sigma}{6L_0} \tag{7.28}$$

We recognize that the bulk material contribution to the interaction terms causes n to collapse to 1,

Table 7.8 van der Waals Interaction Potentials for Various Configurations

Configuration	Schematic	Expression of the Interaction Energy U(r)
Molecule-molecule		$U(r) = -\frac{C}{r^6}$
Molecule-plane		$U(r) = -\frac{\pi C\rho}{6L_0^3}$
Sphere-plane		$U(r) = -\frac{H\sigma}{6L_0}$
Sphere-plane		$U(r) = -\frac{H\sigma}{12L_0}$
Plane-plane		$U(r) = -\frac{H}{12\pi L_0^2} = -2\gamma$

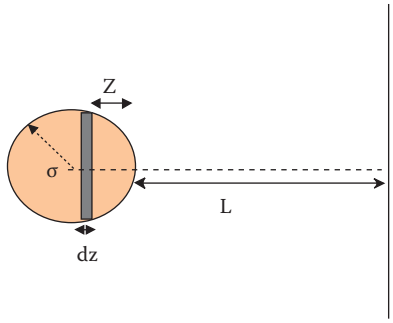


FIGURE 7.12 The van der Waals interaction energy, $U(r)$ for a sphere/plane system.

as a result of the integration of a large number of individual sphere elements with the plate. This also explains the longer range of van der Waals forces, tens of nanometers instead of a few nanometers, when considering solid planes and spheres instead of molecules.

In large objects with small surface area A to volume V ratio (A/V), the physical and chemical properties are primarily defined by the bulk (inside), whereas in small objects with a large A/V ratio, the properties are strongly influenced by the surface.

Two important references to consult on intermolecular forces are Rigby et al.[1] and Stone.[2]

Electrostatic Forces at the Solid/Liquid Interface

Introduction

Most chemical sensors eventually involve a solid in direct contact with an analyte in its environment. Therefore, the better one can control the sensor surface, the better one can control that interface and thus say something more definite about the analyte. Here we will further elaborate on solid/liquid interfaces before we discuss any types of electrochemical sensors in detail. Liquid/solid interfaces are very complex, often involving fractals* (e.g., beach, trees, snowflakes; see Figure 7.13) rather than atomically sharp transitions. This implies, as we will see, that perfect sensor selectivity will be hard to achieve (too many different types of binding sites with different binding constants/selectivities).

* A fractal, we will see in Volume III, Chapters 7 and 9, maximizes surface area while minimizing the amount of work involved in reaching all points on that surface.

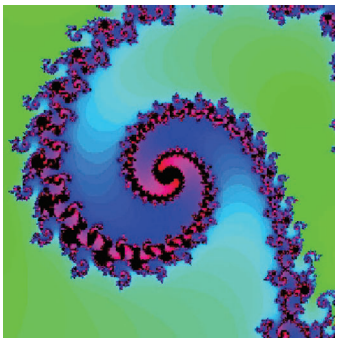


FIGURE 7.13 Solid/liquid interface as a fractal.

There is a tendency for charged species to be attracted to or repelled from the solid/solution interface. This gives rise to a separation of charge, and the layer of solution with a different composition from the bulk solution is known as the *electrical double layer* (EDL), which is sketched for a metal in an aqueous solution in Figure 7.14.

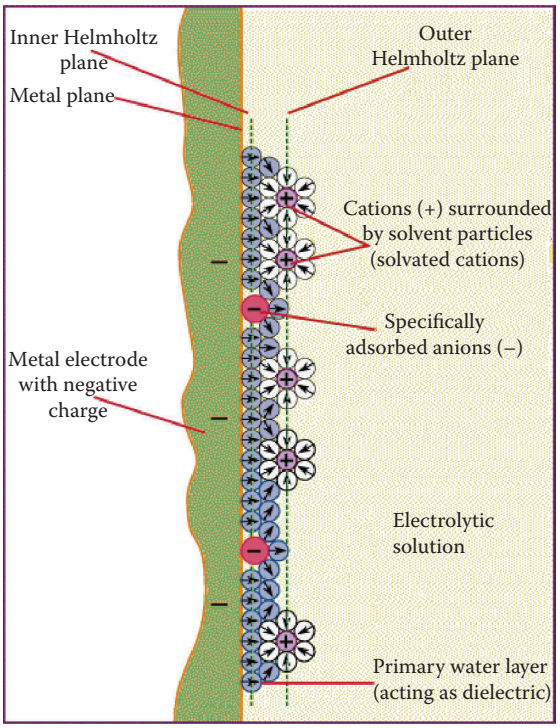


FIGURE 7.14 The electrical double layer (EDL) for a metal in an aqueous solution. The metal/electrolyte interface, showing fully and partially solvated ions. Completely solvated ions are located in the outer Helmholtz plane (OHP). Ions with weakly bound solvation shells (mostly anions) may lose part of their solvation shell and form a chemical bond with the surface (so-called *specific adsorption*) and are located in the inner Helmholtz plane (IHP).

There are a number of theoretical descriptions of the structure of the double layer, including 1) the Helmholtz model (1879), 2) the Gouy-Chapman model (1910–1913), 3) the Stern model (1924), and 4) the Grahame model (1947).

The Electrical Double Layer Models

Helmholtz Model Helmholtz proposed that a layer of oppositely charged ions in the solution balances the metal surface charge. When a metal is in contact with an ionic solution, such as water (polar) with dissolved electrolytes, the metal surface has a high density of electrons, giving it a locally negative charge. The solvated positive ions (cations), such as H^+, Ca^{++}, and Mg^{++}, align themselves along the metal surface, producing a rigid electric double layer—no thermal motion of ions is considered. In the Helmholtz model, the potential dies off both sharply and linearly as we move away from the electrode, i.e., no interactions occur further away from the first layer of adsorbed ions (Figure 7.15). The electrolyte has no effect in this model.

The relation between the stored charge density ρ in the electrical double layer and the potential drop V across its plates is derived from the Poisson equation (slightly modified Equation 5.37):

$$\nabla \cdot E = \frac{\rho(r)}{\varepsilon_0 \varepsilon_r}$$

where E = electrical field
ε_o = permittivity of free space
ε_r = relative permittivity (which is dimensionless); relative permittivity effectively indicates the extent to which a dielectric has greater permittivity than a vacuum (ε_r is 1 for vacuum and very close to 1 for air but 78–80 for water at room temperature)
r = distance
ρ(r) = charge density

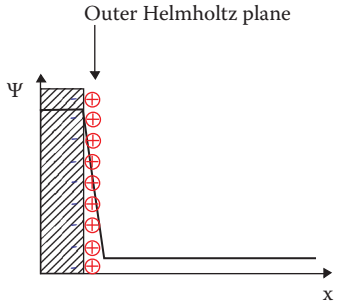


FIGURE 7.15 The Helmholtz model showing the outer Helmholtz plane (OHP).

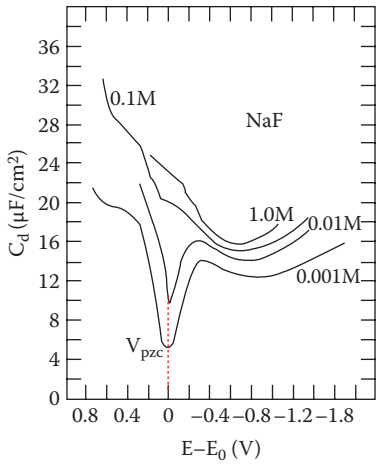


FIGURE 7.16 Differential capacitance as a function of applied voltage for a Hg electrode in NaF solutions at various concentrations.

The electrical field E is the differential of the electrical potential ψ [remember $E = -\nabla V$, where ∇ is the gradient operator (also called the *del vector operator*)]; thus, Equation 5.37 may also be rewritten as (replacing the potential V with ψ*):

$$\nabla^2 \cdot \Psi(r) = -\frac{\rho(r)}{\varepsilon_0 \varepsilon_r}$$

In the Helmholtz model $E = \psi/d$, or:

$$\rho(r) = \frac{\varepsilon_0 \varepsilon_r V}{d} \qquad (7.29)$$

with d the distance between the plates. The differential capacitance, C_d, in this model is given as:

$$C_d = \frac{\partial \rho(r)}{\partial V} = \frac{\varepsilon_0 \varepsilon_r}{d} \qquad (7.30)$$

The weakness of this model is that it predicts a constant differential capacitance, which contradicts experimental evidence gleaned from Figure 7.16, where we show the differential capacitance C_d as

* Note that in this electrochemistry section we use E, V, ϕ, and ψ interchangeably to symbolize a voltage or potential. Use of these different symbols arose in different application fields, and touching on all these application fields in this section we decided to retain the symbol most familiar to the practitioners in each of these fields.

a function of voltage and observe a strong voltage and electrolyte concentration dependence. At low electrolyte concentrations, the dip or U-shape in the C_d versus V curve occurs near the point of zero charge (pzc), V_{pzc}. The pzc of a solid surface (of an electrode or of a dispersion of colloidal particles) is the voltage or pH (pH_{pzc}) at which there are no net charges left at the solid's surface. The terms isoelectric point (IEP) and pzc are often used interchangeably.

The Helmholtz model only approximates the real picture illustrated in Figure 7.16 for very high electrolyte concentration and at very low temperatures where Brownian motion and diffusion are negligible; for most cases, however, we clearly need a more sophisticated model.

Gouy-Chapman In the Gouy-Chapman (G-C) model, because of the thermal motion of the ions in the solution, the population of positive charges (i.e., cations like H^+, Ca^{++}, and Mg^{++}) decreases exponentially with increasing distance from the metal surface (balance between Coulomb attraction and Brownian motion) (see Figure 7.17). The double layer is not compact anymore but of variable thickness with ions that are free to move. The model assumes a Poisson-Boltzmann (PB) distribution of the ions away from the surface and pictures ions as point charges that do not interact with each other. This model explains the effects of an applied potential and the nature of the electrolyte on the differential capacitance, C_d.

To introduce the notion of thermal motion of ions in an aqueous solution, a Maxwell—Boltzmann distribution (see Chapter 3, Equation 3.12) is introduced in the Poisson's equation to obtain the Poisson-Boltzmann distribution.

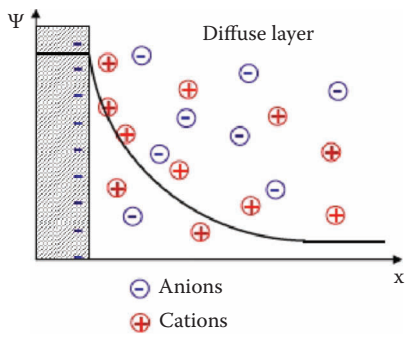


FIGURE 7.17 The Gouy-Chapman model. Thermal motion of the ions causes ions in the outer Helmholtz plane to diffuse away from the surface.

The number of ions at a point x close to the surface, where the field is $\psi(x)$, relative to a point far from the surface, where the potential is zero [$\psi(\infty) = 0$], in a Boltzmann distribution, is given by:

$$\frac{n_i(x)}{n_i(\infty)} = \exp\left(\frac{-w_i}{kT}\right) = \exp\left(\frac{-z_i e\psi(x)}{kT}\right) \quad (7.31)$$

where $n_i(x)$ is the number of ions at x, and w_i is the work required to bring an ion from an infinite distance away from the surface to the point x, i.e., $z_i e\psi(x)$ with z_i the charge on ion i. Combining the Boltzmann equation with the Poisson equation results in the Poisson-Boltzmann expression, whose solution is the potential profile $\psi(x)$:

$$\frac{d^2\psi}{dx^2} = -\frac{1}{\varepsilon_r \varepsilon_0}\sum_i z_i e n_i(\infty)\exp\left(\frac{-z_i e\psi}{kT}\right) \quad (7.32)$$

When dealing with a single symmetrical electrolyte ($|+z| = |-z|$), the PB equation has only two terms: one for the $+z$ and one for the $-z$ ions:

$$\frac{d^2\psi}{dx^2} = -\frac{zen(\infty)}{\varepsilon_r \varepsilon_0}\left[\exp\left(\frac{-ze\psi}{kT}\right) - \exp\left(\frac{ze\psi}{kT}\right)\right] \equiv \frac{2zen(\infty)}{\varepsilon_r \varepsilon_0}\sinh\left(\frac{ze\psi}{kT}\right) \quad (7.33)$$

where z = charge of the symmetrical electrolyte
n = electrolyte concentration (number of ions/m^3)
ε_r = dielectric constant of medium
ε_0 = permittivity of vacuum (F/m)
k = Boltzmann constant (J/K)
T = temperature (K)
e = elementary charge (C)

The two exponential terms in Equation 7.33 combine to form a hyperbolic sine function, and although this differential equation is readily solved, it is helpful to make a further assumption, i.e., that the potential is sufficiently low so that the sinh function may be replaced by its argument, i.e., for low potentials

$|z\psi| \leq 25$ mV, $\sinh y \approx y$ (this is accurate whenever the argument < 1), and it follows that:

$$\frac{d^2\psi}{dx^2} = \frac{2z^2e^2n(\infty)}{\varepsilon_r\varepsilon_0 kT}\psi = \kappa^2\psi;$$

$$\kappa = \text{“Debye parameter”} = \sqrt{\frac{2z^2e^2n(\infty)}{\varepsilon_r\varepsilon_0 kT}} \quad (7.34)$$

This is called the *Debye-Hückel approximation*. The resulting differential equation may be solved using the following boundary conditions: at x = 0 (the surface), $\psi = \psi_0$ and at $x = \infty$ (in the bulk of the solution), $\psi = 0$. This yields an exponential function for $\psi(x)$, with a decay constant made up of a cluster of constants captured by κ and referred to as the *Debye parameter*. It describes the decay length [(κ^{-1}) or the Debye length in units of length]:

$$\psi = \psi_0 e^{-\kappa x} \quad (7.35)$$

The potential profile $\psi(x)$ is illustrated in Figure 7.18.

Once the potential profile is known, it may be substituted into the Boltzmann equation (7.31) to yield the ion concentration profiles:

$$C_i(x) = C\exp\left[-\frac{z_i e}{kT}\psi_0\exp(-\kappa x)\right] \quad (7.36)$$

The concentration profiles for both cations and anions in the case of C = 10^{-2} M NaCl and ψ_0 = 25 mV are shown in Figure 7.19. The location of $x = \kappa^{-1}$ is marked, and this is regarded as the “thickness of the double layer.”

The effect of the electrolyte concentration on the potential ψ as a function of distance from the surface for two different concentrations is illustrated in Figure 7.20.

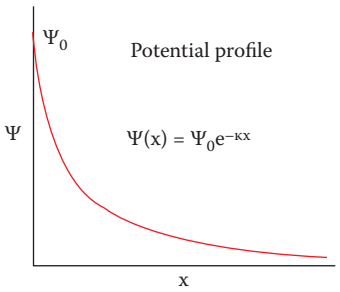


FIGURE 7.18 Potential profile in the double layer.

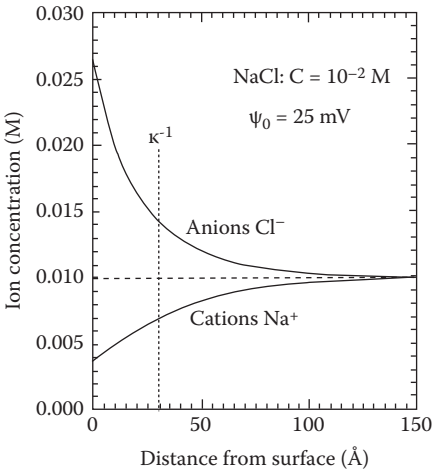


FIGURE 7.19 Concentration profile for cations and anions with ψ_0 = 25 mV and C = 10^{-2} M.

The range of values for the Debye length in water at 25°C can be evaluated from $\kappa^{-1} \approx \frac{0.30}{z\sqrt{C(M)}}$ nm [where $C(M)$ is the concentration in moles], which give values of $0 < \kappa^{-1} < 1000$ nm. The lower the electrical conductivity of a solution, the higher the Debye-Hückel length, although it is hard to make $x = \kappa^{-1}$ much longer than a few tens of nanometers (value at 1 μM of salt in water). This is because higher-resistivity water samples are hard to prepare.

The Debye-Hückel length for a multivalent electrolyte is given by:

$$\kappa = \sqrt{\frac{e^2}{\varepsilon_r\varepsilon_0 kT}\sum_{i=1}^{n} C_i z_i^2} \quad (7.37)$$

where z_i = electrolyte valence

C_i = ion concentration (number of ions/m^3)

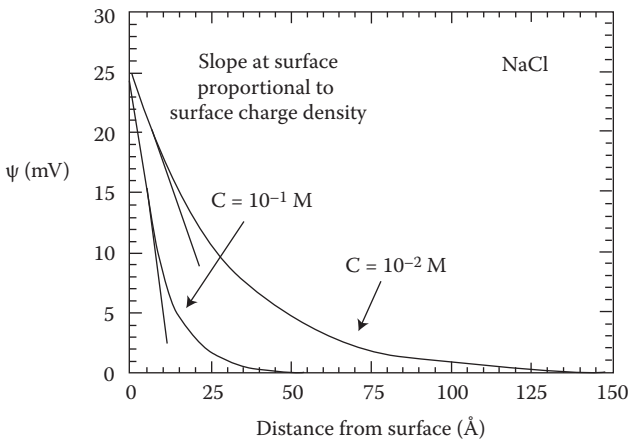


FIGURE 7.20 Potential as a function of distance from the surface for two different concentrations with ψ_0 = 25 mV.

ε_r = dielectric constant of the medium
ε_0 = permittivity of vacuum (F/m)
k = Boltzmann constant (J/K)
T = temperature (K)
e = elementary charge (in C)

Higher-valency ions are more effective in screening surface charge; therefore, a 3-3 electrolyte has a shorter Debye length at the same concentration than a 1-1 electrolyte.

Whereas the Helmholtz model represents the solid/electrolyte interface well at the high concentration limit, the Gouy-Chapman (G-C) model best represents the high dilution limit. Because ions are regarded solely as point charges, the G-C model permits them to pack without limit into the layer adjoining the surface. Because ions are characterized only by their charge (i.e., they have no chemistry), they cannot specifically adsorb to surfaces in excess of the charge required for neutralization, nor can they adsorb to surfaces of the same charge. In the G-C model, the differential capacitance C_d (Equation 7.30) varies with potential in a U-shape as illustrated in Figure 7.21. Here, we see that the differential capacitance versus voltage takes on at least one aspect found in the experimental data presented in Figure 7.16, i.e., a U-shaped minimum at the *pzc*, V_{pzc}. The G-C model still leads to an unlimited increase in differential capacitance C_d with applied bias, although again in violation of experimental evidence. A yet more sophisticated model is needed.

Stern Model The Stern model combines the concepts of the Helmholtz and G-C models. Based on minimizing the free energy at the interface, one obtains a double layer with the positive ions of a finite size closest to the metal surface constrained into a rigid outer Helmholtz plane (OHP), whereas outside this plane, the positive ions are dispersed as in the G-C model (see Figure 7.17). The monolayer of ions in the OHP (also the "Stern layer") directly adjacent to the surface generally effects partial neutralization of the surface charge. The remainder of the surface charge is neutralized in a G-C diffuse layer adjacent to the Stern layer. The Stern layer is assumed to reside between the surface and a plane (the Stern plane), passing through the centers of the first layer of hydrated ions, a distance δ from the surface (OHP). The distance δ is generally 2–3 Å, and the potential at this location is called the *Stern potential* or ψ_H. The Stern potential is the effective surface potential. The differential capacitance C_d (see Equation 7.30) is now the sum of two series capacitors, i.e., C_H, the voltage-independent capacitance of the charge at the OHP, and C_D, a truly diffuse charge, or:

$$\frac{1}{C_d} = \frac{1}{C_H} + \frac{1}{C_D} \tag{7.38}$$

With large ion concentrations or at large polarizations, C_D is large, and the voltage-independent C_H dominates the differential capacitance. Typical double-layer capacitance (C_H) for a metal is 10–40 μF/cm². At low concentrations, C_D varies with potential in a U-shape and so does C_d, as illustrated in Figure 7.22. Here we see that the differential capacitance versus voltage starts to look more like the experimental data presented in Figure 7.16.

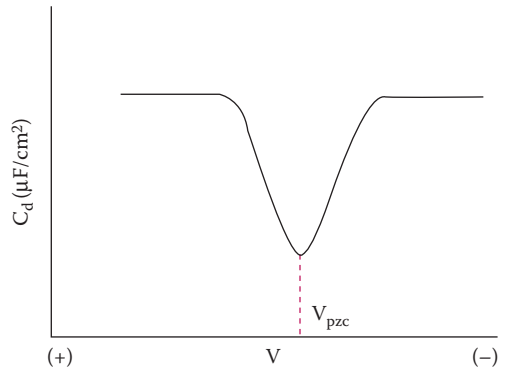


FIGURE 7.21 Variation of the differential capacitance C_d with voltage *V*, showing the minimum at the point of zero charge, V_{pzc}.

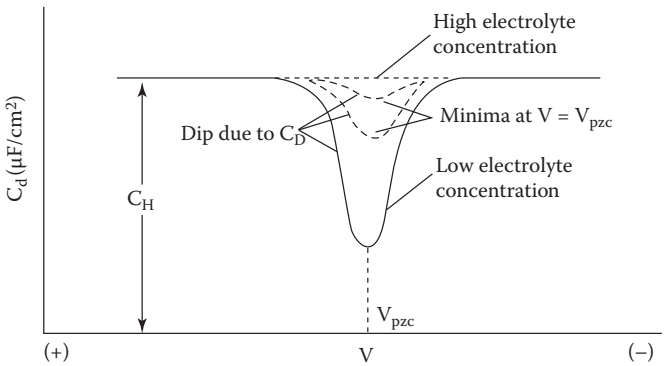


FIGURE 7.22 Differential capacitance versus applied voltage according to the Stern-Gouy-Chapman models.

Grahame Model The Grahame model further refines the Stern model. It assumes that specifically adsorbed ions are desolvated and may approach the electrode surface closer than the outer Helmholtz plane (OHP), and thus feel a greater potential, than the solvated ions. In this model, an inner Helmholtz plane (IHP) passes through the center of specifically adsorbed ions, and an OHP passes through solvated and nonspecifically adsorbed ions (see Figure 7.14). In the case of a metal immersed in pure water, the electrons within the metal cause the polar water molecules to adsorb to the surface and orient themselves so as to create two thin planes of positive and negative charge. Most materials spontaneously develop a thin native oxide coat (e.g., SiO_2 on Si) and surface charges when exposed to air, or even more so when submerged in an electrolyte. In the case of glass or Si (with its native SiO_2 film), the Si-OH (silanol) groups are partially ionized, and depending on the pH, the surface becomes more or less negatively charged. The adsorbed charges and ionized charges are located in the IHP, and the compensating solvated cations are in an opposing rigid OHP and the diffuse G-C layer. Both the IHP and OHP layers are very narrow (in the range of atomic dimensions). The diffusion layer is larger and extends from the OHP to the bulk (~100 Å in a solution $>10^{-2}$ M). With the Grahame model, shifts in the potential dip toward more positive or more negative values than the pzc can be explained.

In Figure 7.23, we illustrate three example mechanisms that generate surface charges at the solid/liquid interface: ion adsorption (e.g., Ca^{2+} onto a lipid layer), ion dissociation of surface groups (e.g., solid-COOH → solid $-COO^-$ + H^+), and isomorphous substitution. When a surface is fully ionized, a charge occupies about 0.5 nm^2.

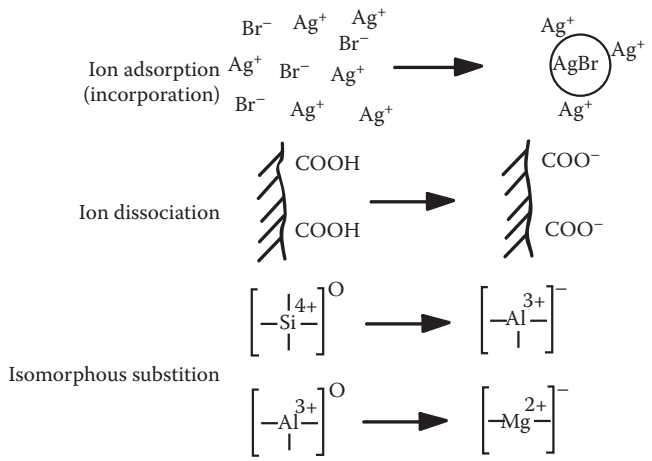


FIGURE 7.23 Examples of surface charges.

The Zeta Potential and Electrokinetic Phenomena When a liquid moves over a solid surface, a thin layer of liquid along that solid surface stays motionless. Liquid movement begins only at a shear plane in the diffuse double layer at a distance, x_0, away from the wall. The shear plane (slipping plane) is an imaginary surface separating the thin layer of liquid bound to the solid surface that shows elastic behavior different from the rest of the liquid showing normal viscous behavior. Adsorbed species are rigidly held inside the shear plane while outside of it there is free movement. The shear plane can therefore roughly be associated with the outer Helmholtz plane (OHP). The value of the electric potential at the shear plane with respect to the bulk solution is called the *electrokinetic* or *zeta potential* (ζ). The electrokinetic potential ζ—from which all electrokinetic effects (Figure 7.24) can be derived—represents the value of the electrostatic potential, ψ, at x_0. Its value at any point is determined from the profile of the electrostatic potential as a function of distance from the wall into the moving fluid. The thickness δ of the diffuse layer is defined as the distance from the immobile OHP or Stern layer to a point at which the electrostatic potential has dropped to 37% of the zeta potential. The amount and type of ionic species in the solution greatly influence the zeta potential; for example, the larger the concentration of added indifferent electrolyte in the moving solution, the smaller the ζ (see Figure 7.25). The electrokinetic effects that are controlled by the zeta potential are electrophoresis, electro-osmosis, streaming potential, and sedimentation potential (see Figure 7.24). As listed in Figure 7.24, the movement of charged particles in an electric field is termed *electrophoresis*. When the charged solid surface is fixed and the electric field causes a movement of the liquid, this is called *electro-osmosis*. Forcing a liquid through a capillary or porous plug induces a difference of electric potentials called *streaming potential*, and forced movement of charged solid particles in a liquid,

- Electrophoresis: Movement of particles (e.g., ions or colloidal particles) in a stationary fluid by an applied electric field.
- Electro-osmosis: Movement of liquid past charged solid surfaces (or perhaps through a membrane) under the influence of an applied electric field. The effects are often studied in a fine capillary to maximize the solid surface area to liquid volume.
- Streaming potential: Creation of an electric field as a liquid moves past a stationary charged surface.
- Sedimentation potential: Creation of an electric field when a charged particle moves relative to stationary fluid due to the gravitational force.

Phenomenon	Electric Field	Solid Phase	Liquid Phase	Property
Electrophoresis	Applied	Mobile	Fixed	Mobility
Electro-osmosis	Applied	Fixed	Mobile	Pressure/ velocity
Streaming potential	Generated	Fixed	Mobile	Potential
Sedimentation potential	Generated	Mobile	Fixed	Potential

FIGURE 7.24 Electrokinetic effects controlled by the zeta potential.

e.g., as a result of gravitation, induces a difference of electric potentials, i.e., the *sedimentation potential*. To a certain approximation, all four of the electrokinetic effects are different manifestations of the same phenomenon. As a rough approximation, the electrophoretic mobility (the ratio of the velocity of particles to the field strength), induced pressure difference in electro-osmosis, streaming potential, and sedimentation potential are all proportional to the zeta potential (ζ). The size of particles that can be calculated from these experiments corresponds to the particle dimensions plus the double-layer thickness.

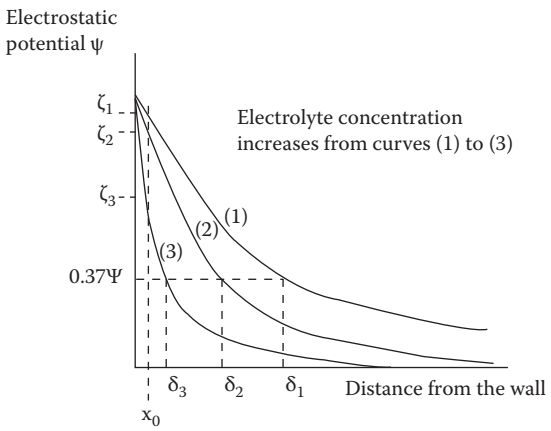


FIGURE 7.25 The electrokinetic potential, ζ, is the electrostatic potential, ψ, at a distance, x_0, from the surface, that is, the shear plane where the fluid movement begins. The concentration and type of ions in the solution influence the position of the curves and thus the electrokinetic potential. Ions and impurities at the surface influence these curves as well.

Marian Smoluchowski (1921) was the first to properly derive an equation to calculate the zeta potential from the electrokinetic mobility, via:

$$v = \frac{\varepsilon_r \varepsilon_0 \zeta}{\eta} E \quad \text{or} \quad \mu_E = \frac{\varepsilon_r \varepsilon_0 \zeta}{\eta} \tag{7.39}$$

where v = velocity
ε_r = media dielectric constant
ε_0 = permittivity of free space
ζ = zeta potential
E = electric field
η = medium viscosity
μ_E = electrophoretic mobility

In this simplified derivation, the electrostatic driving force is opposed by the frictional force only, and it is assumed that $\kappa a \gg 1$, where κ is the Debye parameter and a is the particle radius. In other words, the electrical double layer thickness is assumed to be much smaller than the particle radius. Only when there is a low zeta potential, large colloidal particles, and high ionic strengths does this equation produce good results. For example, with values of zeta potentials with practical meaning, i.e., $\zeta <$ 120 mV, the error is negligible when $\kappa a > 100$. The zeta potential of an aqueous solution in contact with glass can have a magnitude as high as 100 mV.[3]

Many commercial zetameters have the above Smoluchowski equation built into the software, usually together with the electric permitivity and viscosity data for water at 25°C.[4]

In the past, electro-osmosis and electrophoresis were the most popular methods to determine zeta potentials. Recently, electro-osmosis also has been studied for implementation as the pumping mechanism in commercial fully automatic instruments. However, pH, ionic strength, and surface impurities all affect the electrokinetic potential and conspire to make electro-osmosis a rather difficult pumping mechanism to control (see Chapter 6).

Electrophoresis and electro-osmosis potential were also covered in Chapter 6 on fluidics (both AC and DC electrokinetic phenomena are discussed). Here we dwell a bit further on the stability of colloidal dispersions as they constitute one of the key avenues toward making nanoparticles.

Colloidal Dispersions Colloidal dispersions are among the most important types of systems encountered both in nature and in the manufacture of a wide variety of products, and their electrokinetic properties are among the most important in effecting their characterization. Colloids also form one of the most important gateways toward bottom-up nanotechnology (i.e., nanochemistry). (See Volume III, Chapter 3 on the comparison of top-down and bottom-up approaches in nanotechnology.)

Colloids are mixtures in which one material is dispersed in another, with the dispersed material (the "dispersoid") divided into particles that range in size from approximately 1 nm to 10 μm. Colloids have many properties distinct from those of solutions of ordinary-sized molecules or from those of mixtures of macroscopic phases (e.g., slurries of visible particles). The dispersed material may either be dissolved (i.e., in the case of solutions of macromolecules), giving a lyophilic colloid, or they may constitute a separate phase, giving a lyophobic colloid. We limit the present discussion to lyophobic colloids, which may be further classified according to the types of phases that compose them, as shown in Table 7.9. We further limit our study to "sols," i.e., solid particles dispersed in liquids.

One of the features distinguishing the behavior of lyophobic colloids is their inherent instability. The stability ("shelf life") of a lyophobic colloid is kinetic in nature rather than thermodynamic, i.e., it depends on how slowly the inexorable processes of self-aggregation take place. Aggregation is often accompanied by sedimentation, which ultimately leads to macroscopic phase separation. The aggregation of colloidal particles is a spontaneous process because of the integrated effect of the attractive van der Waals interactions between the molecules in the different particles. As we saw above, these van der Waals forces consist chiefly of London or "dispersion" forces. Although these interactions are extremely short ranged, with a potential that varies as $-C/r^6$ (where C is a constant dependent on molecular properties, and r is the distance between the molecules; see Equation 7.18), the integrated effect between all the molecules of one colloid particle with those of a nearby colloid particle is much longer ranged, with a potential that varies as $-H\sigma/12L_0$ (Equation 7.25). In this expression, H is the Hamaker constant and L_0 is the equilibrium distance between repulsive and attractive forces between the particle surfaces. The constant H depends on C and the molecular density ρ in the particles (see Equation 7.24). These

Table 7.9 Lyophobic Colloids

Dispersed Phase	Dispersion Medium	Common Name	Examples
Liquid	Gas	Aerosol	Fog, fine sprays
Solid	Gas	Aerosol	Smoke, viruses, tear "gas"
Gas	Liquid	Microfoam	Shaving cream, beer foam
Liquid	Liquid	Emulsion	Milk, mayonnaise
Solid	Liquid	Sol	Paint, ink, clay
Gas	Solid	Solid microfoam	Foam plastics, marshmallow
Liquid	Solid	Solid emulsion	Butter, cheese, opal, pearl
Solid	Solid	Micro(nano) composite	Injection-molded objects

integrated attractive forces between colloid particles are referred to as *long-range van der Waals forces* (see Figure 7.26a). They are generally important when ordinary colloid particles are within 100 nm of each other and become large as the particle surfaces approach one another. The result is that colloid particles aggregate quickly as external effects, such as Brownian motion and/or bulk fluid motion (e.g., by stirring), bring them within range. In aqueous media, one effect that can slow down the aggregation process is the presence of electric charge on the surfaces of the particles as shown in Figure 7.26b. Solid surfaces in contact with water generally acquire a charge by one or a combination of the mechanisms we illustrated earlier in Figure 7.23 (e.g., oxide particles acquire charge via the pH-dependent ionization of their superficial –OH groups). These charged solid surfaces attract a cloud of ions of opposite charge (counterions) to their vicinity. As the particles approach, these clouds of like charge begin to overlap, setting up repulsion.

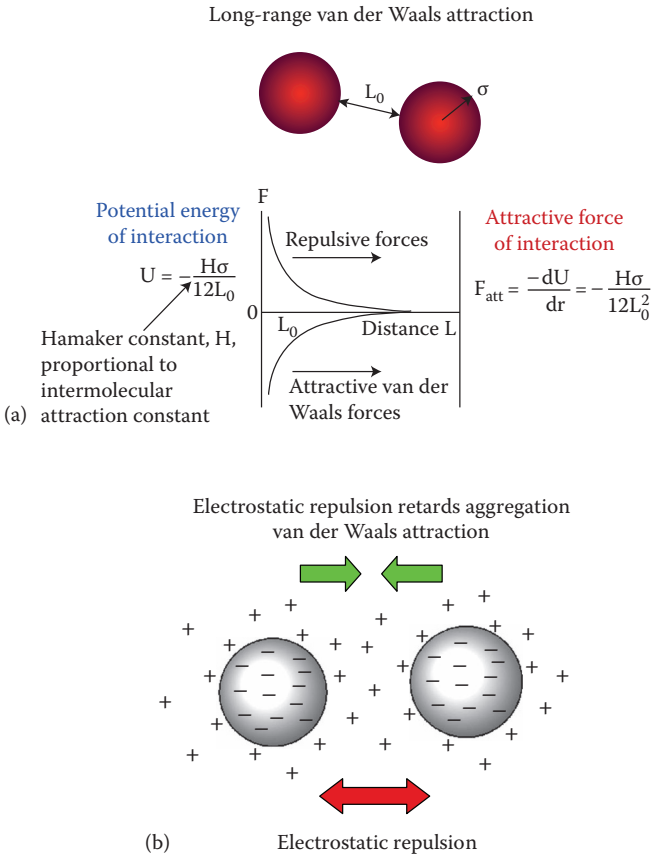


FIGURE 7.26 (a) The aggregation of colloidal particles is a spontaneous process because of the integrated effect of the attractive van der Waals interactions between the molecules in the different particles. These consist chiefly of London or "dispersion" forces, that is, forces caused by synchronous oscillations of the electron clouds of nearby molecules causing strong attraction between their temporal dipoles. (b) In aqueous media, one effect that may slow down the aggregation process is the presence of electric charge on the surfaces of the particles. Such charged surfaces attract a cloud of ions of opposite charge (counterions) to their vicinity. As the particles approach, these clouds of like charge begin to overlap, setting up repulsion.

When this electrostatic repulsion is put together with the long-range van der Waals attraction, a total interaction curve of the type shown in Figure 7.27 results. This adding of these forces was first carried out by Derjaguin and Landau in Russia and Verwey and Overbeek in the Netherlands in the 1940s in the so-called the *DLVO theory*. The curve that results when combining these forces has an energy barrier, and approaching particles must possess sufficient kinetic energy to overcome that barrier if they are to aggregate. If the barrier is large relative to *kT*, the probability of aggregating (the sticking probability) during a collision is low. In other words, the rate of aggregation of colloidal particles depends on the height of the DLVO barrier. For a given colloid (given Hamaker constant *H* and given particle radius σ), the electrostatic repulsion (barrier height) depends on the zeta potential and the double-layer thickness, which in turn depends on the ion concentration and valency of the electrolyte.

The stability of hydrophobic colloids thus depends on the zeta potential: when the absolute value of the zeta potential is above 50 mV, the dispersions are very stable because of mutual electrostatic repulsion. When the zeta potential is close to zero, the coagulation (formation of larger assemblies of particles) is very fast, causing fast sedimentation. For a 1-1 electrolyte and a given colloid, the effect of increasing the salt concentration on the barrier is shown in Figure 7.28. Small changes in concentration *C* have a large effect on the barrier height, and it is possible to identify, for any given case, a particular concentration (called the *CCC*, or *critical coagulation concentration*) at or above which rapid aggregation occurs. For a monovalent electrolyte, the CCC is generally between 50 and 200 mM. For counterions of higher

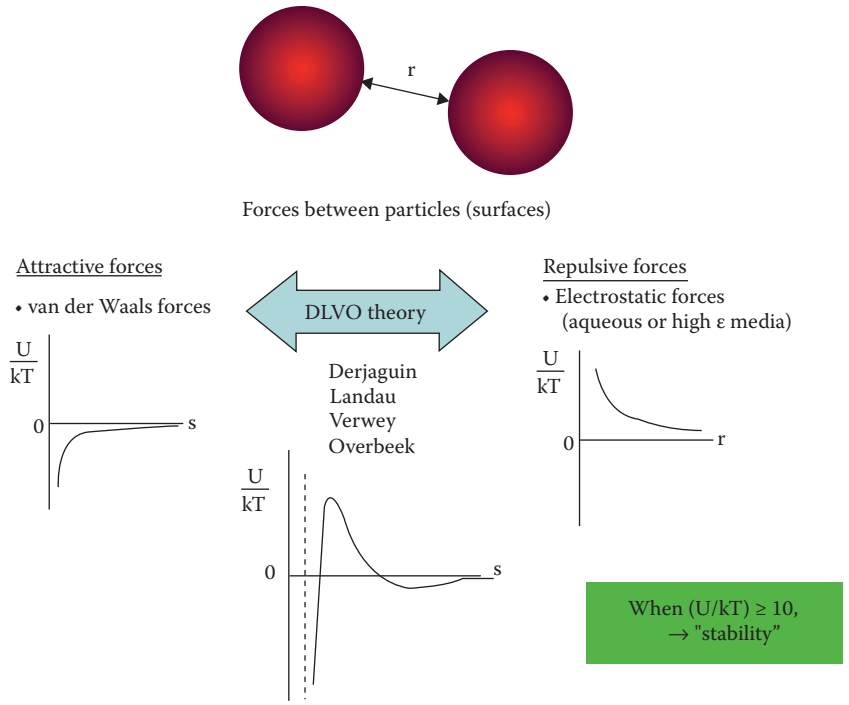


FIGURE 7.27 When electrostatic repulsion is put together with long-range van der Waals attraction, a total interaction curve of the type shown here results.

valence, the CCC decreases drastically, in accord with the Schulze-Hardy rule:

$$\text{CCC} \propto \frac{1}{Z_i^6} \tag{7.40}$$

where Z_i is the valence of the counterions.

Casimir Forces

An important prediction of quantum theory is the existence of irreducible fluctuations of electromagnetic fields even in a vacuum and at 0 K, i.e., the existence of a so-called *zero point energy* (ZPE). The

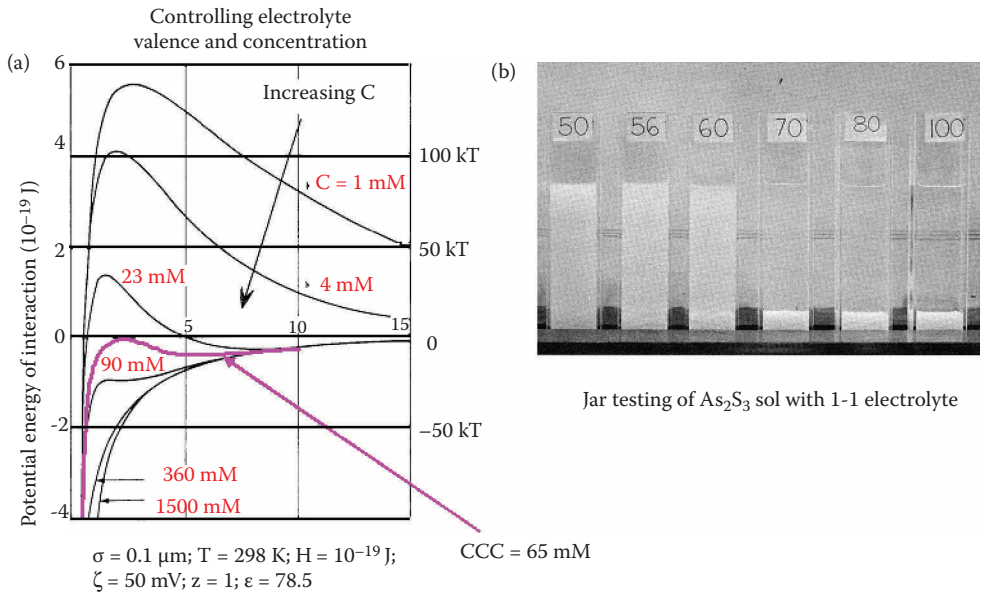


FIGURE 7.28 (a) For a 1-1 electrolyte and a given colloid, the effect of increasing the salt concentration on the barrier is shown. Small changes in *C* have a large effect on the barrier, and it is possible to identify, for any given case, a particular concentration (called the *CCC*, or *critical coagulation concentration*) at or above which rapid aggregation will occur. This can be determined experimentally by (b) "jar testing." For a monovalent electrolyte, the CCC is generally between 50 and 200 mM. For counterions of higher valence, the CCC drops drastically, in accord with the Schulze-Hardy rule, as shown.

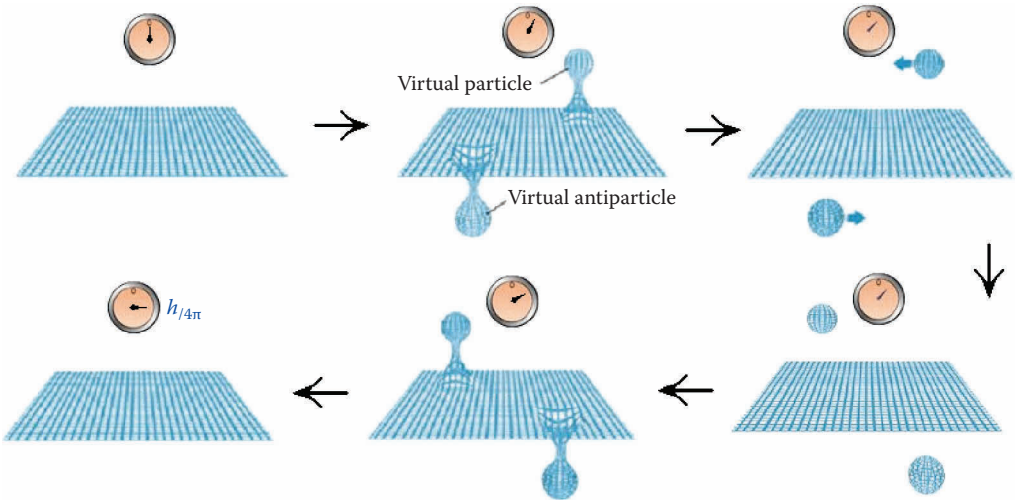


FIGURE 7.29 Virtual particles pop out of a vacuum and wander around for an undefined time and then pop back, thus giving the vacuum an average zero point energy. (From http://www.casimir.rl.ac.uk/zero_point_energy.htm.)

latter can be described as a turmoil of virtual particles that come in and out of existence, violating the energy-momentum conservation of the system for very short periods of time, as described by Heisenberg's uncertainty principle (Equations 3.106 and 3.107) (Figure 7.29). In Chapter 3, we saw how this zero point energy can be measured through the Casimir force, a small attractive force between two close parallel, uncharged conductive plates in a vacuum (Figures 3.69 and 7.30). This force, inversely proportional to the fourth power of the distance between the plates (see Volume III, Chapter 8 on actuators), comes about because the fluctuating virtual particles exert a "radiation pressure" on the plates, which on average is greater outside the plates than between them, as shown in Figure 7.30. The Dutch theoretical physicist Hendrik Casimir predicted this startling phenomenon in 1948 while he was working at Philips Research Laboratories in Eindhoven on, of all things, a problem concerning colloids!

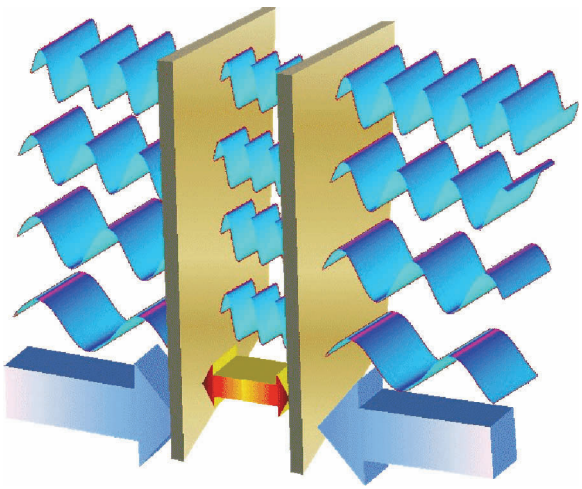


FIGURE 7.30 The Casimir force. Fluctuating virtual particles exert a "radiation pressure" on the plates that on average is greater outside the plates than between them. (Based on http://www.casimir.rl.ac.uk/zero_point_energy.htm.)

At the time of his discovery, Casimir was studying the properties of "colloidal solutions." The properties of such solutions, we know now, are determined by van der Waals forces. One of Casimir's colleagues, Theo Overbeek (of the DLVO theory fame), realized that the theory that was used at the time to explain van der Waals forces did not properly explain some of the experimental evidence on colloids. The prevailing theory Overbeek found lacking was the one by Fritz London, who, in 1930, described the interaction potential between two atoms as a result of fluctuations of the atomic electric dipole moments as $U \sim 1/r^6$ (see Equation 7.18). Overbeek asked Hendrick Casimir to investigate this problem, and by 1948, Casimir and Polder[5] had extended the London force to larger separations and found that retardation effects weaken this force, and that the attractive atom-surface force at large distances comes with a different power law than in the simpler short-range case. If the distance between fluctuating charges is large enough, it takes a finite amount of time for the

electromagnetic field from one charge to reach the other (i.e., the speed of light c is finite). By the time the second charge responds to the field, the momentary charge configuration at the first position will have changed, and the charge fluctuations will fall out of step with each other. The strength of interaction is therefore always weakened, and its distance dependence is changed; for example, $1/r^6$ energy between small particles becomes $1/r^7$. This retarded van der Waals force is also known as the *Casimir-Polder force*.

Casimir noticed that this result could be interpreted in terms of vacuum fluctuations. He asked himself what would happen if there were two mirrors—rather than two molecules—facing each other in a vacuum. It was this work that led to his famous prediction of an attractive force between reflecting plates. In the late 1950s and early 1960s, Evgenii Lifshitz took into account the dielectric properties of the media and thermal fluctuations and developed a general theory for van der Waals type forces, of which the Casimir-Polder force is a limiting case.[6]

Although experimental evidence of an attractive force between neutral atoms and between neutral atoms and surfaces has existed for a long time, it was not until 1993 that Ed Hinds' group at Yale was able to measure that atom-surface force at larger distances or in the retarded or Casimir-Polder regime.[7]

Although the Casimir force is too small to be observed for plates that are several meters apart, it can be measured if the plates are within micrometers of each other. For example, we calculate in Volume III, Chapter 8 on actuators that two plates with an area of 1 cm^2 separated by a distance of 1 μm have an attractive Casimir force of about 10^{-7} N, roughly the weight of a water droplet that is half a millimeter in diameter. Although this force might appear small, at distances below 1 μm the Casimir force becomes the strongest force between two neutral objects. Indeed, at separations of 10 nm—about 100 times the typical size of an atom—the Casimir effect produces the equivalent of 1 atm. Although we do not deal directly with such small distances in everyday life, they are important in nano- and microscale structures (NEMS and MEMS). With the decrease in device dimensions, additional forces have to be considered that are normally neglected in macrosystems. For example, one of the principal causes of malfunctioning in surface micromachined MEMS devices, we discuss in Volume II, Chapter 7, is stiction, i.e., the collapse of movable elements onto nearby surfaces, resulting in their permanent adhesion. This can occur during fabrication, especially because of capillary forces, or during operation. The Casimir effect is potentially an important underlying mechanism causing this phenomenon. The problem is technologically important because it adversely affects the production yield on batch-fabricated devices.

An overview of recent measurements of the Casimir force is presented in Volume III, Table 8.21.

Electrochemistry

Introduction

Electrochemistry is the study of reactions taking place at the interface of a solid conductor and an electrolyte (an ionic conductor with ions as the mobile charge-carrying species). Using sensor examples, we introduce some of the most important concepts in electrochemistry, i.e., potentiometry, voltammetry, two- and three-electrode systems, Marcus theory of electron transfer, reaction rate- and diffusion rate-controlled electrochemical reactions, and ultramicroelectrodes (UMEs). In this same section, we also learn how different materials such as metals, semiconductors, solid electrolytes, and mixed conductors exposed to a solution offer the possibility for building different types of electrochemical sensors. As summarized in Table 7.10, different types of mobile charge carriers characterize these different materials. Metal electrodes (e.g., Pt) feature electrons as the mobile charge carriers, whereas semiconductors (e.g., n-Si) have electrons and/or holes to play that role. Solid electrolytes (e.g., LaF_3) possess mobile ions. Insulators (e.g., SiO_2), however, have no free charge carriers, whereas mixed conductors (e.g., IrO_x) have both ions and electrons.

We conclude this section by providing more electrochemical sensor examples, including ion-sensitive field-effect transistors (ISFETs), potentiometric

TABLE 7.10 Charge Carriers for Different Electrode Materials and Different Types of Sensors Based on Their Interface with a Solution

Material (Examples)	Charge	Type of Electrochemical Sensor
Metals (Pt, Au, etc.)	Electrons	Electrodes of the first kind sensor for redox species with the highest electron exchange current $i_{0,e}$
Semiconductors (GaAs, Si, etc.)	Electrons and holes	pH sensor for the least conductive semiconductors (wide bandgap), redox couple sensitive for the more conductive semiconductors (narrow bandgap)
Insulators (SiO_2, Al_2O_3, etc.)	No mobile charges	pH-sensitive surface in an ISFET
Ionic conductor (LaF_3, ZrO_2(s), $RbAg_4I_5$, sea water, etc.)	Ions as mobile carriers	For detection of ionic species with the highest ion exchange current $i_{0,i}$
Mixed conductor (Pd, plasma, IrOx, etc.)	H_2 in Pd [$H^+ + e^-$], IrO_x[$H^+ + e^-$],	Sensor application depends on the relative magnitude of electron exchange and ion exchange currents $i_{0,e}$ and $i_{0,i}$

and amperometric immunosensors, and glucose sensors.

Interface Processes

In electrochemistry, one studies reactions in which charged particles (ions or electrons) cross the interface between two phases of matter. Typically this involves a solid phase (the electrode) and a conductive solution (or electrolyte). These interface processes are represented as chemical reactions and are known generally as *electrode reactions*. Electrode reactions take place within the double layer and produce a slight unbalance of the electric charges in both the electrode and the solution. External manipulation of the interfacial potential difference affords an important way to exert control over an electrode reaction. The interfacial potential difference at a solid/electrolyte interface is limited to only a few volts at most. This does not seem like very much until one realizes that this potential spans only within a layer of water molecules and ions that attach themselves to the electrode surface—normally only a few atomic diameters thick. Thus, a very small voltage produces a very large potential gradient. For example, a potential difference of 1 V across a typical 10^{-8}-cm interfacial boundary layer amounts to a potential gradient of 100 million V/cm—a very significant electrical field and hence an important potential influence on the charge transfer reaction.

Charge transport in electrodes occurs via motion of electrons (or holes), and charge transport in the electrolyte occurs via motion of ions (positive and negative). With reference to Figure 7.31, it is easily recognized that for current to flow across the solid/liquid interface (in this case across a metal/electrolyte solution interface) an electrochemical reaction must occur. A process at the solid/electrolyte interface where charges (e.g., electrons) are transferred across this interface is called a *Faradaic process*. For example, an *oxidant* O (say, Ag^+) gets reduced in a *cathodic* reaction (reaction on the *cathode*) to become a *reductant* R (say, $Ag^+ + e^- \rightarrow Ag$):

$$O + ne^- \rightarrow R \qquad \text{Reaction 7.1}$$

An oxidant O is defined as an oxidizing agent, i.e., a reactant that oxidizes another reactant and which is itself reduced. A reductant R is a reducing agent, i.e., a reactant that reduces another reactant and which is itself oxidized. For a complete circuit, a *counter electrode* (CE) must also be present in the *electrochemical cell* of Figure 7.31, so that an *anodic reaction* (on the *anode*) may occur (e.g., $Cd - 2e^- \rightarrow Cd^{2+}$):

$$R' - ne^- \rightarrow O' \qquad \text{Reaction 7.2}$$

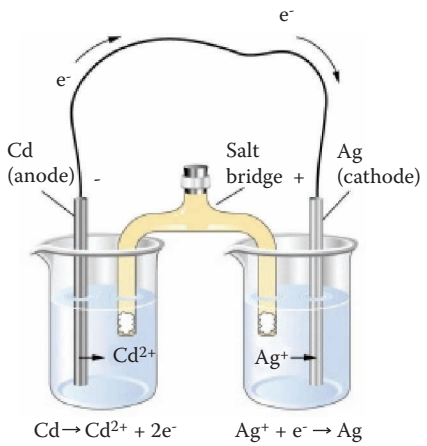


FIGURE 7.31 An electrochemical cell with an anode and a cathode and a salt bridge connecting the two half-cells.

Redox reactions (Reactions 7.1 and 7.2) are electron transfer processes where an oxidizing agent receives electrons, becoming a reducing agent; and a reducing agent gives off electrons, becoming an oxidizing agent. Oxidizing (O and O′) and reducing agents (R and R′) form so-called *redox couples* ($ox + ne^- \Leftrightarrow red$, e.g., $Ag^+ + e^- \rightarrow Ag$). The *anode* of the electrochemical cell in Figure 7.31 is the electrode where the oxidation occurs, and the *cathode* is the electrode where the reduction takes place.

An *electrochemical cell* typically consists of two electronic conductors (also called *electrodes*) connected through an external wire and two *half-cells* filled with an ionic conductor (called an *electrolyte*) and a redox couple. A *salt bridge* is an ionic solution connecting the two half-cells (half-reactions) that completes the circuit. No further electrode reaction would take place if electrical neutrality were not maintained via this salt bridge. The salt bridge (also *Luggin capillary*) maintains electrical neutrality in the two half-cells by migration of ions through the porous material and also prevents mixing of the half-cell solutions.

If the anode and cathode solutions were allowed to mix, the chemical driving force would lead to the transfer of electrons directly between reactants rather than through the external circuit (wire, load). A salt bridge is often made by equipping a U-shaped glass tube with Vycor frits at both ends and filling it with a viscous, aqueous ionic solution that allows charge transport without permitting the two solutions to mix with each other. The solution may be an agar gel loaded with KCl or KNO_3. The K^+, Cl^-, and NO_3^- ions have similar mobilities and migrate in opposite directions with similar velocities. This minimizes the appearance of *junction potentials* in the bridge. Junction potentials occur at the interface of solutions with different ion concentrations as, for example, between the electrolyte solutions of a reference electrode and the sample. These potential differences are caused by the different mobilities of ions at the interface of the two solutions (Figure 7.32). There is no need for concern about junction potentials as long as they are kept small and constant during an analysis.

The salt bridge is one part of an electrochemical sensor that, we will see further below, is particularly cumbersome to miniaturize (see ISFET section).

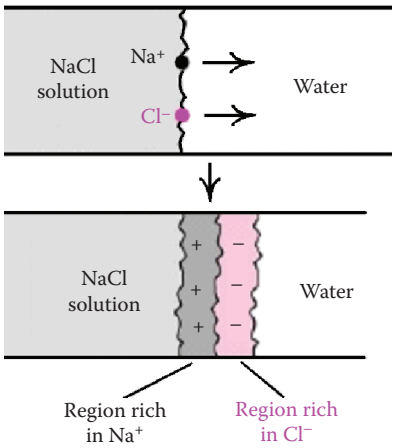


FIGURE 7.32 Ions diffuse from a region of high ionic strength to a region of lower ionic strength and will migrate (diffuse) at different rates, depending on their relative charge and size (z/r). This creates regions with higher net negative charge and regions of higher net positive charge at the interface (or junction). The difference in local charge represents some work (ΔG free energy), or chemical potential, called a *junction potential*.

At each electrode, an electrochemical reaction occurs. This reaction is called a *half-cell reaction* (because there are two electrodes in a typical cell at which reactions occur). There are two fundamental types of half-cell reactions: an oxidation reaction (in Figure 7.31, the oxidation reaction is $Cd \rightarrow Cd^{2+} + 2e^-$) and a reduction reaction (in Figure 7.31, the reduction reaction is $Ag^+ + e^- \rightarrow Ag$). In writing the overall reaction the charge must be conserved so that the number of electrons lost in oxidation equals the number of electrons gained in reduction:

$$Cd(s) + 2Ag^+(aq) \rightarrow 2Ag(s) + Cd^{2+}(aq) \qquad \text{Reaction 7.3}$$

Reactions occur pairwise; one cannot have oxidation without reduction. Because all redox reactions occur pairwise, i.e., reduction and oxidation always occur at the same time, we cannot measure the cell potential for just one half-cell reaction. This means that we must establish a relative scale for cell potentials.

The standard hydrogen electrode (SHE) serves as the ultimate reference electrode (REF E) for calculating the potential of a single electrode. Hydrogen gas is bubbled at 1 atm into a 1 M hydrochloric acid solution at 25°C (298 K) with a platinum electrode (Pt black) providing the reaction surface (Figure 7.33).

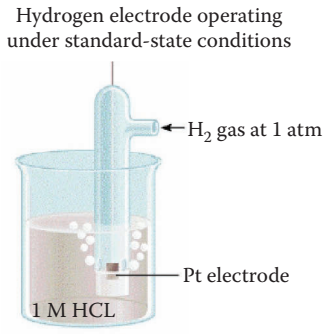


FIGURE 7.33 The standard hydrogen electrode (SHE).

Under these standard-state conditions, the potential for the reduction of H^+ is taken to be exactly zero:

$$2\ H^+\ (1\ M) + 2\ e^- \rightarrow H_2\ (1\ atm)\ E^0 = 0\ V \qquad \text{Reaction 7.4}$$

The voltage associated with a reduction reaction at an electrode with all solutes present at 1 M and all gases at 1 atm is called the *standard reduction potential*. Now recall that the cell potential is the difference between two electrode potentials (the anode and the cathode):

$$E^0_{cell} = E^0_{red}(\text{cathode}) - E^0_{red}(\text{anode}) \qquad (7.41)$$

Table 7.11 lists a number of standard reduction potentials, all measured relative to the SHE. The redox reactions are spontaneous, in the forward direction as written in this table, if the standard EMF of the cell is positive [E^0_{red}(cathode) > E^0_{red}(anode)]. Comprehensive listings of E^0_{red} values for most half-cell reactions are available in the *Lange's Handbook of Chemistry* and in the *CRC Handbook of Chemistry and Physics*.

Because half-cell potentials are measured relative to SHE, they reflect the spontaneity of redox reactions relative to SHE, with more positive potentials resulting in more potent oxidants (oxidants that want to be reduced) and more negative potentials resulting in more potent reductants (reductants do not want to be reduced; they spontaneously oxidize). Using a table of standard reduction potentials, any species on the left of a given half-reaction will react spontaneously with any species appearing on the right of any half-reaction that appears below it, where reduction potentials are listed from highest and most positive to lowest and most negative (see Figure 7.34).

TABLE 7.11 Number of Standard Reduction Potentials All Measured Relative to the Standard Hydrogen Electrode (SHE)

Standard Potential (V)	Reduction Half-Reaction
+2.87	$F_2(g) + 2e^- \rightarrow 2F^-(aq)$
+1.51	$MnO_4^-(aq) + 8H^+(aq) + 5e^- \rightarrow Mn^{2+}(aq) + 4H_2O(l)$
+1.36	$Cl_2(g) + 2e^- \rightarrow 2Cl^-(aq)$
+1.33	$Cr_2O_7^{2-}(aq) + 14H^+(aq) + 6e^- \rightarrow 2Cr^{3+}(aq) + 7H_2O(l)$
+1.23	$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$
+1.06	$Br_2(l) + 2e^- \rightarrow 2Br^-(aq)$
+0.96	$NO_3^-(aq) + 4H^+(aq) + 3e^- \rightarrow NO(g) + H_2O(l)$
+0.80	$Ag^+(aq) + e^- \rightarrow Ag(s)$
+0.77	$Fe^{3+}(aq) + e^- \rightarrow Fe^{2+}(aq)$
+0.68	$O_2(g) + 2H^+(aq) + 2e^- \rightarrow H_2O_2(aq)$
+0.59	$MnO_4^-(aq) + 2H_2O(l) + 3e^- \rightarrow MnO_2(s) + 4OH^-(aq)$
+0.54	$I_2(s) + 2e^- \rightarrow 2I^-(aq)$
+0.40	$O_2(g) + 2H_2O(l) + 4e^- \rightarrow 4OH^-(aq)$
+0.34	$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$
0	$\mathbf{2H^+(aq) + 2e^- \rightarrow H_2(g)}$
−0.28	$Ni^{2+}(aq) + 2e^- \rightarrow Ni(s)$
−0.44	$Fe^{2+}(aq) + 2e^- \rightarrow Fe(s)$
−0.76	$Zn^{2+}(aq) + 2e^- \rightarrow Zn(s)$
−0.83	$2H_2O(l) + 2e^- \rightarrow H_2(g) + 2OH^-(aq)$
−1.66	$Al^{3+}(aq) + 3e^- \rightarrow Al(s)$
−2.71	$Na^+(aq) + e^- \rightarrow Na(s)$
−3.05	$Li^+(aq) + e^- \rightarrow Li(s)$

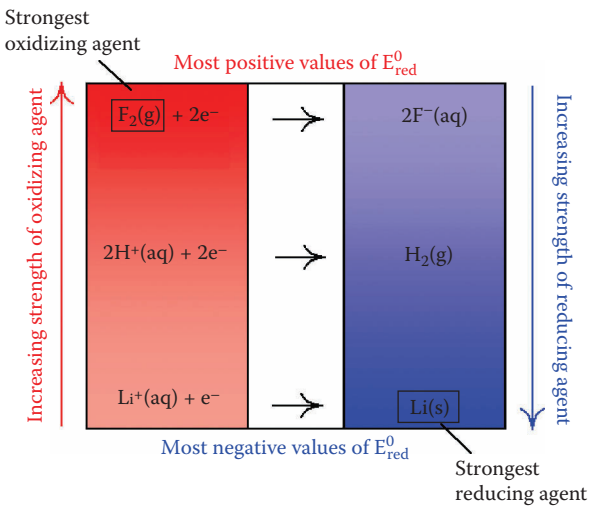


FIGURE 7.34 In a table of standard reduction potentials, any species on the left of a given half-reaction will react spontaneously with any species appearing on the right of any half-reaction that appears below it when reduction potentials are listed from highest and most positive to lowest and most negative.

Half-cell reactions are reversible, i.e., depending on the experimental conditions, any half reaction can be either an anode or a cathode reaction. Oxidation potentials can be obtained from reduction potentials by changing the sign, so Equation 7.41 can be rewritten as:

$$E^0_{cell} = E^0_{red}(\text{cathode}) + E^0_{ox}(\text{anode}) \qquad (7.42)$$

The standard reduction potential is an intensive property, and changing a stoichiometry coefficient of a reaction does not change the reduction potential [e.g., both half-cells $Zn^{2+}(aq, 1\ M) + 2e^- \rightarrow Zn(s)$ and $2Zn^{2+}(aq, 1\ M) + 4e^- \rightarrow 2Zn(s)$ come with a standard reduction potential of $E^0 = -0.76$ V]. This intensive property is also not affected by either the size of the electrode or by the volume of electrolyte, but the concentration of Zn ions must be 1 M.

The potential difference at any concentration is called the *electromotive force* (EMF = cell voltage = cell potential = E_{cell}). This electromotive force depends on:

- The reactions that occur on the electrodes
- The concentrations of reactants and products (see Nernst equation further below)
- The temperature (assumed to be 25°C unless otherwise stated)

A shorthand electrochemical cell notation has been developed for convenience:

- The anode is always on the left and the cathode on the right.
- All redox forms of reagents present should be listed with phase and concentration in parentheses, $ZnSO_4$ (*aq*, 1 M).
- Single lines represent a change of phase: single vertical line "|" (e.g., *s* to *l* or *l* to *g*).
- A comma should be used to separate two components in the same phase.
- Double lines represent a salt bridge or membrane: double vertical line "||".

Example 7.3:

- Fe(s) | Fe^{+2} (1M) || H^+ (1M) | H_2(g) | Pt (1M)
- Cd^{+2} (1M) | Cd(s) || Ag^+ (1M) | AgCl(s) | Ag(s)

Exercise 7.1: A spontaneous redox reaction produces energy that can do electrical work as in the Daniell cell (1836) with a zinc electrode (in 1 M Zn^{2+}) and a copper electrode (in 1 M Cu^{2+}) as shown in Figure 7.35. Calculate the cell EMF (a), and represent this cell in shorthand notation (b).

From Table 7.11: (a) $Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$ $E^0_{red} = +0.34$, and from the same table $Zn^{2+}(aq) + 2e^- \rightarrow Zn(s)$ $E^0_{red} = -0.76$ V; the latter reaction will spontaneously run to the left or $Zn(s) - 2e^- \rightarrow Zn^{2+}(aq)$ $E^0_{ox} = +0.76$ V, so the overall reaction (using Equation 7.42):

$$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s) \quad E^0_{red} = +0.34$$

$$Zn(s) - 2e^- \rightarrow Zn^{2+}(aq) \quad E^0_{ox} = +0.76\ V$$

$$Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s)\ EMF = +1.1\ V$$

(b) The Daniell cell in shorthand notation:

$$Zn(s)\ |\ Zn^{2+}\ (1M)\ ||\ Cu^{2+}\ (1M)\ |\ Cu(s)$$

With a measured EMF of = 1.10 V at 25°C.

Exercise 7.2: Is the following redox reaction spontaneous?

$$Mg^{2+} + 2Ag \rightarrow Mg + 2Ag^+$$

Given that:

$$Ag^+ + e^- = Ag \quad E^0_{red} = +0.80\ V$$

$$Mg^{2+} + 2e^- = Mg \quad E^0_{red} = -2.37\ V$$

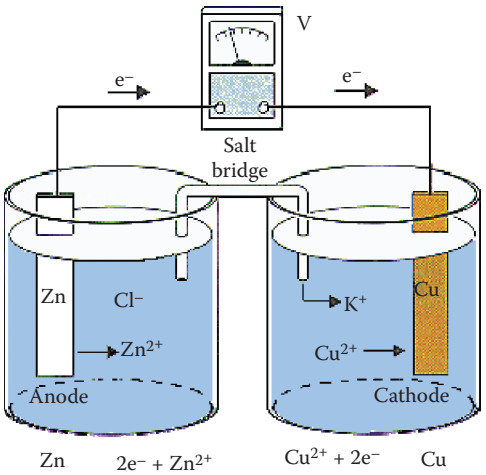


FIGURE 7.35 The Daniell cell.

Summing the two half-cells to obtain the above total cell reaction, we obtain (using Equation 7.42):

$$2Ag - 2e^- = 2Ag^+ \quad E^0_{ox} = -0.80\ V$$

$$Mg^{2+} + 2e^- = Mg \quad E^0_{red} = -2.37\ V$$

$$Mg^{2+} + 2Ag \rightarrow Mg + 2Ag^+ \quad EMF = -3.17\ V$$

This is not a spontaneous reaction because the EMF of the cell is negative.

The above examples all concern Faradaic processes in which electrons cross the solid/electrolyte interface. In a *non-Faradaic* process, there are no charge transfer reactions across the interface because they are thermodynamically and kinetically unfavorable. Under these conditions, processes such as adsorption and desorption can still take place and modify the structure of the electrode/solution interface. During the charging of the double layer, a transient external capacitive current flows:

$$C_d = \frac{dQ}{dV} = \frac{idt}{dV} \qquad (7.43)$$

Further below, we will see that the value of C_d can be obtained from a cyclic voltammogram by measuring the current, *i*, at a given voltage sweep rate, *dV*/*dt*.

Faradaic and non-Faradaic processes occur simultaneously during electrode reactions. The electrochemical stability window of an electrochemical cell is defined as the potential range over which an electrode in a solvent with an inert/indifferent electrolyte (salt) can be polarized without substantial Faradaic currents. These stability ranges depend on the nature of the solvent, the salt, the electrode, and the presence of contaminants. Nonaqueous electrochemistry often provides stability windows as wide as 4–5 V, whereas in aqueous electrochemistry the hydrolysis of water limits the range typically to 1.5 V. Thermodynamically, the voltage difference between hydrogen and oxygen evolution at ideally reversible electrodes is expected at 1.23 V.

Faraday's Law

Electron transfer causes oxidation (electrons jump from a dissolved species onto the solid) or reduction (electrons transfer from the solid onto a species in solution) to occur at the interface. This process is governed by Faraday's law, which dictates the amount of chemical produced (in moles) by the amount of charge passed through the interface:

$$\text{moles} = \frac{Q}{nF} = \frac{i \cdot t}{nF} \qquad (7.44)$$

where i = current passed (in amperes, A)
t = time of current flow (in seconds, s)
n = number of electrons involved in the reaction
F = Faraday (= 96,500 C mol^{-1})
Q = charge (in coulombs, C)

The current, *i*, multiplied by the time the current flows, *t*, gives us the total number of coulombs ($Q = i \cdot t$). The charge on 1 mol of electrons is called the *Faraday* (*F*), i.e., 1.6×10^{-19} C multiplied by the number of Avogadro (N_A) (eN_A) 6.02×10^{23} or 96,500 C/mol.

Thermodynamic Interpretation of the Electromotive Force

Starting with some standard thermodynamic relations, we find from dU = dQ + dW (U = internal energy, Q = heat, and W = work) and dQ = TdS that $dU = TdS - PdV + dW_{electrical}$ (*S* is entropy, –PdV is mechanical work, and $+dW_{electrical}$ is electrical work). Because $dG_T = dH_T - TdS$ (*H* is enthalpy) and $dG_T = dU_{T,P} + PdV - TdS$ (*U* is internal energy, *P* pressure, and *V* volume); therefore, $dG_T = TdS - PdV + dW_{electrical} + PdV - TdS$, and we arrive at the important relationship between Gibbs free energy and work done in an electrical system (most electrochemical cells work at constant *P* and *T*), i.e., the so-called *Gibbs function*:

$$dG_{T,P} = dW_{electrical} \qquad (7.45)$$

The Gibbs function is at the heart of electrochemistry because it identifies the amount of work we can electrically extract from a system. Now the maximum electrical work done by an electrochemical cell equals the product of the charge flowing and the potential difference across which it flows or $W_{electrical} = EQ$. The charge, $Q = n \cdot N_A \cdot e$ where *n* is the number of moles of electrons transferred per mole of reaction, N_A is Avogadro's number (6.02×10^{23}), and *e* is the charge on an electron (-1.6×10^{-19} C). As we just calculated above, $N_A \cdot e = F$ (1 Faraday

is 96,485 C/mol), and we obtain $W_{electrical} = nFE$. By convention, we identify work, which is negative with work being done by the system on the surroundings. Negative free energy change is identified as a spontaneous process. Therefore, we must choose a convention for the sign of the cell potential consistent with these equations, the convention regarding reaction spontaneity, and the sign for ΔG. If we choose to associate a positive potential with a spontaneous process, then we need a negative sign in the equation relating ΔG to E:

$$\Delta G_{T,P} = -nFE \tag{7.46}$$

Note how a measurement of a cell potential directly calculates the Gibbs free energy change for the process. With ΔG negative and E positive, the reaction is spontaneous, and one defines a voltaic (galvanic) electrochemical cell as a cell that generates an electrical current (e.g., battery or fuel cell). With ΔG positive and E negative, the reaction is nonspontaneous, as encountered in an electrolytic electrochemical cell, and it occurs as a result of passage of current from an external power source (e.g., electrolysis of water and the charging of a battery). A galvanic cell is one that is allowed to operate in its natural direction. An electrolytic cell is one operated in reverse. A galvanic cell produces a potential to drive an external load. An electrolytic cell is one that consumes external electricity. Note that the electrode polarity changes between galvanic and electrolytic. The nomenclature of the electrode also changes. However, the definition for reduction and oxidation is constant: Oxidation Is Loss of electrons and Reduction Is Gain of electrons, or "OIL RIG."

All thermodynamic measurements are based on differences between states; there is no absolute value for any property (with the exception of entropy). To quantify thermodynamics, we choose by convention a reference state called the *standard ambient temperature*, pressure, and concentration state [temperature = 298 K (25°C), pressure = 1 bar (10^5 Pa), concentration = 1 molal (mol of solute/kg of solvent)]. The use of atmospheres as a unit of pressure is widespread, and it is so close to the value of 1 bar (1.0134 bar = 1 atm) that the reference state often used is that of 1 atm. The differences are seldom significant. Concentrations are more reliably measured in molality (amount/mass of solvent) rather than molarity (amount/volume of solvent). However, for the following reasons: 1) volume is easier to measure than mass, and 2) density of water (the most common solvent) is close to 1, the most commonly used reference state for concentration is that of 1 M (mol/L). Reference states are indicated by the superscript °, e.g., C° or P°.

The propensity for a given material to contribute to a reaction is measured by its activity, a. This describes how "active" this substance is in the reaction compared with how it would behave if it were present in its standard state. Activity scales with concentration or partial pressure:

$$a \propto C/C^\circ \text{ or } a \propto P/P^\circ \tag{7.47}$$

Generally speaking, a substance's activity increases as the amount of that substance available for reaction increases. However, because of intermolecular interactions (which increase as a substance's concentration or partial pressure increases), there are deviations from a direct correspondence as pressure and concentration increase. These differences can become very significant at very high concentrations. We then define activity as:

$$a = \gamma C/C^\circ \text{ or } a = \gamma P/P^\circ \tag{7.48}$$

Activity coefficients γ are close to 1 for dilute solutions and for low partial pressures, but they change with concentration, temperature, other species, and so on. Generally, we ignore activity coefficients for educational simplicity, but careful work requires its consideration. In the case of pressure, the coefficient used is often called *fugacity*, but it is the same thing and behaves in the same way. Activity *a* is unitless, and because activity coefficients are close to 1 for dilute solutions, the reference states for partial pressure and concentration have a numerical value of 1. Therefore, we often approximate activity by concentration (M) or partial pressure (atm). The reference state for a solid material is itself, and the same is true for a pure liquid. A solvent is usually present in huge excess (compared with the solutes); therefore, it is mostly like a pure liquid. In all these cases, $a = 1$. In other words, if we increase the amount of a pure solid, a pure liquid, or a solvent, the reaction goes on for a longer time (but it does not go any faster).

If you increase the amount of these materials in a system, it will not increase the reaction rate, but it may increase the duration of a reaction. The point is that the "activity" of the system does not change with amount of these materials, so they are always referenced to themselves.

The Nernst Equation

We now want to know what happens with the cell potential if the concentrations and pressures are not in the standard state. We first derive how the chemical potential, μ, changes with activity by integrating the expressions for the dependence of amount of material on the Gibbs function, leading to the following relationship:

$$\mu = \mu^0 + RT \ln a \tag{7.49}$$

When analyzing a chemical reaction such as aA + bB → cC + dD mathematically, we form the following reaction quotient:

$$Q = \frac{a_C^c a_D^d}{a_A^a a_B^b} \tag{7.50}$$

where products appear in the numerator and reactants in the denominator. It requires the activity of each reaction participant, where each term is raised to the power of its stoichiometric coefficient. Terms in the reaction quotient that arise from the participation of pure solids, pure liquids, or solvents are left out as they all equal to 1. If we ignore activity coefficients, then activities for solutes and gases are numerically equal to their concentration or partial pressures (because the activity is the ratio of these terms with the number 1). One often writes the reaction coefficient Q with concentration and pressure shown explicitly, but remember that we are really dealing with activities. The reaction quotient is unitless because it has only activity in its expression, and all activities are unitless. The Gibbs free energy has the same activity dependence as the chemical potential:

$$G = G^0 + RT \ln a \tag{7.51}$$

When we apply this to a reaction, the reaction quotient comes into play, which gives us:

$$\Delta G = \Delta G^0 + RT \ln Q \tag{7.52}$$

Q contains all the activities of both reactants and products, all raised to their proper stoichiometric coefficients. When all participants have unit activity (a = 1), then Q = 1 and ln Q = 0, or:

$$\Delta G = \Delta G^0 \tag{7.53}$$

The reaction proceeds and Q changes until finally ΔG = 0. At that point, the reaction stops as equilibrium is reached with the reaction quotient Q = Q*. The reaction does not stop at Q^* but simply has reached the situation where the forward and reverse reaction rates are equal. Chemical activity is still ongoing with just as much vigor as before, but it is happening equally in both directions—products are reacting to turn into reactants just as fast as reactants are reacting to turn into products. This is called a *dynamic equilibrium*, rather than a *static equilibrium*.

$$\begin{aligned} 0 &= \Delta G^0 + RT \ln Q^* \\ \Delta G^0 &= -RT \ln Q^* \\ \text{and } Q^* &= K_{eq} \end{aligned} \tag{7.54}$$

This special Q^* is renamed K_{eq}, the equilibrium constant. From the expression for the Gibbs free energy as a function of activity (Equation 7.51), we can derive an expression in terms of the cell potential by using the relation between the free energy and the cell potential E (Equation 7.46):

$$\Delta G = \Delta G^0 + RT \ln Q = -nFE = -nFE^0 + RT \ln Q \tag{7.55}$$

where n is the number of electrons transferred in the process. By rearranging, we obtain the very important Nernst equation:

$$E = E^0 - \frac{RT}{nF} \ln Q = E^0 - \frac{0.0592}{n} \log Q \tag{7.56}$$

This expression relates the dependence of the cell potential E on the reaction quotient Q. The logarithmic term tells us that if Q changes by a factor of 10, this will change the cell potential by 59.2 mV, or about 60 mV for a one-electron process (n = 1). This "rule of thumb" relationship will show up often in sensors, batteries, fuel cells, and so on.

The above equations allow us to relate measured cell potentials to standard Gibbs free energies of reaction. These, in turn, are related to a reaction's equilibrium constant (see Equation 7.54).

Example 7.4:

Consider the Daniell cell of Figure 7.35 again. What is the potential of the cell if $[Cu^{2+}] = 0.01$ M and $[Zn^{2+}] = 1.00$ M? (We will assume that the activity coefficients are = 1.)

$$E = 1.10 - \frac{0.0592}{n} \log\frac{1.00}{0.01} = 1.10 - 0.030 \log(100)$$

$$= 1.041\,V$$

Note that the cell potential is decreased by about 60 mV from the standard cell potential. This accounts for a decrease in Cu^{2+} concentration of two orders of magnitude, where our rule of thumb predicted 120 mV; however, because the reaction involves two electrons (n = 2), we obtain a 60-mV change in potential only.

Example 7.5:

The Nernst equation demonstrates that potential depends on concentration. A cell made of the same materials, but with different concentrations, will also produce a potential difference, e.g.:

$$Cu \mid Cu^{2+}\,(0.001M) \parallel Cu^{2+}\,(1.00M) \mid Cu$$

What is standard cell potential E^0 for this cell? What is the cell potential E? What is n, the number of electrons transferred? Which electrode, anode, or cathode will be in the numerator?

$$E = E^0 - \frac{0.0592}{n} \log\left(\frac{[Cu^{2+}]_{anode}}{[Cu^{2+}]_{cathode}}\right)$$

$$= 0 - \frac{0.0592}{2} \log\frac{0.001}{1.00} = +0.089\,V$$

The standard cell potential in a concentration cell clearly must be 0 for the two electrodes, because at standard conditions both are identical. The number of electrons transferred is two. Remember that we are calculating the cell potential for the cell as written in the definition. The right side is the cathode and thus refers to a reduction process. The other is for oxidation. If we write those two reactions, we see that the aqueous solution from the cathode appears as a reactant. Hence the cathode aqueous solution is in the denominator, and conversely the anode side is in the numerator.

Example 7.6:

Consider the cell:

$$Pt \mid I^-\,(1.00M),\ I_2\,(1.00M) \parallel Fe^{2+}\,(1.00M),\ Fe^{3+}\,(1.00M) \mid Pt$$

with a standard cell potential given by = 0.771 V – 0.536 V = +0.235 V.

$$\Delta G^0 = -nFE^0 = -2(96,485\ C/mol)(0.235\ J)$$

$$= -45,348\ J/mol$$

This is the free energy change leading to the equilibrium constant, K_{eq}, for the reaction as follows:

$$\ln K_{eq} = -\frac{\Delta G^0}{RT} = -\frac{(-45,348\ J/mol)}{(8.314\ J/Kmol)(298\ K)} = 18.3034$$

$$K_{eq} = e^{18.3034} = 8,8910^7$$

Exercise 7.3: A metal in contact with a sparingly soluble salt of the metal, we will see below, is called an *electrode of the second kind*. Another common name for such an electrode is an anion electrode. Examples include reference electrodes (REF Es), such as Ag/AgCl(s) and $Hg/Hg_2Cl_2(s)/Cl^-$ [saturated calomel electrode (SCE)]. These reference electrodes are often more convenient to use than the above-reviewed SHE. Consider the Ag/AgCl electrode shown in Figure 7.36. This REF E is made from an Ag wire coated with AgCl(s) immersed in an NaCl or KCl solution.

The reactions involved are $AgCl \leftrightarrow Ag^+ + Cl^-$ and $Ag^+ + e^- \leftrightarrow Ag(s)$ [$E^0_{red} = +0.80$ V]. Derive the Nernst equation for this electrode.

$$Ag^+ + e^- \leftrightarrow Ag(s) \quad E^0_{red} = +0.80\ V$$

Chemical reaction;

$$AgCl \leftrightarrow Ag^+ + Cl^- \text{ no electrons involved}$$

$$AgCl + e^- \leftrightarrow Ag(s) + Cl^-$$

$$E = E^0 - \frac{0.0592}{1} \log\frac{a_{Ag}a_{Cl^-}}{a_{AgCl}} \text{ but } a_{Ag} \text{ and } a_{AgCl} = 1$$

$$\text{or } E = 0.80 - 0.0592 \log a_{Cl^-}$$

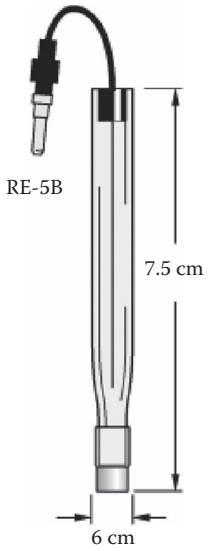


FIGURE 7.36 A typical commercial Ag/AgCl reference electrode. (From http://www.bioanalytical.com.)

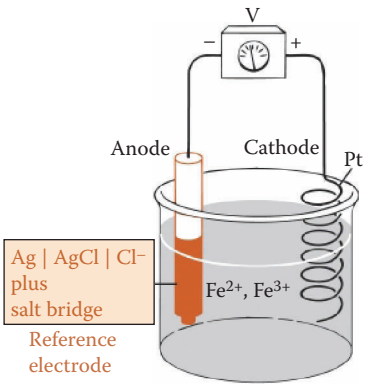


FIGURE 7.37 Potentiometric sensor in a two-electrode cell. The entire left half-cell is made into a self-contained reference electrode containing appropriate solutions and a salt bridge. The chloride concentration in the reference electrode is held constant by the fixed solubility of KCl.

From this expression, Cl^- (an anion) activity determines the potential of this electrode, and the electrode can be used as a Cl^- sensor (60-mV voltage change per decade of Cl^- concentration change). The purpose of a reference electrode is to provide a stable potential, independent of solution composition, against which other potentials can be reliably measured. Filling a glass envelope, which is equipped with a small pinhole to make ionic contact with the outside solution, with a saturated Cl^- solution that surrounds a Ag/AgCl wire produces a reliable reference electrode. The constant a_{Cl^-} in the above equation provides for a constant E (E = 0.22 V vs. SHE at 25°C). Only when enough solvent or ions from the outside solution enter the glass envelope does the reference potential start drifting. It is obvious that the larger the REF E electrolyte reservoir, the longer the reference electrode will last without need to refill the reservoir with fresh saturated Cl^- solution. This explains why miniaturization of a reference electrode is such a challenge; smaller means less electrolyte capacity and thus a shorter lifetime.

Two-Electrode and Three-Electrode Cell Configurations

There are two common electrode configurations used in electrochemistry: two-electrode and three-electrode cells. In a two-electrode configuration, an electrochemical cell is equipped with one working electrode (WE) and one reference electrode (REF E) (Figure 7.37). The electrode at which the reaction of interest occurs is called the *working electrode*. An *amperometric measurement* with such a cell occurs when current is passed between the WE and REF E by applying a potential between the two. The kinetics of the electrochemical reaction are driven by the applied potential, and we measure the diffusion-controlled current flowing across the electrode/solution interface. This current is directly proportional to the bulk concentration of the analyte present in the solution (see Equation 7.124). Passage of a significant current through a reference electrode shifts the electrode away from equilibrium and may even destroy its materials (e.g., in the case of Ag/AgCl, the AgCl could all be reduced to Ag). Moreover, a significant resistance decrease (iR) results for less conductive media (in a salt water solution, this is not a problem since R is typically 100 Ω, so with $i < 10$ μA, $iR < 1$ mV) or when the WE is very large. As a consequence, a two-electrode cell should only be used when the iR drop is less than 1 mV or when using *micro-* or *ultramicroelectrodes*. In *potentiometric measurements*, only very small currents are drawn, and a two-electrode system can always be used confidently in this case (as long as a high-impedance voltmeter is used so that a measurement of the potential difference will be accurate). In potentiometry, the electron transfer reaction is kinetically facile, i.e.,

the equilibrium electron exchange current $i_{0,e}$ (on an electron conductor) or ion exchange current $i_{0,i}$ (on an ion conductor) is large. The cell potential responds to changes in the activity of the analyte species present in the solution in a well-defined manner described by the Nernst equation. The cell potential varies in a linear manner with the logarithm of the analyte activity.

Microelectrochemical sensors are often based on a two-electrode system with a WE and a pseudoreference electrode because a real reference electrode is too complicated and expensive to implement. This is possible because the small working electrode draws so little current that a larger pseudoreference electrode (i.e., a Ag/AgCl film) is not shifted from equilibrium enough to cause large errors.

In a three-electrode electrochemical cell (Figure 7.38) one deals with a WE, a REF E, and a counter electrode (CE) or auxiliary electrode. A Luggin capillary is often used to position the sensing point of a reference electrode at a desired point in a cell close to the working electrode.

With three-electrode systems, a *potentiostat* is used. A potentiostat is an instrument that controls the voltage difference between a WE and a REF E (Figure 7.39). The potentiostat implements this control by injecting current into the cell through the auxiliary electrode or CE. In almost all applications, the potentiostat measures the current flow between

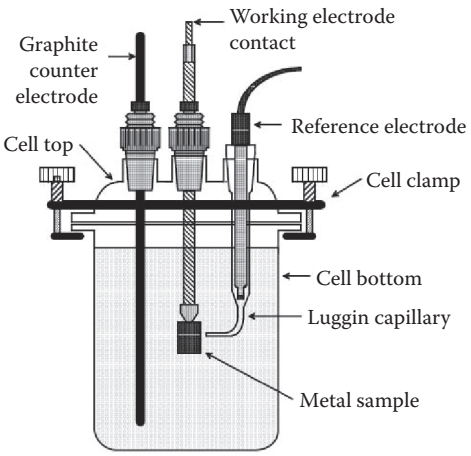


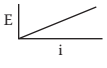
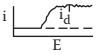
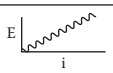
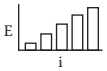
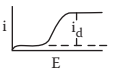
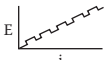
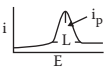
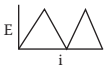
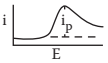
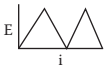
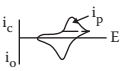
FIGURE 7.38 In this three-electrode cell, the counter electrode (CE) is a graphite rod, the working electrode (WE) is located in the center, and the reference electrode (REF E) is brought close to the working electrode surface via a Luggin capillary.

FIGURE 7.39 Solartron 1286 potentiostat and 1250 frequency response analyzer. (See also http://www.bioanalytical.com.)

the working and auxiliary electrodes. A high impedance is placed in front of the reference electrode to limit the current so that the reference potential remains constant. A cell like this avoids the internal polarization of the reference electrode and its potential degradation. The system also compensates for a major portion of the cell *iR* drop.

A potentiostat can be converted to work in *galvanostat* mode. Whereas in potentiostat mode one controls the potential and measures the ensuing current, in galvanostat mode one controls the current and measures the ensuing voltage. *Voltammetry* is an electroanalytical method in which the potential of an indicator electrode is varied in a definite manner with time, and the current flowing through the indicator electrode is measured. Voltammetric techniques are classified according to the type of voltage perturbation applied to the working or indicator electrodes, i.e., the way in which the voltage signal input varies with time. The form of the input $V(t)$ function determines the form of the resulting current response (see some examples in Table 7.12). The current/potential response curve is called a *voltammogram*. The voltammetry method relies on the fact that the current measured reflects rate-determining diffusion of the analyte species from the bulk solution to the surface of the indicator electrode, where it is readily oxidized or reduced. The solution contains supporting electrolyte, i.e., an excess of nonreactive electrolyte (e.g., KCl or NaCl), to conduct current besides the redox couple. This ensures that the redox species reaches the electrode by diffusion but not by migration. Under such conditions of diffusion control, the measured current is linearly proportional

TABLE 7.12 Voltammetry Examples

Name of Techniques	Potential Excitation Signal		Mass Transfer	Measurement	Analytical Relation to Bulk Concentration	Typical Display
Polarography (DC or normal)	Slow linear scan (or constant E)		Diffusion	i vs. E	$i_d \propto C$	
AC Polarograph	Slow linear scan + low amplitude sine wave		Diffusion	i_{ac} vs. E	$i_a \propto C$	
Pulse polarography	Square voltage pulses of increasing amplitude		Diffusion	i vs. E	$i_d \propto C$	
Differential pulse polarography	Square voltage pulses of constant amplitude + linear ramp		Diffusion	Δi vs. E	$i_p \propto C$	
Single-sweep voltammetry	Linear scan E		Diffusion	i vs. E	$i_p \propto C$	
Cyclic voltammetry	Triangular scan E		Diffusion	i vs. E	$i_p \propto C$	

to the bulk concentration of the analyte species (see Equation 7.124).

A most important example of voltammetry is cyclic voltammetry, an electrochemical method in which information about an analyte is obtained by measuring current (i) as a function of applied potential using a triangular waveform for excitation (Figure 7.40). We will detail the features of a cyclic voltammogram after we elucidate the difference between kinetic and diffusion-controlled electrochemical reactions.

In Figure 7.41 we compare two- and three-electrode systems.

Types of Electrodes

Introduction

We distinguish electrodes based on the nature and number of phases between which electron transfer occurs and/or based on the nature of the current-carrying species within them. Based on the nature and number of phases, one typically recognizes electrodes of the first kind and electrodes of the second kind. Based on the charge carrier in the solid, we can also distinguish between metals, semiconductors, insulators, solid electrolytes, and mixed conductors. As summarized in Table 7.10, different types of mobile charge carriers characterize the different types of electrode materials. Metal electrodes (e.g., Pt) feature electrons as the mobile charge carriers, whereas semiconductors (e.g., n-Si) have electrons and/or holes that play that role. Solid electrolytes

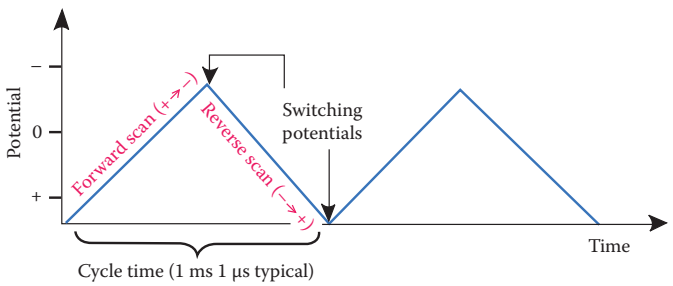
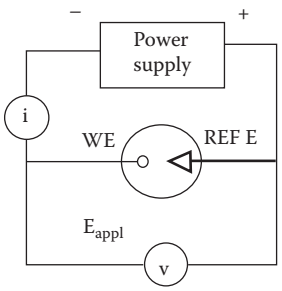
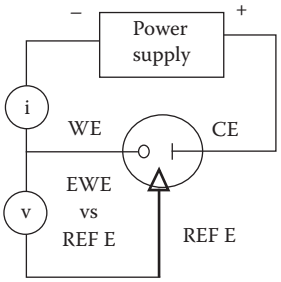


FIGURE 7.40 Triangular waveform used to register a cyclic voltammogram.



Two-electrode cell: Current passes through reference (can cause iR drop and degradation of the reference)

R in aqueous solutions is ~100 Ω so when i < 10 μA, iR < 1 mV



Three-electrode cell: Current is passed through the counter (auxiliary) electrode

Useful for high-resistance (nonaqueous) solutions

FIGURE 7.41 Two-electrode versus three-electrode system.

(e.g., LaF_3) have mobile ions, whereas insulators (e.g., SiO_2) possess no free charge carriers. Mixed conductors (e.g., IrO_x) have both ions and electrons for conduction.

We treat ion-selective electrodes (ISEs), also called *membrane electrodes*, as a separate category of electrodes. Membrane electrodes include pH glass membrane electrodes, crystalline solid electrolyte electrodes, polymeric electrodes, and membrane-based gas-sensing probes.

Electrodes of the First and Second Kind

Electrodes of the First Kind A metal in contact with its cations or a metal in contact with a redox couple in solution is called an *electrode of the first kind*. Examples include $Cu^{2+}/Cu(s)$, $Zn^{2+}/Zn(s)$, the standard hydrogen electrode (SHE) (i.e., Pt black in contact with H_2/H^+), Ag^+/Ag [nonaqueous reference electrode (REF E)], and $Cl^-/Cl_2(g)/Pt$. A working electrode (WE) (e.g., Pt) that changes potential against a suitable reference electrode and senses the redox couple concentration present in solution is an indicator electrode of the first kind. The electrode response of this type with respect to a REF E is given by the Nernst equation:

$$E = E^0 + \frac{0.0592}{n} \log a_{M^{n+}} \qquad (7.57)$$

Often these sensors respond to cations, so they also may be called *cation electrodes*. With an electrode of the first kind, the fastest electron-exchange reaction, i.e., the reaction with the largest exchange current density $i_{0,e}$ (the current at zero external current and with no net reaction), determines the potential of the electrode. Often there are different redox species involved in establishing the equilibrium potential, in which case we speak about a mixed potential: zero external current but net reaction (e.g., corrosion). The inert species in solution (e.g., 1 M NaCl), also called *indifferent electrolyte*, do not, at zero or low bias, exchange electrons with the platinum electrode. Instead, they provide solution conductivity. All redox couples have a known redox potential as measured against a standard reference electrode (e.g., SHE). Platinum is the most common metal indicator electrode to measure these redox potentials. Pt is mostly inert, does not participate in reactions, and allows electron transfer to and from solution. Gold is also an inert metal indicator electrode used for the same purpose. Carbon electrodes are often used because many redox reactions are very fast on a carbon surface (see also C-MEMS in Volume III, Chapter 5). Note that not all metals are electrodes of the first kind; metals such as Fe, Al, and W are not electrodes of the first kind because they have relatively thick surface oxide coatings that are pH-sensitive.

Electrodes of the Second Kind Metals in contact with their sparingly soluble salt are *electrodes of the second kind*, also called *anion electrodes*. Examples include Ag/AgCl(s) and $Hg/Hg_2Cl_2(s)/Cl^-$ (saturated calomel electrode; SCE). As seen from the Nernst expression for an electrode of the second kind (see also Exercise 7.3 on a Ag/AgCl electrode):

$$E = E^0 - \frac{0.0592}{n} \log a_{A^{n-}} \qquad (7.58)$$

In this case, anion activity determines potential. These electrodes make great reference electrodes because of the low solubility of their salt (potential very stable). As said before, a reference electrode is an electrode that does not change its potential with solution composition. By making the anion concentration around the sensor element constant, one obtains such a situation. Fixing the anion concentration is often accomplished by mounting the electrode in a glass envelope and putting a saturated solution of the anion in contact with the sparingly soluble salt of the metal wire, which itself is encapsulated in a very slowly leaking glass envelope. This is illustrated for a Ag/AgCl REF E in Figure 7.42.

The purpose of a reference electrode is to provide a stable potential against which other potentials can be reliably measured. Criteria for the selection of a good reference electrode include stability (time, temperature) and reproducibility. Importantly, the potential should not be altered by passage of a small current, i.e., the electrode should be perfectly nonpolarizable (Figure 7.43). The definition for an ideal nonpolarizable electrode is that there is no charge transferred across the electrode/electrolyte interface (in the case of a Ag/AgCl electrode, all the charge transfer is between the Ag and the

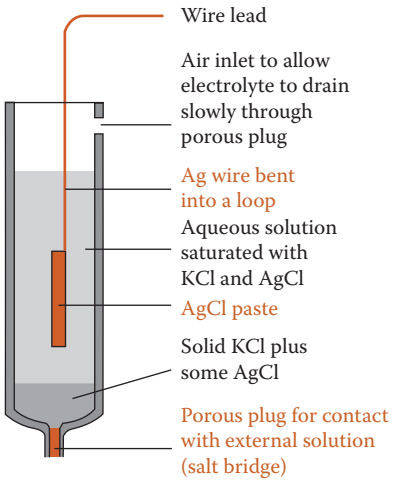
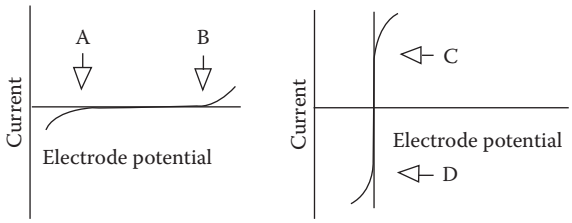


FIGURE 7.42 Construction of a Ag/AgCl reference electrode.

AgCl instead), and ΔV remains = 0 for a wide range of currents. An ideal polarizable electrode (IPE) is one where the charge transfer resistance is very large, so that Δi = 0 for a large range of voltages. The behavior of such an interface when the potential across the electrode is changed resembles that of a plain capacitor. An IPE follows the standard capacitor equation. The reference electrode should also be easily constructed and convenient for use. The standard hydrogen electrode is difficult to use because Pt black is easily poisoned by organics, sulfide, cyanide, and so on. Also, hydrogen is explosive, and sulfuric and hydrochloric acids are very strong acids; however, SHEs feature one of most reproducible potentials (±1 mV). More convenient reference electrodes include the Ag/AgCl and the saturated calomel electrode (Hg/Hg_2Cl_2) (which has problems with mercury toxicity though).



Ideal polarized electrode (INPE; electrode of fixed i)

Δi = 0 over a wide range of V (A to B)

Δi/ΔV = 0

Ideal nonpolarized electrode (INPE; electrode of fixed V reference electrodes)

ΔV = 0 over a wide range of i (C to D)

Δi/ΔV = finite value

FIGURE 7.43 Concept of polarization. A good reference electrode is ideally nonpolarizable.

When the above reference electrodes are used in nonaqueous media, there might be significant junction potentials. Nonaqueous media can also cause greater noise, and electrolyte precipitation may clog the electrode frit. Finally, the chemistry may be water-sensitive.

Electrochemical reference electrodes are very difficult to miniaturize, and when miniaturized they do not last very long because the internal electrolyte quickly mixes with the external solution. The development of miniature chemical and biological sensors has been hampered by this problem. Because of these problems, many researchers have chosen to implement pseudo-reference electrodes in miniature sensors. These pseudo-reference electrodes, for example, a Ag/AgCl wire without the saturated anion solution, only mimick the behavior of a real thermodynamically justified electrode. These problems will be illustrated below when introducing the ISFET.

Earlier, we pointed out that some metals with a relatively thick surface oxide coating may act as electrodes of the second kind, with the oxide coat setting up a pH response. This may be interesting as an alternative for pH glass electrodes. Glass pH electrodes have their shortcomings, such as extreme fragility, relatively slow response times, short lifetime, and very high impedance. Moreover, glass electrodes cannot be used in strong alkaline, HF, or organic solutions and are very difficult to miniaturize. As a consequence, various metal oxides have been tested as alternate, fast, robust, long-lifetime, low-impedance pH sensors that are easier to miniaturize. Most studies have focused on Sb/Sb_2O_3 (widely used for acid-based titrations or in solutions containing HF; however, voltage drift with this electrode is considerable), Bi/Bi_2O_3 (used in alkaline media), Pt/PtO_2, Ir/IrO_x, Ru/RuO_2, Os/OsO_2, Rh/RhO_2, Ti/TiO_2, and Sn/SnO_2. These oxides all represent mixed conductors with both electrons and ions (protons) exchanging rapidly at the interface. Depending on the relative magnitude of $i_{0,e}$ versus $i_{0,i}$, the electrode acts as a redox probe, a pH sensor, or both. For most mixed conductors $i_{0,e} \gg i_{0,i}$, i.e., one expects severe redox interference when used as pH sensors. Of the metal oxides listed, IrO_x holds the most promise as a very fast, stable,

low-impedance pH sensor with surprisingly small redox interference.[8] As an example, we introduce some of our own work on Ir/IrO_x as a solid-state ion-selective electrode (ISE). Whereas this oxide previously had been prepared by electrochemical growth (potential cycling), anodic electrodeposition, AC reactive sputter coating (Ir target and oxygen gas), and thermal oxidation, we showed for the first time that an electrode prepared by melt-oxidation in a lithium carbonate (Li_2CO_3) exhibits a very reproducible, fast pH response (faster than a few tens of seconds) with low redox interference. The Nernstian slope (58 mV/pH) is close to ideal. Other good characteristics of this electrode include stability, even in solutions containing F^-. The ease of miniaturization using integrated circuit (IC) techniques is an additional benefit. The world's smallest CO_2 sensor, based on thin-film IrO_x pH sensing, is shown in Volume II, Figures 4.92–4.97. An example IrO_x-based pH probe (A) and some experimental results (B) are shown in Figure 7.44. From the SEM picture of an electrode cross-section (C), the IrO_x coating on the Ir wire appears as a 20-µm-thick, nonporous film with a columnar grain structure. In Figure 7.50 we use a Ir/IrO_x wire as described here to make a CO_2 gas sensor.

SensIrOx, a private enterprise, demonstrated effective pH measurement in extreme temperature and pressure conditions using IrO_x probes as shown in Figure 7.44 (http://www.sensirox.com). For example, in food production or biological process monitoring, IrO_x probes provide a safe, nonglass solution. They are also well-suited for measurements in high salt solutions, viscous sludges and slurries, or for wide-ranging pH. A pH sensor based on IrO_x, suitable for measuring total acidity number in non-aqueous applications (e.g., engine oil), was also developed by SensIrOx.

Ion-Selective Electrodes

Introduction Ion-selective electrodes (ISEs) are membrane electrodes that exchange ions at a membrane/solution interface. Ideally, these electrodes respond to one ion only, and their equilibrium potential does not involve redox reactions. ISEs use a membrane that exchanges ions fast, where the electric potential across the membrane/liquid interface depends on the concentration of the ion with the fastest (largest) ion exchange current, $i_{0,i}$. The difference in voltage across the membrane is measured between two REF Es: an internal and an external one, as sketched for the case of F^- sensing with a LaF_3 ISE (developed in 1967 by Ross) in Figure 7.45. The ion from solution that exchanges sites within the single-crystal LaF_3 is F^-, which is mobile in the LaF_3 crystal lattice, even at room temperature.

The electrical potential difference across the membrane boundaries of an ISE can be described by:

$$\triangle E = \frac{0.0592}{n} \log \frac{[C^+]_{outer}}{[C^+]_{inner}} \qquad (7.59)$$

where n is the charge on the analyte ion (1 in the case of F^-), $[C^+]_{outer}$ is the ion concentration in the

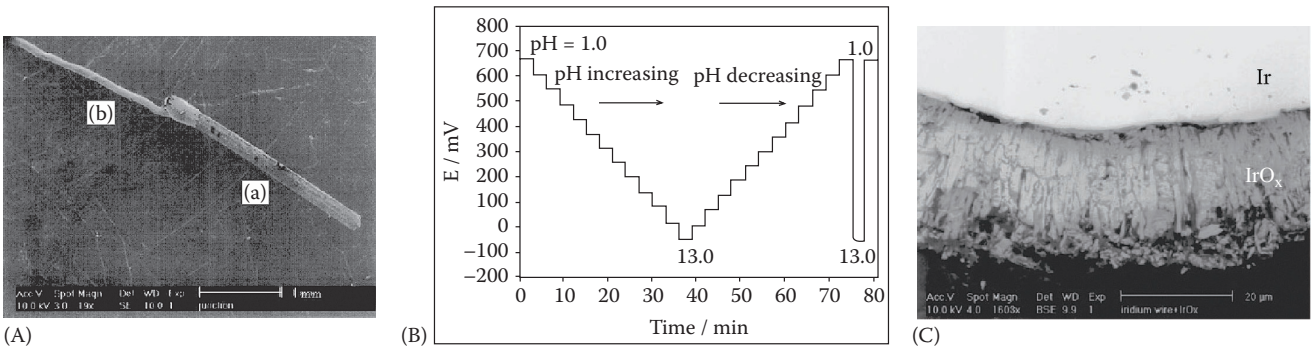


FIGURE 7.44 (A) Solid-state pH sensor: Ir wire with IrO_x coat (a) and Au wire contact (b). (B) A pH titration curve with IrO_x, no hysteresis. (C) SEM of an electrode cross-section. Note this same type of pH sensor was used to construct the CO_2 gas sensor in Figure 7.50.

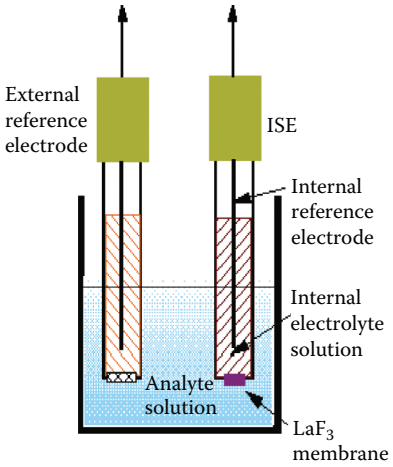


FIGURE 7.45 A F^- ion-selective electrode based on a single crystal of the solid electrolyte LaF_3 as the membrane.

unknown solution, and $[C^+]_{inner}$ is its concentration in the inner solution. Because the $[C^+]_{inner}$ is constant, the above equation can be simplified to:

$$\triangle E = \frac{0.0592}{n} \log\left[C_{outer}\right] \tag{7.60}$$

There is no change in the inner reference solution because there is no contact between the inner reference electrolyte and the unknown sample solution. Contact with the sample solution is established with the outer membrane surface and the external reference electrode, which is in ionic contact via a Luggin capillary. The potential of the whole electrochemical cell (at 25°C) is then given as:

$$E_{cell} = E_{const} + \frac{0.0592}{n} \log\, a_{M^{n+}} \tag{7.61}$$

where n is the valence of the ion that can be – (for anions) or + (for cations) and a is the activity of the ion.

For univalent ions, the slope is 59.16 mV/decade, and for divalent ions, the slope is 29.58 mV/decade. From the slope, it is possible to tell whether the electrode is behaving properly.

The selectivity coefficient describes the relative response of any electrode to interfering species and is defined as:

$$k_{A,X} = \frac{\text{response to interfering X}}{\text{response to principal ion A}} \tag{7.62}$$

The smaller the value of k, the lesser the response to X. For example, for F^- sensing with a LaF_3 sensor in the presence of Cl^-, the selectivity coefficient is $k_{F^-,Cl^-} = 10^{-4}$; in other words, Cl^- ions must be present in a concentration 10,000 times higher than F^- to set up a similar signal strength. At pH > 8, OH^- anions become a major interference for F^- measurement. At low pH, F^- exists only as HF (which this sensor will not pick up). Therefore, the sample must be buffered at a pH between 5 and 7 for reliable F^- analysis.

Some requirements that all membranes in ion-selective electrodes must meet include minimal solubility, a finite electrical conductivity, and selectivity for one ion only. The solubility of an ISE electrode in analyte solutions typically approaches zero. In general, the finite conduction of a membrane takes the form of migration of singly charged ions within the membrane (e.g., F^- in LaF_3 and Na^+ and/or Li^+ in a glass electrode). Finally, a membrane or some species contained within the membrane matrix must be capable of selectively exchanging ions with analyte ions.

The types of membrane electrodes include glass electrodes, crystalline solid-state electrodes, polymer electrodes, and gas-sensing probes.

The selectivity of a membrane electrode is linked with the uniqueness of the binding site for the intended ion from solution. Ideally, there would be only one type of binding site in the membrane with exclusive selectivity for the intended ion. In practice, surfaces are very complex, often involving fractals as encountered on the beach, trees, snowflakes, and so on, rather than atomically sharp transitions (see Figure 7.13). This implies that a binding site at the surface might find itself in a variety of energetic configurations with different associated binding constants/selectivities for the analyte. As a consequence, absolute selectivity with one sensor material cannot be achieved. In Volume III, Chapter 10 on MEMS and NEMS, we illustrate how using an array of sensors in an electronic tongue or in an electronic nose might present a possible alternative to more selective sensing.

Glass Electrodes A traditional pH (pondus Hydrogenii; literally hydrogen exponent)

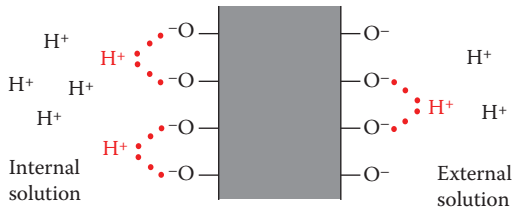


FIGURE 7.46 The Donnan potential on a thin pH-sensitive glass membrane is established on both sides of the membrane.

measurement with a glass electrode qualifies as a potentiometric ion-selective electrode (ISE) (e.g., a thin glass layer with a composition of 22% Na_2O, 6% CaO, and 72% SiO_2), and it is the most common solid-state ISE. The H^+ ion is the only one that significantly exchanges with the outer hydrated glass layer (see Figure 7.46). The proton exchange at the glass/solution interface results in a charge buildup, establishing the so-called *Donnan potential*. The Donnan potential is also established at the glass membrane/internal solution interface and is kept constant by the fixed pH of the internal electrolyte (inner reference), whereas the Donnan potential on the other side of the membrane is determined by the analyte solution pH (Figure 7.46).

Na^+ or Li^+ ions that are present in the glass carry the electrical current by migrating from one side to the other. A combination electrode used for pH measurements utilizes a combination of a silver reference electrode and a glass electrode as shown in Figure 7.47.

The pH response of a glass electrode, based on Equation 7.61 and the fact that pH = –log $[H^+]$, is given by:

$$E_{cell} = E_{const} = 0.0592 \text{ pH} \qquad (7.63)$$

A glass electrode is calibrated with several buffers close to the pH of the unknown solution. It is also important that the glass electrode be kept hydrated when not in use. If not, the hydrated gel layer of the glass dries out, and the electrode has to be reconditioned/rehydrated for several hours before use.

The selectivity of a glass electrode strongly depends on the glass composition, with a Li_2O-containing glass having a selectivity coefficient k_{H^+,Na^+} of $\approx 10^{-11} - 10^{-13}$ and a Na_2O-containing glass having a k_{H^+,Na^+} of $\approx 10^{-10}$. An "alkaline error" can result when cations other than H^+ are present in large excess in the solution. These cations can exchange for H^+ in the outer glass gel layer. However, the glass can be carefully selected to reduce the effect of interfering ions. On the other hand, by adding alumina to the glass composition, selectivity can be changed for the determination of other univalent ions such as Na^+.

The resistance (impedance) of a glass pH sensor is very high (100 MΩ to more than 1000 MΩ), which requires the input amplifier of the pH meter to be very high; the input impedance of the meter must be at least 100 times greater than that of the sensor (high-input impedance amplifier with field effect transistor input stage). However, very high

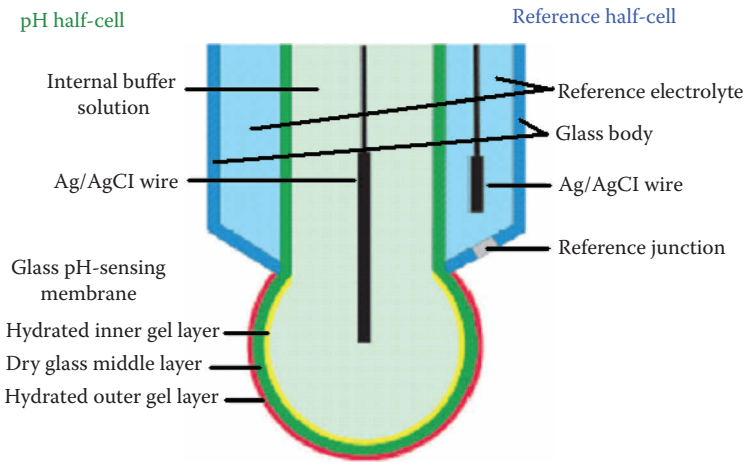


FIGURE 7.47 A combination pH sensor.

impedance can make the measurement noisy, and the smaller the sensor, the bigger this problem becomes. We will come back to this aspect of miniaturized electrochemical sensors when covering ISFETs. We saw earlier how glass pH electrodes have many shortcomings, such as extreme fragility, relatively slow response time, short lifetime, and very high impedance. We have also shown how IrO_x holds promise as a very good alternative, as a stable, low-impedance pH sensor with small redox interference (although hydrogen is detrimental).

Crystalline Solid-State ISEs Solid-state ISEs are based on inorganic crystals such as single-crystal LaF_3, as we explored earlier for F^- sensing. Doping of the LaF_3 with EuF_2 creates anion vacancies in the crystal, allowing F^- anions to jump back and forth between the solution and the vacancies in the solid. This leads to a dynamic equilibrium that establishes a Donnan potential at the crystal/solution interface (F^- has a larger ion exchange current $i_{0,i}$ than any other ion at the LaF_3/electrolyte interface). The F^- sensor cell in Figure 7.45 can be annotated as:

Ag(s) | AgCl(s) | NaCl, NaF(aq) | LaF_3 crystal | ext. soln || saturated KCl and AgCl(aq) | AgCl(s) | Ag(s)

The response of this electrode is given by:

$$E_{cell} = E_{const} - 0.0592 \log[F^-] = E_{const} + 0.0592\, pF^- \quad (7.64)$$

The solid-state membrane may be a single crystal, polycrystalline, or a mixed pellet of an ionic salt. Mixed pellets of Ag_2S (with a solubility constant k_{sp} of 10^{-31}) and AgX respond to X^-, where X = Cl, Br, I. The charge transporting species in Ag_2S electrodes is Ag^+. Ion-selective electrodes for Cu^{2+}, Cd^{2+}, and Pb^{2+} can be made by mixing CuS, CdS, or PbS with Ag_2S. All mixed pellet Ag_2S electrodes also respond to Ag^+, so this will be an interferent. Some example solid-state ISEs for anion sensing are shown in Table 7.13.

Polymer-Based ISEs Liquid- and polymer-based ISEs use a mobile carrier (also ion exchanger, chelator, or ionophore = ion carrying, from the Greek phoros, meaning "bearing") in an organic solvent or a polymer matrix, respectively, to render that liquid or polymer ion-selective. We will restrict ourselves here to the discussion of polymer-based ISEs because they are easier to miniaturize than liquid-based ISEs. The ionophore is typically a hydrophobic species that is a chelator of the species of interest. An ion-selective polymeric membrane, say for a potassium ion sensor, is made by mixing a K^+ ionophore such as 1% valinomycin (a natural-occurring antibiotic) with 33% polyvinyl chloride (PVC) and 66% plasticizer (to make the rigid plastic more flexible: plasticized PVC acts like a sponge). The valinomycin (Figure 7.48) in the membrane selectively exchanges potassium ions from the solution (with a selectivity for K^+ over Na^+ of $k_{K^+,Na^+} = -4.3$) setting up a Donnan potential at the interface. The atomic radius of K^+ is 1.33 Å and that of Na^+ is 0.95 Å. With only this difference to work with, valinomycin manages to select for the K^+ ion over the Na^+ ion by a factor of nearly 5000.

No significant amount of counterions from solution may enter the membrane phase, or selectivity will be lost. To achieve this so-called *Donnan exclusion of counterions*, counterions must also be present

Table 7.13 Examples of Solid-State ISEs for Anions

Ion	Concentration Range (M)	Membrane Material	pH Range	Interfering Species
F^-	10^{-6} – 1	LaF_3	5–8	OH^- (0.1 M)
Cl^-	10^{-4} – 1	AgCl	2–11	CN^-, S^{2-}, I^-, $S_2O_3^{2-}$, Br^-
Br^-	10^{-5} – 1	AgBr	2–12	CN^-, S^{2-}, I^- S^{2-}
I^-	10^{-6} – 1	AgI	3–12	
SCN^-	10^{-5} – 1	AgSCN	2–12	S^{2-}, I^-, CN^-, Br^-, $S_2O_3^{2-}$
CN^-	10^{-6} – 10^{-2}	AgI	11–13	S^{2-}, I^-
S^{2-}	10^{-5} – 1	Ag_2S	13–14	

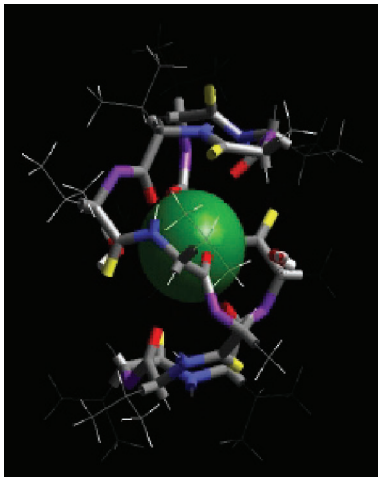


FIGURE 7.48 Valinomycin ionophore for K^+ ion.

in the membrane itself. To accomplish this, lipophilic counterions are added to the membrane cocktail [e.g., from sodium tetrakis[3,5-*bis*(trifluoromethyl) phenyl]borate or NaTFPB]. Addition of lipophilic salts to the membrane improves selectivity and also reduces the impedance of the probe. In a 3D ISE embodiment, as shown in Figure 7.49a the membrane potential is measured between an internal and an external reference electrode. Equation 7.61 gives the Nernstian response of polymeric membrane electrodes.

Ionophores for anions are harder to find than ionophores for cations because anions come in all types of shapes (e.g., SCN^- is linear but SO_4^{2-} is stetrahedral), whereas cations are mostly spherical (see Figure 7.48). Anion exchangers are typically R_4N^+ salts, and they respond according to the hydrophobicity of the anions; this is the Hofmeister series:

$$ClO_4^- > SCN^- > I^- > NO_3^- > Br^- > Cl^-$$

By using a combination of thin- and thick-film deposition techniques, one can fabricate planar ISEs on an inexpensive substrate as shown in Figure 7.49b. Thin-film silver electrodes are vapor deposited through a stencil mask on a polyimide substrate. The top layer of the Ag is converted to AgCl by dipping in a $FeCl_3$ solution. Silk screening of hydrogels with the appropriate electrolytes completes the internal and external reference electrodes. The internal reference side is now covered with the ISE membrane using a drop delivery technique, whereas the external reference side is covered with a dielectric membrane (also using drop delivery) with a small laser-cut hole for Luggin capillary. By simply changing the ionophore in the membrane, different ISEs can be fabricated. Although we show only one ISE in this figure, many planar sensors can be made in parallel (batch fabrication!). By switching from a 3D to 2D architecture, mass production of minute, polymeric ISEs (e.g., 2×3 mm per sensor) becomes possible, resulting in inexpensive (disposable), reproducible biosensors for ions such as Ca^{2+}, Na^+, K^+, NO_3^-, Cl^-, and so on. For ease of manufacturing, planar ISEs often do not use an internal reference solution. This makes the electrodes easier to handle and store, but voltage drift tends to be larger because there is no pinning of the Donnan potential on the internal reference side with a fixed reference ion concentration. The manufacturing approach sketched here is contrasted further below (under ISFETs) with an approach where the electronics are integrated. In Volume III, Chapter 1 we elaborate on the MEMS batch and continuous mode manufacture of arrays

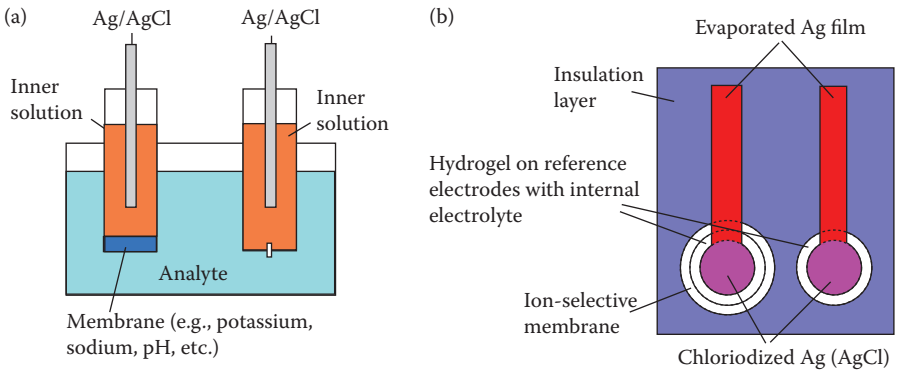


FIGURE 7.49 (a) A 3D measuring setup for polymeric ISEs. (b) A miniaturized 2D polymeric ISE.

of ISEs. In Table 7.14 we list membrane compositions and selectivities of important ISEs.

Membrane-Based Gas-Sensing Probes Gases that react with water, which causes freeing or absorbing of protons in the electrolyte, can be detected by a pH-sensitive element, e.g., a glass or an IrO_x electrode. A gas-permeable membrane (microporous), such as a 100-µm-thick Gore-Tex® film, separates the unknown sample from a pH probe and a reference electrode immersed together in a small internal solution (no ions can cross the gas-permeable membrane). The pH of the internal solution provides one with a measure of the gas content of the analyte solution. Gases that can be quantified this way include CO_2, NH_3, and H_2S. A direct proportionality exists between the concentration of the neutral gas and the measured pH. For example, in the case of the Severinghaus-type CO_2 sensor, depicted in Figure 7.50, the indicator electrode is only H^+-sensitive, and the potential of the reference is constant because of the constant Cl^- concentration in the internal electrolyte. Therefore, the potential of the pH sensor in the internal reservoir is given as:

$$E = E^0 + 0.059 \log a_{H^+} \tag{7.65}$$

The CO_2 gas penetrates through the gas-permeable membrane and reacts with the inner electrolyte,

Table 7.14 Membrane Compositions and Selectivities of Important Ion-Selective Electrodes

Analyte	Membrane Type	Membrane composition	Linear Range ($-\log a_i$)	log $K^{pot}_{i,j}$
H^+	Glass	72.2% SiO_2, 6.4% CaO, 21.4% Na_2O (mol%)	2–12	Na^+:–11 K^+:–11
H^+	Polymer	Tri-*n*-dodecylamine, PVC, *bis*-2-ethylhexylsebacate, tetraphenylborate	3–12	Na^+: –10.4 K^+: –9.8 Ca^{2+}: –11
Na^+	Glass	11% Na_2O,18% Al_2O_3, 71% SiO_2	1–6	K^+: –2 Ag^+: +2.6 NH_4: –4.2 H^+: 1–2.5
Na^+	Polymer	ETH 227, PVC, 2-nitrophenyloctyl ether, tetraphenylborate	1–3	Li^+: +0.3 K^+: –1.7 Ca^{2+}: –0.4 Mg^{2+}: –2.4
K^+	Polymer	Valinomycin, PVC, dioctyladipate	0–5	Na^+: –4.2 NH_4^+: –1.9 Ca^{2+}: –3.3 Mg^{2+}: –4.4
Li^+	Polymer	ETH 1810, PVC, 2-nitrophenyloctyl ether, tetraphenylborate	<2–4	Na^+: –2.2 K^+:–2.3 NH_4^+: –2.3 Ca^{2+}–27 Mg^{2+}: –4 H^+: +1
Li^+	Polymer	Dodecylmethyl-14-crown-4, PVC, 2-nitrophenyloctyl ether, tetraphenylborate	0–5	k^+: –1.8 Na^+:–2.1 NH_4^+: –2.8 Ca^{2+}: –4.2 Mg^{2+}: –4.6 H^+: –3.2
Ca^{2+}	Polymer	ETH 1001, PVC, 2-nitrophenyloctyl ether, tetraphenylborate	2–7	Na^+:–5.5 K^+:–5.4 Mg^{2+}: –4.9
Ca^{2+}	Polymer	Ca-di-(*n*-decyl) phosphate, PVC, di-(*n*-octylphenyl) phosphonate	0–5	Na^+:–4.4 K^+:–4.5 Mg^{2+}: –4.9
Mg^{2+}	Polymer	ETH 5282, PVC, 2-nitrophenyloctyl ether, KTpClPB	<2–5	Na^+:–3.8 K^+:–1.4 Ca^{2+}: –2.4
Ag^+	Solid state	Ag_2S	1–7	Cu^{2+}: –6 Pb^{2+}: –6 to –9 H^+: –5 Hg^{2+}: –2

which contains a high fixed amount of Cl^- and bicarbonate ($NaHCO_3$) in a hydrogel matrix:

$$CO_2 + H_2O \rightleftharpoons H^+ + HCO_3^- \tag{7.66}$$

and:

$$a_{H^+} = \frac{Kp_{CO_2} a_{H_2O}}{a_{HCO_3^-}} \tag{7.67}$$

Because the activities of water and HCO_3^- are constant, one obtains:

$$E = E^{0'} - 0.059\ pH \quad \text{and also} \quad E = E^{0''} - 0.059\ pCO_2 \tag{7.68}$$

The Severinghaus sensor shown in Figure 7.50a was fabricated by the author and his team using the same Ir/IrO_x wire illustrated in Figure 7.44.[8] For the reference electrode, a Ag/AgCl wire was used. Figure 7.50a is a schematic of the CO_2 sensor, whereas Figure 7.50b is a photo of the PVC tube-based sensor.

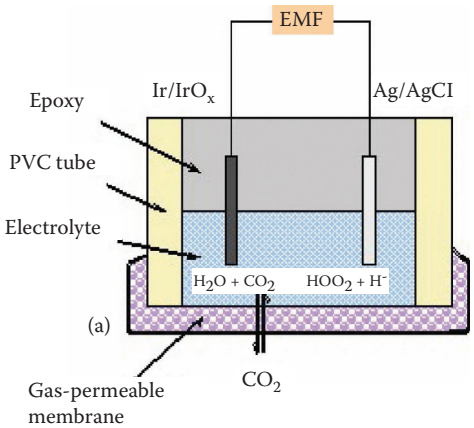
The world's smallest MEMS-fabricated electrochemical CO_2 sensor, incorporating thin-film Ir/IrO_x and Ag/AgCl electrodes, mounted in a catheter, together with an oxygen and a pH sensor, also made by this author and his team, is shown in Volume II, Figures 4.92–4.97. These miniaturized sensors are used clinically to determine blood CO_2 concentration.

Mass and Charge Transport in Solutions

As we saw in Chapter 6 on fluidics there are basically three modes of mass transport in solution: diffusion, migration, and convection (natural and mechanical). Diffusion from a region of high concentration to a region of low concentration occurs whenever the concentration of an ion or molecule at the surface of an electrode is different from that in the bulk of the solution. This will occur whenever a chemical change takes place at a surface. Convection takes place when mechanical means are used to carry reactants toward the electrode and to remove products from the electrode. The most common means of convection is to stir the solution using a stirring bar. Other methods include rotating the electrode and incorporating the electrode into a flow cell. Migration happens when charged particles in solution are attracted or repelled from an electrode that has a surface charge. The flux of material to and from the electrode surface is a complex function of all three modes of mass transport, and for most analytical techniques, one wants diffusion to dominate over migration. The reason, as we will find out further below, is that at the limit in which diffusion is the only significant means for the mass transport, the diffusion-limited current, i_l, in a voltammetric cell is simply related to the bulk concentration c_{bulk} ($x = \infty$) by:

$$i_l = nFAD_0\left[\frac{c_{bulk}(x=\infty)}{\delta}\right] \tag{7.124}$$

where n = number of electrons transferred in the redox reaction
F = Faraday's constant
A = area of the electrode



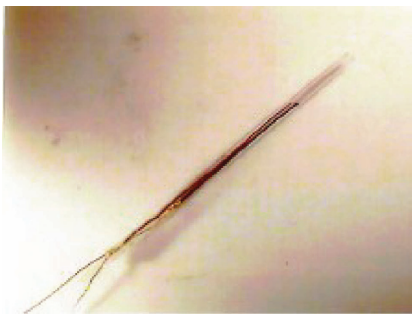


FIGURE 7.50 (a) Schematic of Severinghaus-type CO_2 sensor. (b) Photo of the sensor in a PVC tube. Notice the two wires: an Ir/IO_x wire and an Ag/AgCl wire. (By Dr. Marc Madou.)

D_0 = diffusion coefficient for the reactant or product
c_{bulk} $(x = \infty)$ = concentration of the analyte in the bulk solution
δ = thickness of the diffusion layer

For the above equation to be valid, migration and convection must not interfere with the formation of the diffusion layer around the electrode surface. Adding a high concentration of an inert supporting electrolyte to the analytical solution can eliminate migration. The large excess of inert ions ensures that few reactant and product ions will move as a result of migration. Although convection may be easily eliminated by not physically agitating the solution, in some situations it is desirable either to stir the solution or to flow the solution through an electrochemical flow cell. Fortunately, the dynamics of a fluid moving past an electrode results in a small diffusion layer, typically of 0.001–0.01-cm thickness; in this stagnant layer the rate of mass transport by convection drops to zero. Below our calculations for δ, the thickness of the steady-state diffusion layer, will reveal it to be only a fraction of the thickness of the stagnant hydrodynamic boundary layer, δ_h (see footnote on page 582).

The conductance *S* of an electrolyte, in units of Siemens (S), is given by:

$$S = \frac{1}{R} = \frac{\rho A}{l} \tag{7.69}$$

where R = resistance of the electrolyte
A = active area of contact
l = length (thickness) of electrolyte layer
ρ = conductivity–intrinsic property of the electrolyte (in S/cm)

The conductivity ρ of the electrolyte has a contribution from every ion in solution that is proportional to the concentration of the ions (*c*), the charge of the ions (*z*), and a property that determines their migration-velocity—the mobility, μ. In Chapter 6 we defined mobility, μ, as the limiting velocity of the ion in an electric field of unit strength. As an electric field (F = eEz, where *e* is the electronic charge) forces ions to move through the solution, an opposing force resulting from frictional drag is generated. The latter is represented by the Stokes equation:

$F_D = 6\pi\eta Vr$ (see Equation 6.26)

where η = viscosity of the solution
r = Stokes radius of the ion
V = ion velocity

When the opposing frictional force exactly counterbalances the force exerted on the ion by the electrical field, the ion attains a steady-state velocity called the *terminal velocity*. This terminal velocity is termed the mobility (μ) when the electric field strength is unity:

$$\mu = \frac{ze}{6\pi\eta r} \tag{7.70}$$

With this expression for mobility μ, we can write the following expression for conductivity ρ:

$$\rho = F\sum_i z_i \mu_l c_i \tag{7.71}$$

This is often expressed as molar ionic conductivity Λ, which is given as:

$$\Lambda = \frac{\rho}{c} \tag{7.72}$$

where *c* is in moles and Λ is in units of S m^2/mol. Some examples of molar ionic conductivity for ions in water at 25°C and at infinite dilution are given in Table 7.15.

At infinite dilution, no interionic interactions take place; thus, the molar conductivity of a salt C_cA_a can simply be calculated from its constituent ions as:

$$\Lambda_{C_cA_a} = c\Lambda_c + a\Lambda_a \tag{7.73}$$

Example 7.7:

Calculate the molar conductivity of KCl and $BaCl_2$ (two good electrolytes):

- $\Lambda_{KCl} = \Lambda_{K^+} + \Lambda_{Cl^-} = (74 + 77) \times 10^{-4}$ S m^2/mol
  $= 151 \times 10^{-4}$ S m^2/cm
- $\Lambda_{BaCl_2} = \Lambda_{Ba^{+2}} + 2\Lambda_{Cl^-} = \left[127 + (2 \times 77)\right] \times 10^{-4}$
  S m^2/mol $= 281 \times 10^{-4}$ S m^2/mol

A useful measure of how much of the current is carried by the cations versus the anions of a salt

Table 7.15 Molar Ionic Conductivity

Ion	$\Lambda(10^{-4}$ S m²/mol) at Infinite Dilution (25°C)
H^+	350
OH^-	200
Ba^{2+}	127
K^+	74
Cl^-	77
HCO_3^-	45
TBA^+	24

is the transference or transport number. The fractions of current carried by the positive and negative ions are given by their transport numbers t_+ and t_-, respectively:

$$t_+ = c_+\Lambda_+ / \Lambda_{C_cA_a} \tag{7.74}$$

$$t_- = a_-\Lambda_- / \Lambda_{C_cA_a} \tag{7.75}$$

where $c_+\Lambda_+ + a_-\Lambda_- = \Lambda_{C_cA_a}$. Because each mole of current passed through the cell corresponds to 1 mole of electrochemical change at each electrode, the amount of ions transported in the electrolyte also equals 1 mole. Thus:

$$t_+ + t_- = 1 \tag{7.76}$$

Example 7.8:

Calculate the transference numbers for the cations in KCl and $BaCl_2$ (two good electrolytes):

- t_+ (KCl) = 74/151 = 0.49
- t_+ ($BaCl_2$) = 127/281 = 0.45

Solvent and concentration affect ionization and therefore both ionic conductivity and transference numbers.

For multiple species, Equation 7.76 is generalized to:

$$\sum_i t_i = 1 \tag{7.77}$$

This equation is valid when more than one type of ion is in the solution. Now the transport numbers of ions are determined by the conductivity (ρ) of the electrolyte (Equation 7.71). The transport numbers may be represented as the ratio of the contribution to conductivity made by a particular ion to the total conductivity of the solution, or based on Equation 7.71:

$$t_i = \frac{\rho_i}{\rho} = \frac{z_i\mu_i c_i}{\sum_i z_i\mu_i c_i} \tag{7.78}$$

This equation gives us a clue how to eliminate migration for an analyte ion of interest, say, redox species *i*. Current flow resulting from electrostatic attraction (or repulsion) of analyte ion *i* by the electrode is reduced to a negligible level by the presence of a high concentration of supporting or indifferent electrolyte. By increasing the concentration for all the other ions but ion *i*, the transport number for ion *i* is minimized. The only way for an analyte ion, present in much lower concentration than the supporting electrolyte, to reach the electrode is by diffusion. Typically, a supporting electrolyte concentration 50–100 times greater than the analyte concentration will reduce electrostatic transport of the analyte to a negligible level and minimize the iR drop in the solution as well. Under the latter circumstances, qualitative and quantitative analysis of an ion analyte, using Equation 7.124 (see further below), is possible.

Energetics of Metals, Semiconductors, Membranes, and Mixed Conductors in Contact with an Electrolyte

Introduction

In Chapters 3 and 4, we introduced the band theory of solids and learned how to draw energy diagrams for solid-state junctions, such as metal/semiconductor and *p*-type and *n*-type semiconductor combinations. In this section, we derive the energy diagrams for solid/liquid junctions that explain the operation of these materials as sensor electrodes, in fuel cells, in batteries, as liquid junction solar cells, ISFETs, during corrosion processes (including the anisotropic and isotropic etching of semiconductors), and so on.

When electrons or ions pass through a solid/electrolyte interface, a reversible electrode potential E_{rev} is set up. This equilibrium is a dynamic equilibrium, meaning that the rates of charge carriers passing

through the interface in either direction are equal. We denote the forward electron/ion transfer and the backward electron/ion transfer rates as $\vec{i}$ and $\overleftarrow{i}$, respectively. At equilibrium, the exchange current density is the same in both directions, or $\vec{i} = \overleftarrow{i} = i_0$. Thermodynamic equilibrium between two phases is reached when the electrochemical potentials on either side of the interface become equal. To refresh the reader's understanding of this process, we briefly reconsider here the relation between chemical potential μ, electrochemical potential $\bar{\mu}_e$, and Fermi level E_F. Every substance has a unique propensity to contribute to a system's energy. We call this property the chemical potential μ. When the substance is a charged particle (such as an electron or an ion with charge z) we must include the response of the particle to an electrical field in addition to its chemical potential. This sum is called the *electrochemical potential*, which in the case of electrons is given by $\bar{\mu}_e = \mu + zF\phi$ with F the Faraday constant and ϕ the potential drop. It can be demonstrated that the Fermi level, E_F, in a given material is equivalent to the electrochemical potential of the electrons, $\bar{\mu}_e$, in that material expressed per particle, or:

$$E_F = \frac{\bar{\mu}_e}{N_A} \tag{7.79}$$

As explained in Chapters 3 and 4, when bringing a solid in contact with another solid, the Fermi levels line up, and a potential drop ϕ is established across the interface. The potential drop is always the largest in the phase with the highest resistance (Ohm's law). For example, for a metal/semiconductor contact the whole potential drop is located in the semiconductor space charge (there are many more charge carriers in a metal than in a semiconductor), but for a metal/electrolyte contact the potential drop is all in the electrolyte (there are many more charge carriers in the metal than in the electrolyte).

Metal/Electrolyte Interface

Introduction In this section we detail the electron exchange current $i_{0,e}$ at a metal/electrolyte contact, as well as the current i as a function of the applied overvoltage η. We first derive these properties of the metal/liquid junction phenomenologically before we approach the same problem semiquantitatively, introducing the energetics of the solid/liquid interface. For simplicity in equations we will abbreviate $i_{0,e}$ as i_0.

Phenomenological Description of the Exchange Current $i_{0,e}$ and the Relationship between Current and Metal Electrode Overpotential We will first assume that the reactions at the metal electrode involve the two partners of the same redox system (say, Fe^{3+} and Fe^{2+}), so that at equilibrium no macroscopic changes take place at the electrode. The forward current $\vec{i}$ ($Fe^{3+} + e^- \rightarrow Fe^{2+}$) and the backward current $\overleftarrow{i}$ ($Fe^{2+} - e^- \rightarrow Fe^{3+}$) cancel each other out, i.e., the exchange current density is the same in both directions, or $|\vec{i}| = |\overleftarrow{i}| = i_0$ and the total current $i = \vec{i} - \overleftarrow{i} = 0$. This situation is called a *dynamic equilibrium at a potential* E^0. When no thermodynamic equilibrium is established for zero total current ($i = 0$ but $i \neq i_0$), $\vec{i}$ and $\overleftarrow{i}$ belong to different redox couples, and macroscopic changes occur at $i = 0$. In this case the potential at $i = 0$ is called a *mixed potential*, E^{mix} (e.g., in iron corrosion two redox couples, one involving oxygen and one iron, are responsible for this mixed potential).

Overpotentials occur when an external current is superimposed on the electrochemical cell. The electrode then assumes a voltage E_i different than the equilibrium value E^0 to an extent governed by the current i. The difference between E_i and the equilibrium voltage is known as the *overpotential*, generally denoted by η:

$$\eta = E_i - E^0 = \phi - \phi^0_{eq} \tag{7.80}$$

The rate of electron transfer between the metal and the redox couple, v, can be simply described as in homogeneous chemical kinetics:

$$v = k_{ox} c^s_{red} - k_{red} c^s_{ox} \tag{7.81}$$

From the absolute rate theory it follows that:

$$k_{ox} = A \exp\left[-\frac{\Delta G^*_{ox}(\phi)}{kT}\right]$$
$$k_{red} = A \exp\left[-\frac{\Delta G^*_{red}(\phi)}{kT}\right] \tag{7.82}$$

where ΔG^*_{ox} and ΔG^*_{red} are activation energies.

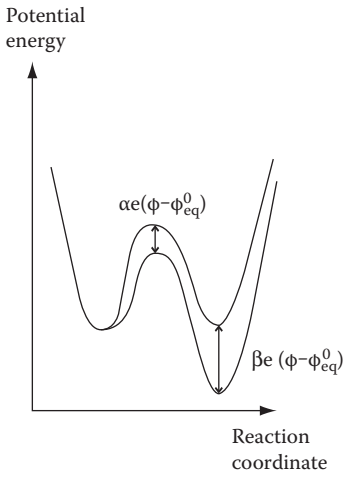


FIGURE 7.51 Reaction coordinate in the metal redox system.

The activation terms are dependent on the electrode potential but not the pre-exponential factor *A*, and with reference to Figure 7.51 we can write:

$$\Delta G^*_{ox}(\phi) = \Delta G^*_{ox}\left(\phi^0_{eq}\right) - \alpha e\left(\phi - \phi^0_{eq}\right)$$

$$\Delta G^*_{red}(\phi) = \Delta G^*_{red}\left(\phi^0_{eq}\right) + \beta e\left(\phi - \phi^0_{eq}\right)$$

$$\text{with } \alpha = -\frac{1}{e}\frac{\partial \Delta G^*_{ox}(\phi)}{\partial \phi}\Big|_{\phi^0_{eq}}$$

$$\text{and } \beta = \frac{1}{e}\frac{\partial \Delta G^*_{red}(\phi)}{\partial \phi}\Big|_{\phi^0_{eq}} \qquad (7.83)$$

The Gibbs energy of activation and reaction are correlated by:

$$\Delta G^*_{ox}(\phi) - \Delta G^*_{red}(\phi) = G_{ox} - G_{red}$$

and

$$\Delta G^*_{ox}\left(\phi^0_{eq}\right) = \Delta G^*_{red}\left(\phi^0_{eq}\right) = \Delta G^*_{eq} \qquad (7.84)$$

Assuming a transfer reaction in which the electrostatic energy of the ionic species is not significantly affected by the electrode potential (referred to as an *outer sphere electrode reaction*), it follows that:

$$\Delta G^*_{ox}(\phi) - \Delta G^*_{red}(\phi) = -e\left(\phi - \phi^0_{eq}\right) \qquad (7.85)$$

Based on Equations 7.83 and 7.85, we obtain:

$$\alpha + \beta = 1 \qquad (7.86)$$

We further develop Equation 7.81 to obtain:

$$i = ek_0 c^s_{red} \exp\left[\frac{\alpha e\left(\phi - \phi^0_{eq}\right)}{kT}\right] - ek_0 c^s_{ox} \exp\left[-\frac{(1-\alpha)e\left(\phi - \phi^0_{eq}\right)}{kT}\right] \qquad (7.87)$$

$$\text{where } k_0 = A\exp\left(-\frac{\Delta G^*_{eq}}{kT}\right)$$

To connect voltages to concentrations we recall the Nernst expression (Equation 7.56), which we rewrite here as:

$$\phi_{eq} = \phi^0_{eq} + \frac{kT}{e}\ln\left(\frac{c^s_{ox}}{c^s_{red}}\right) \qquad (7.88)$$

Using the expression for the overpotential η (Equation 7.80), we can finally derive the so-called *Butler-Volmer equation* for charge transfer reactions at a metal/electrolyte interface as:

$$i = i_0\left[\exp\left(\frac{\alpha e\eta}{kT}\right) - \exp\left(-\frac{(1-\alpha)e\eta}{kT}\right)\right] \qquad (7.89)$$

This relationship is illustrated in Figure 7.52a for three different values of α. The transfer coefficient

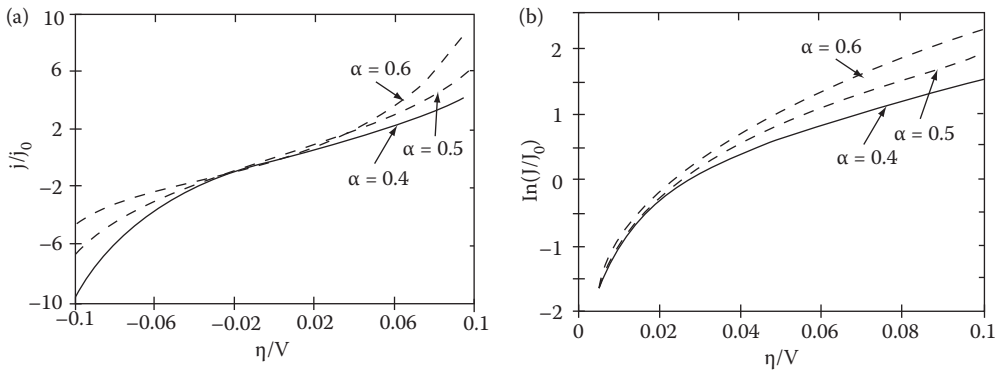


FIGURE 7.52 (a) The Butler-Volmer expression for three different values of α. (b) The Tafel law.

α determines the symmetry (or lack thereof) of the current potential curve. Notice that the Butler-Volmer analysis is based on surface concentrations. The exchange current density i_0 in Equation 7.89 is given by:

$$i_0 = ek_0\left(c^s_{red}\right)^{1-\alpha}\left(c^s_{ox}\right)^{\alpha} \tag{7.90}$$

For low overpotentials (η) the system behaves as an ohmic resistance:

$$i = i_0\frac{e\eta}{kT} \tag{7.91}$$

It follows that the charge transfer resistance, R_{ct}, is given by:

$$R_{ct} = \frac{N_A kF}{N_A ei_0} = \frac{RF}{Fi_0} \tag{7.92}$$

For large overpotentials the Butler-Volmer equation becomes logarithmic (Tafel regime):

$$\eta = a + \log i \tag{7.93}$$

The Tafel regime is illustrated in Figure 7.52b.

Semiquantitative Description of the Exchange Current $i_{0,e}$ and the Relationship between Current and Metal Electrode Overpotential

Equilibrium Current Exchange Density $i_{0,e}$ To derive the electron transfer rate at the metal/redox electrolyte interface semiquantitatively we must establish the energy density of empty and filled states that electrons can tunnel to and from on both sides of the interface. The question, as in the case of electron transfer between two solids, is whether there will be electron transfer from the metal electrode to the redox molecules in solution (or vice versa). This, we find, depends on the alignment of the density of states of the redox molecules relative to the Fermi level of the metal electrode. In Chapter 3, we learned how to calculate the density of states in solids, and here we do the same for redox couples in solution. In molecules, the lowest empty states are the lowest unoccupied molecular orbitals (LUMOs), and the highest filled states are the highest occupied molecular orbitals (HOMOs). The Fermi level in a solid metal is like the HOMO in a molecule (Figure 7.53).

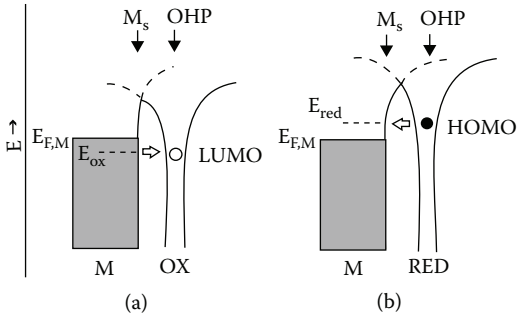


FIGURE 7.53 Solid metal/redox molecule interface energy levels. Fermi level of a solid metal and (a) energy levels of an oxidized (LUMO) and (b) a reduced (HOMO) redox species.

We can understand the concept of density of states of a redox couple in a solution in the following way. When a molecule is in the oxidized state, it has empty levels centered at E_{ox}. Once an electron occupies the oxidized state, the molecule relaxes to a lower energy level, marked as E_{red} in Figure 7.54. The relaxation includes the reorientation of the surrounding solvent molecules and changes in the chemical bonds of the reacting molecule. For example, consider the simple redox couple, Fe^{3+}/Fe^{2+}. The solvent polarization for Fe^{3+} is different from that of Fe^{2+}. Following the electron transfer $Fe^{3+} \rightarrow Fe^{2+}$, the dipoles of the solvent molecules reorient. The extra negative charge on the central ion causes the negative poles of the solvent dipoles to rotate away from the central ion, making the potential at the central ion higher and more attractive to electrons. This explains why E_{red} is always below E_{ox}. The shift in energy level

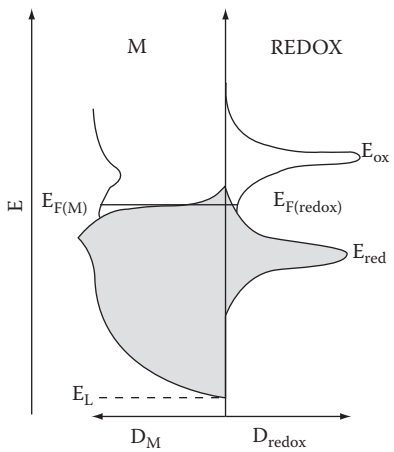


FIGURE 7.54 Time-fluctuating energy levels of the redox system can be regarded as bands of energy levels.

between oxidized (unoccupied) and reduced (occupied) levels is called the *Frank-Condon shift*. The Frank-Condon principle states that during electron transfer from one level to another, atoms and ions are effectively frozen in their positions; thus, they can move or relax only before or after the electron transfer. The amount of Frank-Condon shift, or $E_{ox} - E_{red}$, is given by:

$$E_{ox} - E_{red} = 2\lambda \tag{7.94}$$

where λ is called the *reorganization energy*. The reorganization energy contains contributions from all involved bonds and solvent polarization and is normally on the order of a few tenths of electron volts. In some cases, such as ions or metal ion complexes, the reorganization caused by bond length and angle changes is small, and the main contribution results from solvent polarization.

According to Morrison,[9] the time-fluctuating energy levels of the redox system can be regarded as bands of energy levels and are described by the densities of the oxidized and reduced states. The state density of electrons in the reduced $[D_{red}(E)]$ and oxidized $[D_{ox}(E)]$ states is given by the product of the probability density distribution, $W(E)$, and the concentration, c, or:

$$D_{red}(E) = c_{red}W_{red}(E) \text{ and } D_{ox}(E) = c_{ox}W_{ox}(E) \tag{7.95}$$

The total density of states for a redox couple is defined as:

$$D_{redox}(E) = D_{red}(E) + D_{ox}(E) \tag{7.96}$$

Similarly, for the density of states of electrons in the metal:

$$D_M(E) = D_{M(e)}(E) + D_{M(h)}(E) \tag{7.97}$$

where *e* means a filled state and h an empty one. The probability density distribution curves in solution, $W(E)$, take the form of:

$$W_{red}(E) = W_{red}(E_{red})\sqrt{\frac{c_{ox}}{c_{red}}}\exp\left(-\frac{\lambda_{red}}{4kT}\right)\exp\left\{-\frac{(1-\beta)\left[E - E_{F(redox)}\right]}{kT}\right\} \tag{7.98}$$

and:

$$W_{ox}(E) = W_{ox}(E_{ox})\sqrt{\frac{c_{red}}{c_{ox}}}\exp\left(-\frac{\lambda_{ox}}{4kT}\right)\exp\left\{-\frac{\beta\left[E - E_{F(redox)}\right]}{kT}\right\} \tag{7.99}$$

where β is the same symmetry factor we encountered in the Butler-Volmer expression; if $\lambda_{ox} = \lambda_{red}$, then $\beta = 0.5$. Equations 7.98 and 7.99 describe Gaussian distributions of energy states around E_{red} and E_{ox}, respectively, and these two curves, for $c_{ox} = c_{red}$, intersect at E^0. This intersection corresponds to the Fermi level of the redox species, $E_{F(redox)}$:

$$E_{F(redox)} = E^0 = (1/2)\,(E_{ox} + E_{red}) \tag{7.100}$$

Based on Equation 7.96 it then also holds that:

$$\frac{1}{2}D_{redox}\left[E_{F(redox)}\right] = D_{red}\left[E_{F(redox)}\right] = D_{ox}\left[E_{F(redox)}\right] \tag{7.101}$$

Consequently, we derive from Equations 7.96, 7.98, 7.99, and 7.101:

$$D_{redox}(E) = \frac{1}{2}D_{redox}\left[E_{F(redox)}\right]\left\{\exp\left\{\frac{\beta\left[E - E_{F(redox)}\right]}{kT}\right\} + \exp\left\{-\frac{(1-\beta)\left[E - E_{F(redox)}\right]}{kT}\right\}\right\} \tag{7.102}$$

This expression is sketched in Figure 7.55.

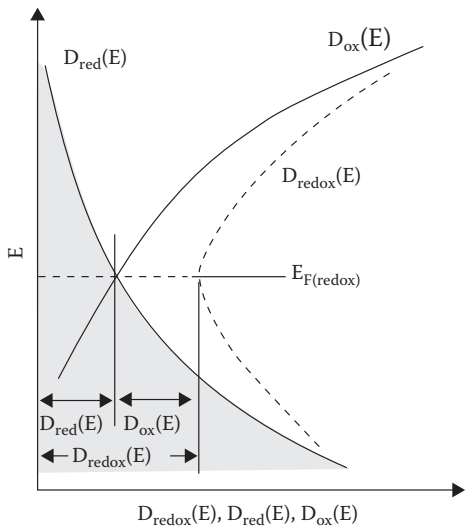


FIGURE 7.55 The state density of electrons in the reduced $D_{red}(E)$ and oxidized $D_{ox}(E)$ and the total density of states for a redox couple $D_{redox}(E) = D_{red}(E) + D_{ox}(E)$.

Because we derived expressions for the density of state functions on both sides of the interface, we are now in a position to calculate the equilibrium current exchange density i_0 across the metal/electrolyte interface. Electron transfer takes place by isoenergetic tunneling across the 0.3–0.5-nm compact layer at the metal electrode interface (OHP, see Figure 7.14). The rate of the reaction will depend on the density of states in the metal and in the redox couple. Using Fermi's golden rule (FGR) approximation (see Chapter 3, Equation 3.285), we calculate the probability of transitions of electrons between the eigenstates associated with the metal and the redox system. In this model one assumes that electron exchange occurs between levels of equal energies in both phases (the metal and the redox species in this case), and an electronic coupling strength, H, between metal electrode and the redox state modulates that transition probability. We can express the forward (cathodic) and backward (anodic) currents as a function of the energy level as:

$$\vec{i}(E) = ek_c D_{M(e)}(E) D_{ox}(E) \tag{7.103}$$

and:

$$\overleftarrow{i}(E) = ek_a D_{M(h)}(E) D_{red}(E) \tag{7.104}$$

where k_c and k_a are electron tunnel constants. Because we are dealing with electrons we also have to introduce the Fermi-Dirac distribution functions:

$$\vec{i}(E) = ek_c(E) D_M(E) f\left[E - E_{F(M)}\right] D_{redox}(E)(1-f)\left[E - E_{F(redox)}\right] \tag{7.105}$$

and:

$$\overleftarrow{i}(E) = ek_a(E) D_M(E)\left\{(1-f)\left[E - E_{F(M)}\right]\right\} D_{redox}(E) f\left[E - E_{F(redox)}\right] \tag{7.106}$$

With the Fermi level of the metal electrode being level with the redox couple's Fermi level [$E_{F(M)} = E^0 = E_{F(redox)}$], the oxidation probability (electron transfer from the molecule to the electrode) equals the reduction probability (electron from the electrode to the molecule), i.e., the exchange current density is the same in both directions, or $|\vec{i}| = |\overleftarrow{i}| = i_0$ and the total current $i = \vec{i} - \overleftarrow{i} = 0$. At equilibrium it also holds that the tunneling constant $k_t = k_a = k_c$, i.e., principle of microreversibility:

$$\begin{aligned} i_0 = \vec{i}_0(E) &= ek_t(E) D_M(E) f\left[E - E_{F(M)}\right] D_{redox}(E)(1-f)\left[E - E_{F(redox)}\right] \\ &= ek_t(E) D_M(E)\left\{(1-f)\left[E - E_{F(M)}\right]\right\} D_{redox}(E) f\left[E - E_{F(redox)}\right] = \overleftarrow{i}_0(E) \end{aligned} \tag{7.107}$$

The overall electron exchange current is obtained by integration of the microscopic current elements over the entire energy range:

$$i_0 = \int_{-\infty}^{\infty} ek_t(E) D_M(E) f\left[E - E_{F(M)}\right] D_{redox}(E)(1-f)\left[E - E_{F(redox)}\right] dE \tag{7.108}$$

Assuming $D_M(E)$ and k_t are energy-independent, Equation 7.108 can be somewhat simplified, and with further development this leads to:

$$i_0 = ek_t D_M\left[E_{F(M)}\right] \frac{1}{2} D_{redox}\left(E_{F(redox)}\right) B(\beta) \tag{7.109}$$

$$B(\beta) = \int_{-\infty}^{\infty} \exp\left\{\frac{\beta\left[E - E_{F(redox)}\right]}{kT}\right\} f\left[E - E_{F(M)}\right] dE \approx \frac{\pi kT}{\sin\left[(1-\beta)\pi\right]} \tag{7.110}$$

Ninety percent of the integrand falls in a zone of 0.25 eV around the Fermi level, and we finally obtain:

$$i_0 \approx \frac{e\pi kT}{2\sin\left[(1-\beta)\pi\right]} k_t\left[E_{F(M)}\right] D_M\left[E_{F(M)}\right] D_{redox}\left[E_{F(redox)}\right] \tag{7.111}$$

This expression is plotted in Figure 7.56. Figure 7.56a repeats the time-fluctuating energy levels of the redox system and the energy bands in the metal. The relation $k_t(E)$ is illustrated in Figure 7.56b, and $i_0(E)$ is shown in Figure 7.56c.

Out of Equilibrium: Current under Polarization

Applying an overvoltage shifts the redox states of the molecule in solution relative to the Fermi level of the metal electrode. This triggers electron transfer between the electrode and the molecule in an energy

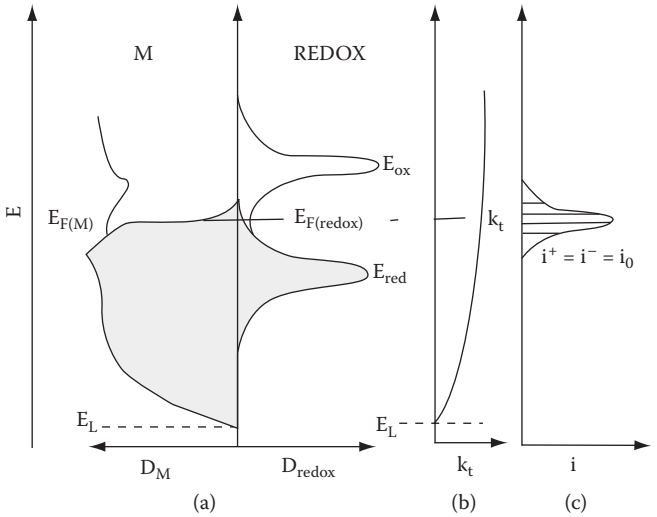


FIGURE 7.56 Using Fermi's golden rule, we can calculate the current exchange density i_0 across the metal/electrolyte interface in the presence of a redox couple. (a) Time-fluctuating energy levels of the redox system can be regarded as bands of energy levels. (b) The relation of tunneling constant k_t and energy E. (c) Electron exchange current density i_0 versus energy. The lowest energy level is E_L.

range close to the Fermi level. Out of equilibrium, the Fermi levels are offset by the overvoltage η or the total current $i(\eta) = \overleftarrow{i}(\eta) - \vec{i}(\eta) \neq 0$. Based on Equation 7.80 we rewrite the overvoltage here as:

$$\eta = E_{F(M)} - E_{F(redox)} \tag{7.112}$$

With this simple expression we can rework Equations 7.105 and 7.106 as:

$$\vec{i}(\eta) = e\int_{-\infty}^{\infty} k_t(E,\eta) D_M(E) f\left[E - E_{F(redox)} + e\eta\right] D_{redox}(E)(1-f)\left[E - E_{F(redox)}\right] dE \tag{7.113}$$

$$\overleftarrow{i}(\eta) = e\int_{-\infty}^{\infty} k_t(E,\eta) D_M(E)\left\{(1-f)\left[E - E_{F(redox)}\right] D_{redox}(E) f\left[E - E_{F(redox)}\right] dE \tag{7.114}$$

Using Equation 7.102, we find another Butler-Volmer-type expression (Equation 7.89: remember $\alpha + \beta = 1$) for the current i versus η:

$$i = i_0 \frac{k_t\left[E_{F(M)}\right]}{k_t\left[E_{F(redox)}\right]}\left[\exp\left(\frac{(1-\beta)e\eta}{kT}\right) - \exp\left(\frac{-\beta e\eta}{kT}\right)\right] \tag{7.115}$$

This expression explicitly contains the electron tunneling rate constant k_t, for which there are semi-quantitative data (see section below). In Figure 7.57, we illustrate forward and backward currents across the metal/electrolyte interface.

Notice that all the potential changes fall within the less conductive medium, i.e., the electrolyte in this case.

Electron Tunneling Transfer Rate k_t According to the Marcus theory (1992 Nobel Prize),[10] the electron transfer rate k_t in Equation 7.115 depends on the activation energy for electron transfer, ΔG^*_{ACT}, and the electronic coupling, H_{DA}, between the two electron-exchanging species (say, a redox enzyme and a metal electrode). The electronic coupling is often very difficult to calculate and is widely treated as a fitting parameter to experimental data. With reference to Figure 7.58, at the interception point of the potential energy curves, the acceptor energy level (A) is equal to the donor level (D). The energy

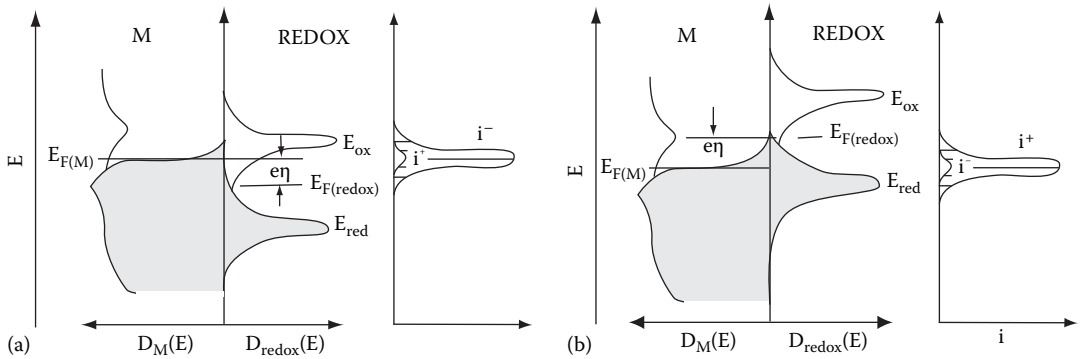


FIGURE 7.57 Applying a voltage across the metal/electrolyte interface that is different from the equilibrium potential leads to cathodic (a) or anodic (b) currents that are different from the equilibrium exchange current i_0.

barrier (activation energy) or charge transfer at this point is given as:

$$\Delta G^{*}_{ACT} = \frac{\lambda}{4}\left(1+\frac{\Delta G^{0}}{\lambda}\right)^{2} \tag{7.116}$$

where ΔG^0 is the free energy difference between initial and final states. In Figure 7.58 we made $\Delta G^0 = 0$ for simplicity so that both curves have a minimum at the same energy level. If thermal fluctuations change the distance between *D* and *A* (see reaction coordinate in Figure 7.58), then the electron transfer between donor and acceptor will be modified with a probability factor of:

$$k_t = \frac{1}{\sqrt{4\pi\lambda k_B T}}\exp\left[-\frac{\left(\Delta G^{0}+\lambda\right)^{2}}{4k_B T\lambda}\right] \tag{7.117}$$

In the case of strong coupling between the donor and acceptor orbitals, i.e., a large H_{DA}, electron transfer takes place as soon as the donor and acceptor levels line up. In this case, the electron transfer rate k_t is simply given as:

$$k_t = \nu\exp\left[-\frac{\left(\Delta G^{0}+\lambda\right)^{2}}{4k_B T\lambda}\right] \tag{7.118}$$

where ν is the frequency of the nuclear coordinate, which is typically between 10^{12} and 10^{14} s^{-1}. This extreme situation represents the so-called *adiabatic electron transfer*. On the other hand, in the weak coupling limit or the nonadiabatic case, the electron transfer rate, obtained by applying Fermi's golden rule (Equation 3.285), is given by:

$$k_t = \frac{2\pi}{\hbar}H^{2}_{DA}\frac{1}{\sqrt{4\pi\lambda k_B T}}\exp\left[-\frac{\Delta G^{*}}{k_B T}\right]$$

$$\text{where } \Delta G^{*} = \frac{\lambda}{4}\left(1+\frac{\Delta G^{0}}{\lambda}\right)^{2} \tag{7.119}$$

The above considerations include only one nuclear coordinate, but the result is essentially the same when multiple nuclear coordinates are considered except that the reorganization energy, λ, contains contributions from all involved bonds and solvent polarization. As mentioned before, the reorganization energy λ is normally on the order of a few tenths of electron volts.

Because H_{DA} in general diminishes exponentially with distance (like any orbital), the distance dependence of the rate of electron transfer thus can be described by:

$$k_t = k_0 e^{-\beta d} e^{-\frac{\left(\Delta G^{0}+\lambda\right)^{2}}{4k_B T\lambda}} \tag{7.120}$$

in which β (the attenuation factor) is a constant scaling the distance dependence and k_0 is the rate at close contact (~10^{13} s^{-1}). The medium between donor and acceptor strongly influences the rate of the electron transfer process through solvation and the value of β. Many researchers have focused on determining the electron transfer rate k_t as a function of the donor-acceptor separation that they systematically controlled experimentally. If donor and acceptor are separated by a vacuum, the interaction (through space interaction) is much less (β ≈ 2.8–3.8 $Å^{-1}$) than when a protein, DNA (β ≈ 1.4 $Å^{-1}$), or a solvent (through solvent coupling), like benzene, is the medium (β ≈ 1 $Å^{-1}$). In contrast, for highly conjugated organic bridges, β is much smaller, lying in the range of 0.2–0.6 $Å^{-1}$. With DNA, proteins immobilized on electrodes, and saturated organic molecules, β is found between these two limits. But if the energy level of the bridge becomes too low, it can work as a trap for electrons, and the transfer stops (now the electrons localize on the bridge).

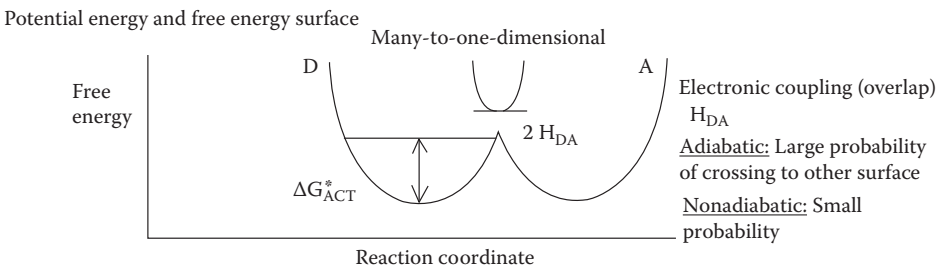


FIGURE 7.58 Potential energy and free energy surface, where H_{DA} is the electronic coupling and the electron transfer (ET) is thermal instead of optical.

Equation 7.120 is of utmost importance when attaching large molecules with reaction sites or redox centers to electrode surfaces. If the redox center is too far removed from the metal surface, no electron transfer can occur. Direct electron transfer between enzymes and metal electrodes, for example, is normally very slow as a consequence of steric insulation of the redox center in the protein. To illustrate this point, the "depth" of the redox center in glucose oxidase enzyme (GOx) is 8.7 Å, whereas for horseradish peroxidase (HRP) it is 4.1 Å. This is too deep below the protein surface for efficient electron transport to the underlying metal electrode. One solution to electronically connect to such a deep-lying redox center [flavin adenine dinucleotide ($FAD/FADH_2$) in the case of GOx (Figure 7.59)] is to rely on an electron acceptor molecule (mediator) to establish electrical communication with the electrode. Enzyme mediators are used in current (second-generation) commercial glucose sensors. Before introducing a second and even a third generation of these sensors, we consider how a first generation of glucose sensors operated.

When an enzyme reacts with its substrate *S*, product *P* is generated. In the case of GOx, this reaction is given as:

$$\text{Glucose (S)} + O_2 \xrightarrow{\underline{\text{GOx}}} \text{glucono-}\delta\text{-lactone} + H_2O_2\ \text{(P)} \qquad \text{Reaction 7.5}$$

As illustrated in Figure 7.60a, in this process the redox center in the enzyme is reduced from FAD to $FADH_2$, and it is the oxygen present in the sample that regenerates the enzyme by acting as an electron acceptor. Oxygen is called a *cosubstrate of the enzyme* in this case. To make a sensor based on this reaction, one might follow the anodic oxidation of H_2O_2 or the cathodic reduction of O_2. A first generation of glucose sensor was made this way.

In a second generation of sensors, the rationale was to monitor the redox center of the enzyme more efficiently. In the case of GOx, that means monitoring the $FAD/FADH_2$ redox couple directly. But, as we just saw, that redox center is located too far from the surface of the enzyme to be electrically accessible. The solution is to rely on a fast mediator redox system (MED_{red}/MED_{ox}) to establish electrical communication with the electrode (see Figure 7.60b). These mediators are chosen to regenerate FAD from $FADH_2$ much faster than oxygen does; in other words, a mediator redox couple where MED_{ox} is a much better electron acceptor than oxygen. The ferrocene/ferrocinium redox couple self-exchanges electrons very rapidly and is used as a diffusing mediator to shuttle electrons between the protein redox center and the electrode. Mediators can also be provided at much larger concentrations than

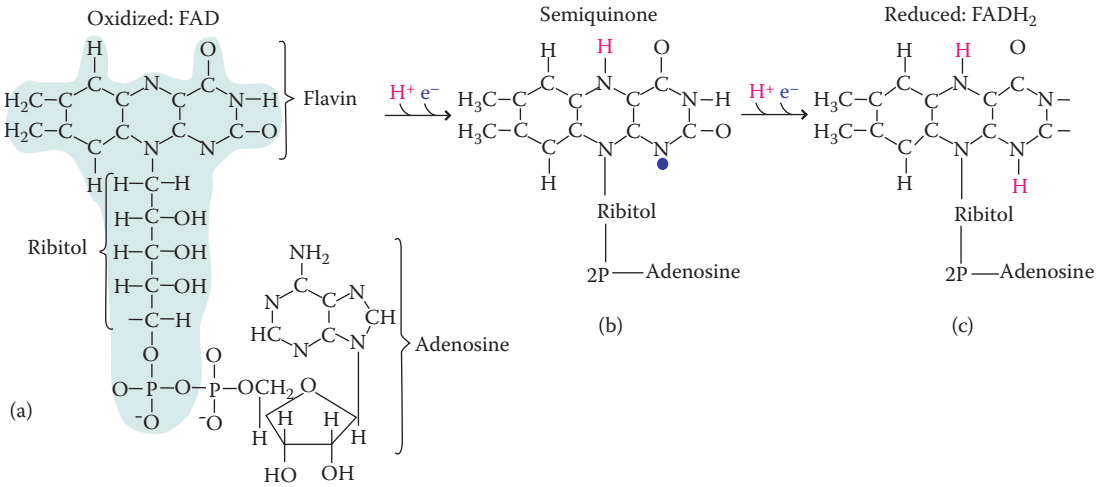


FIGURE 7.59 The redox center, flavin adenine dinucleotide ($FAD/FADH_2$). In the case of glucose oxidase (GOx) this center is buried 8.7 Å deep within the enzyme. The reduced form of the redox center or hydroquinone form is $FADH_2$ (a), and the oxidized form or the quinone form is FAD (c). Flavin adenine dinucleotide (FAD) can accept one or two hydrogen atoms. The addition of one electron together with a proton (i.e., a hydrogen atom) generates a semiquinone intermediate (b). The semiquinone is a free radical because it contains an unpaired electron (denoted by a blue dot), which is delocalized by resonance to all the flavin ring atoms. The addition of a second electron and proton (i.e., a second hydrogen atom) generates the reduced form, $FADH_2$.

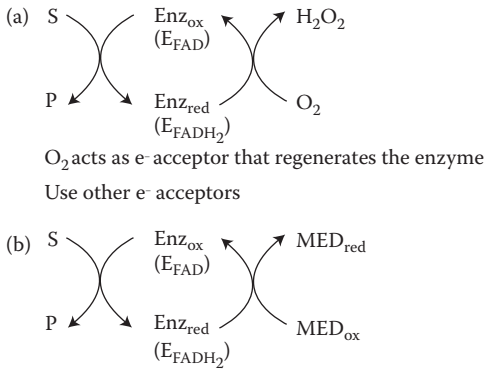


FIGURE 7.60 Glucose oxidase (GOx) reaction with its substrate (*S*) glucose in the presence of oxygen (a) and a mediator (b).

oxygen, which also makes electron transfer faster. Besides a higher current, another added benefit of these mediated enzyme-based sensors is that they are independent of the oxygen content in the sample, a factor that sometimes limits Reaction 7.5. During the regeneration of FAD, ferrocene (MED_{red}) is produced and can be measured by anodic oxidation. Alternatively, the decrease in ferrocinium (MED_{ox}) can be monitored by cathodic reduction. Medisense/Abbott uses a ferrocene derivative (carboxy ferrocene) in their commercial glucose sensor for diabetes management. This sensor is also used as an example of an amperometric device at the end of this section.

Third-generation glucose sensors were pioneered by Adam Heller and involve wiring the enzyme redox center directly to the electrode.[11,12] The $FADH_2$ redox centers of GOx can be electrically "wired" to an electrode by precipitation on the surface of thin films of the electrostatic adduct of GOx, which is a polyanion at neutral pH, and polycationic copolymers of vinyl ferrocene. This approach has the benefit of immobilizing the mediator molecules. The best approach, also invented by Heller, is currently the use of an electroactive enzyme-containing hydrogel. This is made by cross-linking nonprecipitating electrostatic adducts of polyanionic GOx and an excess of a polycationic osmium complex redox center containing polyvinyl pyridine-based redox polymer.[13,14] This cocktail creates a 3D matrix that is reversibly electroreduced and electrooxidized, extending electrochemistry from the 2D surface of the metal electrode to the 3D volume of the electron-conducting hydrogel. These wiring gels are highly permeable to facilitate the transport of reactants and products. In the volume of the hydrogel, electrons and holes are transported through collisions between Os^{2+} and Os^{3+} redox centers, which are tethered to the polyvinyl pyridine backbone, but with polymer segments that remain mobile within the polymer. A more recent optimized version of such a polymer, designed by Fei Mao,[15] is shown in Figure 7.61. In this example, the enzyme is wired using a polymer chain that folds along the enzyme structure.

The state-of-the-art glucose sensors are reviewed as the biggest bio-MEMS commercial success in Volume III, Chapter 10.

Semiconductor/Electrolyte and Insulator/Electrolyte Interface The semiconductor/electrolyte interface behaves significantly different than the metal/electrolyte interface. First, because there are many more free charge carriers in the solution than in the semiconductor, all potential changes in the electrical double layer occur on the semiconductor side, not in the electrolyte. Second, because of the

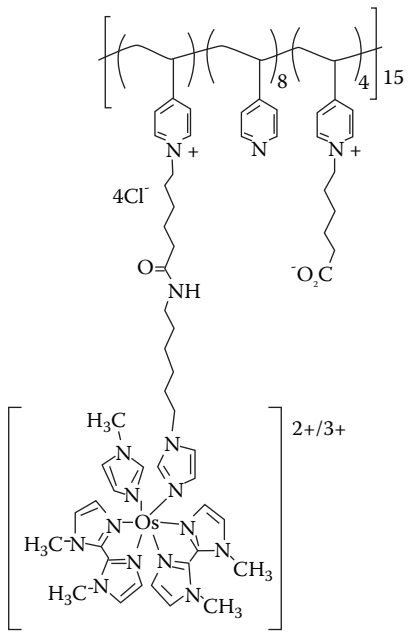


FIGURE 7.61 Os^{2+} and Os^{3+} redox centers are tethered to the polyvinyl pyridine backbone, but polymer segments remain mobile within the polymer. A more recently optimized version of such a polymer, designed by Fei Mao. (Mao, F., N. Mano, and A. Heller. 2003. *J Am Chem Soc* 125:4951.[15])

bandgap in the semiconductor, many electron transitions between redox levels and semiconductor are forbidden. In the case of an insulator/electrolyte interface, all electron transitions are forbidden. In Figure 7.62, we show what happens when we bring the same redox couple in contact with an n- and p-type semiconductor of the same material (e.g., *n*- and *p*-type Si). As in the case of the metal/electrolyte interface, equilibrium is reached when Fermi levels have lined up across the interface, which leads to a space-charge layer in the semiconductor. The reductant of a redox couple (filled electron levels) may inject electrons in the empty conduction band of the semiconductor if those energy levels overlap. Similarly, the oxidant may inject holes in the valence band if there is energy overlap in those bands.

A redox couple with energy levels that overlap only with the semiconductor bandgap does not lead to any current. Mathematically, this is expressed in the overlap integrals of Equations 7.113 and 7.114. In the case of semiconductors, there is a small range, known as the *forbidden bandgap*, where the overlap integral vanishes. In the case of an insulator that gap is large, and the overlap integral is always zero. To illustrate the previous points, consider Figure 7.63, where we show an n-type semiconductor/electrolyte contact for three different redox systems. For redox system 1, with its standard redox potential (E^0_{redox}) close to the conduction band edge, a high rate of electron exchange can be expected, and an accumulation layer forms in the semiconductor. For redox system 2, a depletion layer is formed in the semiconductor, and the electron exchange is expected to be low as the redox system significantly overlaps with the forbidden bandgap. In the case of redox system 3, the barrier height in the semiconductor is even higher (close to E_g, the bandgap), but (E^0_{redox}) is now close to the valence band edge and electron exchange is possible again. In the latter case, empty levels (holes) must be available in the valence band for electrons of reductant molecules to jump into. These holes can be generated, for example, by shining light on the semiconductor or by a strong oxidant. These holes react with the semiconductor itself and, in the presence of the appropriate electrolyte solution, may etch/corrode the semiconductor material. In Volume II, Chapter 4, in the section on wet bulk micromachining, we will expand on the concepts developed here to explain the difference between isotropic and anisotropic etching of silicon. We will also learn that current/overvoltage curves (i-η) on semiconductor/electrolyte contacts take on distinct diode characteristics. In liquid-junction solar cells, appropriate reductants compete effectively with the oxidation/corrosion of the semiconductor because they react more avidly with the holes in the valence band than the semiconductor material itself.[16]

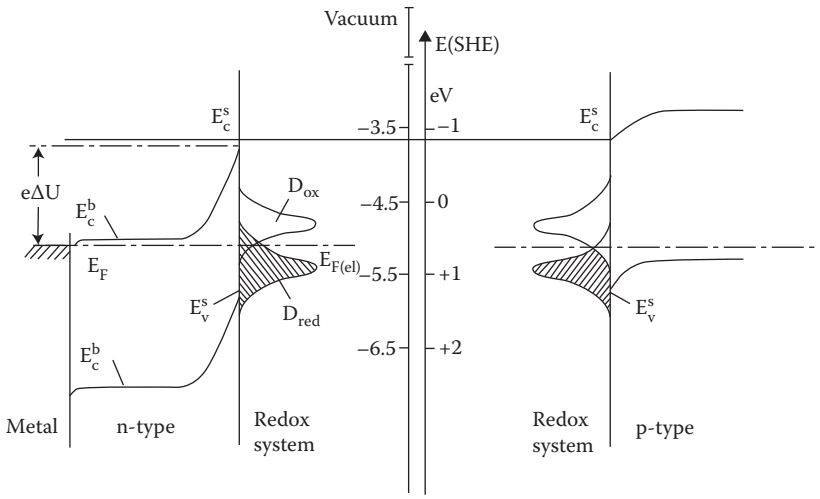


FIGURE 7.62 Energy diagram of the same redox couple in contact with *n*- and *p*-type Si (two energy scales are shown: electron in vacuum at infinity and SHE).

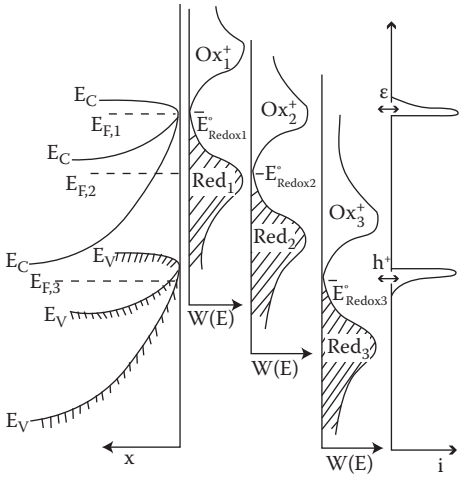


FIGURE 7.63 Energy diagram of three different redox couples in contact with n-type silicon.

For a certain value of the applied potential, there is equality between the number of electrons removed from and supplied to the semiconductor electrode. In this situation there will be no space-charge region in the semiconductor, and the potential is called the *flat band potential* E_{fb} (= E_F in the semiconductor). This potential can be established from impedance measurements on the semiconductor/electrolyte interface. From the flat band potential, the band edges of the semiconductor can be calculated if the dopant concentration is known (see Equations 4.38 and 4.39) (for details, see Madou and Morisson[15]). In Figure 7.64, we show a number of common semiconductor band edges and redox couples all referenced to the same reference electrode (REF E) on the right and with respect to an electron at infinity in vacuum on the left. Using this figure one can, in principle, predict the expected electron exchange for any semiconductor redox couple system by evaluating the overlap integral of energy levels on both sides of the interface (again using Equations 7.113 and 7.114). For example, in liquid-junction solar cells, for the reductant to compete effectively for holes in the valence band and avoid corrosion of the solar cell, one chooses a redox system where the maximum density of states of the reductant (the peak of the Gaussian) is at the same level as the valence band edge of the semiconductor.[15]

In a flat band situation, a potential drop remains on the solution side, and this remaining potential drop can be changed by ion adsorption on the semiconductor surface. Semiconductors in general develop a thin oxide coat on their surface. The hydroxyl groups at the surface of these oxide films will be more or less ionized, depending on the pH of the solution. This makes for a flat band potential,

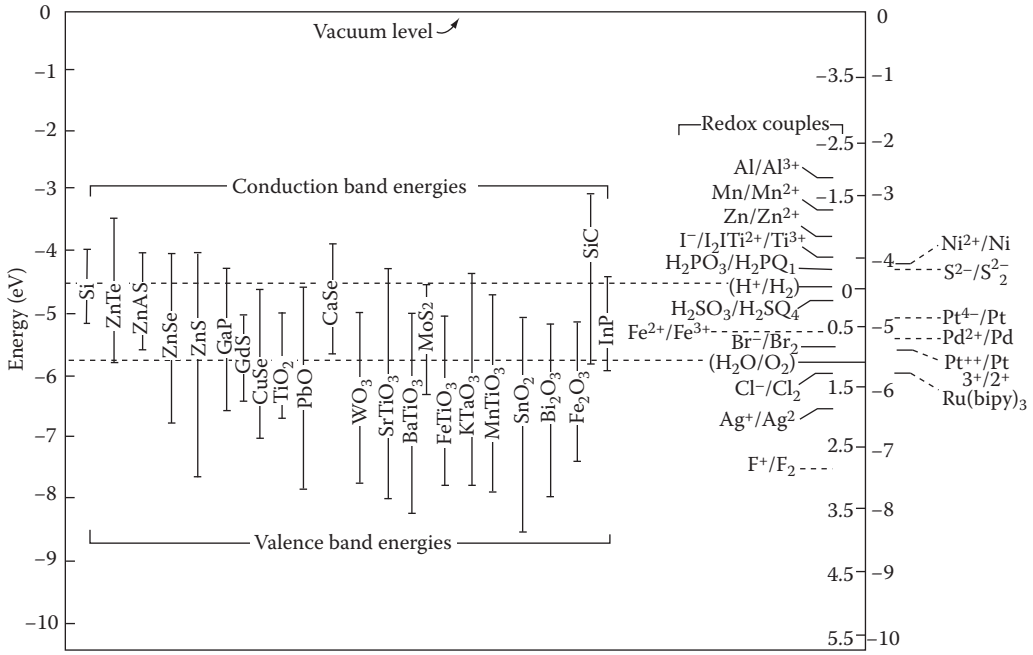


FIGURE 7.64 Relative energy levels of some common semiconductor and redox systems. The scale on the left is the energy with respect to an electron at infinity in vacuum. The scale on the right is the standard hydrogen electrode (SHE).

E_{fb}, that is a function of the pH of the solution (typically about 59 mV/pH unit change). If the redox system is not pH-dependent, this means that we can shift the semiconductor band edges with respect to the redox levels in solution. This is exploited in semiconductor electrochemistry to probe the density of energy levels of redox couples by changing the pH in small steps and observing the measured currents across the interface.[9]

With semiconductors, considerable electron transfer can only occur if the redox potential is located close to the band edges. In such a case, the electron equilibrium exchange current, $i_{0,e}$ is high but lower than for a metal. Any ion exchange processes, e.g., proton exchange, at the semiconductor surface would have difficulties competing with the electron exchange for establishing the open circuit voltage ($i_{0,e} >> i_{0,i}$). In principle, the pH dependence of the flat band potential of a semiconductor could be used as a good pH sensor, but only if the electron exchange current does not overwhelm the ion exchange current. In other words, in the presence of redox couples, a wide bandgap semiconductor or, better yet, an insulator, would make the better solid-state pH sensor. However, the pH sensitivity of an insulator normally cannot be measured because a thick insulator has too high of an impedance. Further below, we will see that the solution is to use an ion-sensitive field-effect transistor (ISFET), where the dielectric is so thin that a substantial electric field across it can influence the drain-source current of the transistor. These devices enable one to measure potentials on insulator/electrolyte contacts, ensuring that no electron exchange current ($i_{0,e}$) will interfere with the ion exchange current ($i_{0,i}$).

Membrane/Electrolyte and Mixed Conductor/Electrolyte Interfaces At an ionically conductive membrane/electrolyte interface, ideally only the fast exchange of one type of ion with the solution determines the open circuit potential. Because ions do not follow Fermi-Dirac statistics, the theory of ion transfer at the solid/liquid differs from that of electron exchange. Despite this obvious difference and the fact that ions must leave behind their hydration shell if they want to cross the phase boundary, the calculations for the equilibrium ion exchange current, $i_{0,i}$, and the polarization curves (i-η) lead to results very similar to those obtained for electron transfer at the metal/electrolyte interface. The ionic species that carries the greatest number of electric charges across the phase boundary is potential determining (largest $i_{0,i}$). For a "good" ion-selective membrane, the ion exchange current $i_{0,i}$ should be 10^{-5} A cm^{-2}. For membranes with an ion exchange current $i_{0,i} < 10^{-7}$ A cm^{-2}, mixed potentials are common. A Butler-Volmer-type expression for reactions under kinetic control and a diffusional expression similar to the ones derived for metals can be also derived for the membrane electrolyte interface but will not be developed here. For a more detailed treatise on ion exchange at solid/electrolyte interfaces consult Madou and Morrison[16] and Camman.[17]

We touched on mixed conductors, with ions and electrons as the charge transport species, when introducing IrO_x as a pH probe on page 557 "Electrodes of the Second Kind." Depending on the relative magnitude of $i_{0,e}$ versus $i_{0,i}$, a mixed conductor acts as a redox probe, a pH sensor, or both. For most mixed conductors $i_{0,e} >>> i_{0,i}$, one expects severe redox interference when used as pH sensors. As mentioned before, for reasons not well understood, from the noble metal mixed conductor oxide electrodes, IrO_x holds the most promise as a very fast, stable, low impedance pH sensor with surprisingly small redox interference.[8]

Cyclic Voltammetry and Ultramicroelectrodes

Cyclic Voltammetry

Earlier, we mentioned that the most important example of voltammetry is cyclic voltammetry, an electrochemical method in which information about an analyte is obtained by measuring current (i) as a function of an applied potential using a triangular waveform for excitation (Figure 7.40). We also explained the need of adding a high concentration of an inert supporting electrolyte to the solution (the large excess of inert ions ensures that few reactant and product ions will move as a result of migration). A typical, potentiostat-controlled, three-electrode system to carry out these experiments was shown in Figure 7.38.

Species in the electrolyte must be transported to and from the electrode before electrode reactions can occur. As species are being consumed or generated at the electrode surface via electrochemical reactions, the concentration of these species at the electrode surface will become smaller or larger, respectively, than in the bulk of the electrolyte. Suppose we are dealing with a redox species of concentration c_{bulk} in the bulk of the electrolyte ($x = \infty$) being consumed at the electrode. The concentration gradient at the electrode is then given as:

$$\frac{dc}{dx} = \frac{c_{bulk}(x=\infty) - c_{surface}(x=0)}{\delta} \qquad (7.121)$$

where $c_{surface}$ ($x = 0$) is the concentration of the redox species at the electrode surface, and δ represents the boundary layer thickness. From the thermodynamic relationship between the potential imposed on the metal electrode by the potentiostat and the concentration of the reacting redox species at its surface (i.e., the Nernst equation, Equation 7.56), we conclude that the concentration difference between surface and bulk leads to an overpotential:

$$\eta_c = E - E^0 = \frac{RT}{nF} \ln \frac{c_{surface}(x=0)}{c_{bulk}(x=\infty)} \qquad (7.122)$$

i.e., the expression for the concentration polarization, η_c, where n is the number of electrons involved in the reaction. For a planar electrode, which is much larger than the diffusion layer thickness, the measured current is given by Faraday's law (Equation 7.44):

$$i = nFAD_0 \left[\frac{c_{bulk}(x=\infty) - c_{surface}(x=0)}{\delta} \right] \qquad (7.123)$$

with D_0 ($cm^2\ s^{-1}$) representing the diffusion coefficient of the electroactive species and n the number of electrons transferred. At a certain overpotential, η_c, all of the reacting redox species arriving at the electrode are immediately consumed, and from that overpotential onward, the concentration of the electroactive ion at the surface decreases to 0, that is, $c_{surface}(x = 0) = 0$, and we reach the limiting current density i_l:

$$i_l = nFAD_0 \left[\frac{c_{bulk}(x=\infty)}{\delta} \right] \qquad (7.124)$$

The largest current (the diffusion-limited current) is established when every arriving species at the electrode is consumed immediately so that the surface concentration of it is zero, as illustrated in Figure 7.65. In the case shown we consider a cathodic reduction reaction (say, of $Fe^{3+} + e^- \rightarrow Fe^{2+}$, at a potential $E \ll E^0$).

Equation 7.124 shows that the limiting current is proportional to the bulk concentration of the ionic species. On this basis, classical amperometric (current-based) sensors are used as analytical devices (such as the amperometric glucose sensors we introduced earlier). We can rewrite $c_{surface}(x = 0)/c_{bulk}(x = \infty)$ in Equation 7.122 in terms of i and i_l to obtain:

$$i = i_l \left(1 - e^{\frac{nF\eta_c}{RT}} \right) \qquad (7.125)$$

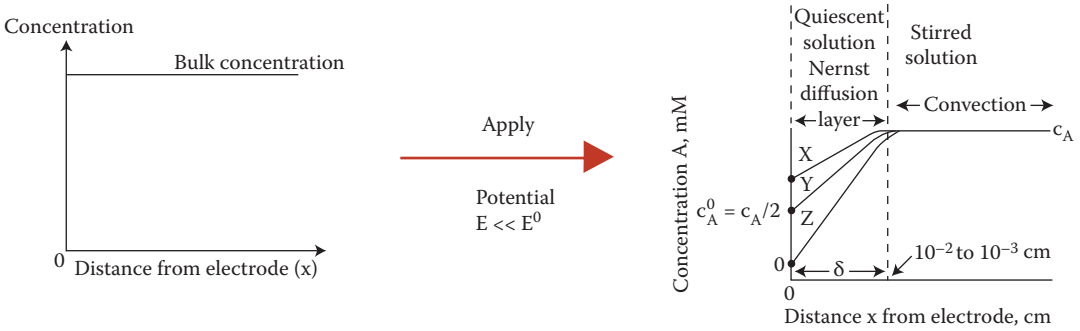


FIGURE 7.65 Applying a potential at a metal electrode that is sufficient to reduce a species in solution ($E \ll E^0$) sets up a concentration gradient that evolves over time (*X*, *Y*, *Z*). The concentration of c_A at various times after application of a voltage sufficient to consume c_A. At *Y*, the concentration of c_A has fallen to half of that in the bulk.

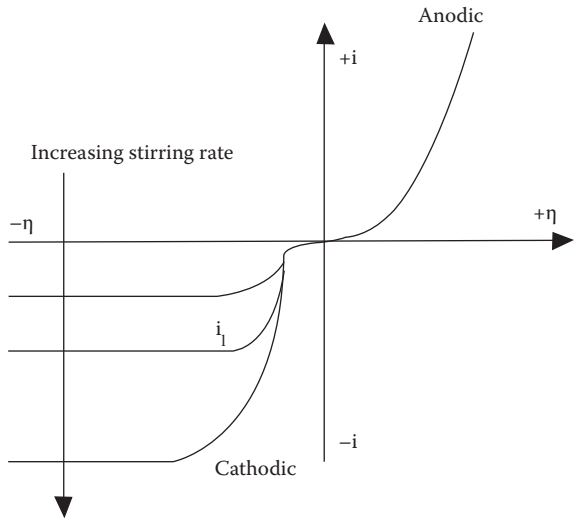


FIGURE 7.66 A diffusion-limited current is indicated for different stirring rates. The limiting current appears as a horizontal straight line limiting the current that can be achieved at a large value of the overpotential, whereas activation-controlled currents keep on increasing. This situation could come about when the concentration of oxidant is much lower than that of reductant.

This expression is illustrated for a cathodic diffusion-limited reaction in Figure 7.66. The cathodic current quickly becomes diffusion-limited, whereas the activation-controlled anodic current keeps on increasing. This situation could come about when there is very little oxidant (say, 10^{-4} M) but an abundance of reductant (say, 1 M), so that a diffusion plateau is not obtained in the anodic branch of the i-η curve.

The concentration overpotential and the activation overpotential have different origins. In the case of activation overpotential, the slope of i versus η increases with η, whereas in the transport-limiting case, the slope of i versus η decreases with increasing η and effectively becomes zero for sufficiently large values of the overpotential.

As shown in Figure 7.66, stirring of the electrolyte strongly affects the limiting current. Higher stirring rates promote convective transport of the ions toward the electrode and result in a smaller boundary layer δ.

For reversible systems, the i-E response shows an anodic and a cathodic plateau as illustrated in Figure 7.67a (assume that O and R are at the same concentrations). This can be ascertained by writing the surface concentration values for oxidant and reductant $c_{o,surface}(x = 0)$ and $c_{r,surface}(x = 0)$ in the Nernst equation in terms of i and i_l to obtain:

$$E = E^0 + \frac{RT}{nF}\ln\frac{c_{o,surface}\left(x=0\right)}{c_{r,surface}\left(x=0\right)}$$
$$= E^0 - \frac{RT}{nF}\ln\frac{D_0/\delta_0}{D_R/\delta_R} + \frac{RT}{nF}\ln\frac{i_l - i}{i}$$

or:

$$E = E_{\frac{1}{2}} + \frac{RT}{nF}\ln\frac{i_l - i}{i} \qquad (7.126)$$

The value of $E_{1/2}$ is the half-wave potential or the potential where the current is one-half the limiting current ($i_l/2$), and it is characteristic for the specific redox couple O/R only (qualitative analysis), whereas the height of the current plateau is proportional to the analyte concentration (quantitative analysis). In most cases, $E_{1/2}$ is the same as the reaction's standard state potential. A voltammetric wave and the half-wave potential $E_{1/2}$ are illustrated in Figure 7.67a. From Equation 7.126, a plot of E

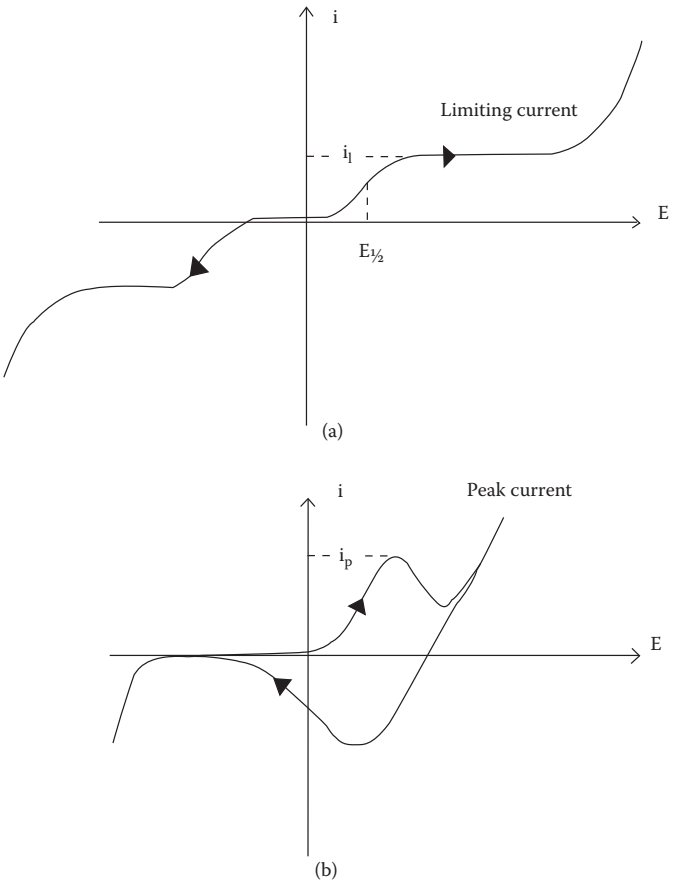


FIGURE 7.67 Cyclic voltammograms with slow (a) and fast (b) sweep rates.

versus ln $(i_l - i)/i$ yields a straight line with a slope of RT/nF (or $0.059/nV$ in a log plot) at room temperature and is used to determine the reversibility of a reaction.

The voltammetric wave in Figure 7.67a applies only when sweeping the voltage ($v = dE/dt$, where v is the voltage sweep rate) relatively slow. At higher sweep rates, cyclic voltammograms are obtained, as illustrated in Figures 7.67b and 7.68. When changes in the potential are made slowly, at such a rate that the rates of diffusion of oxidized and reduced species to and from the electrode can always keep the system in equilibrium with the bulk of the solution at all times (the steady-state value for δ is achieved in a short period compared with the measurement), the cyclic voltammogram exhibits a plateau corresponding to i_l as shown in Figure 7.67a. Most amperometric sensors work within this diffusion plateau, where operation is more forgiving for small potential drifts because the current is independent of the applied potential in that plateau. This also explains why commercial amperometric sensors mostly come without real reference electrodes (less expensive two-electrode system rather than the more complicated and more expensive three-electrode system): the correct knowledge of the voltage is not that important in a current plateau. The time necessary for diffusion to and from the electrode is crudely given by $t \approx \delta^2/2D$ (for a more correct expression, see Equation 7.128), where δ is the thickness of the diffusion layer. The diffusion layer δ will grow with $t^{1/2}$ so that a diffusion layer thickness of 10^{-4} cm is built up in 1 ms and of 10^{-2} cm in 10 s (using a typical value for the diffusion coefficent D of 5×10^{-6} cm^2s^{-1}). In the absence of convection δ keeps on increasing, but with convection present it finally approaches a steady-state value.

At fast sweep rates, the potential rapidly moves into the diffusion-limited regime. As a consequence, the concentration gradient is very large, and the initial current is very high. The rate of potential change is too rapid for diffusional processes to maintain equilibrium with the bulk of the solution. As a result, when sweeping the potential the current will not be maintained. Instead, it will peak and then decay as illustrated in Figure 7.67b. The peak value of the current, i_p, for a reversible reaction is given by the Randles-Sevcik equation:

$$i_p = 2.69\ 10^5 n^{\frac{3}{2}} A (Dv)^{\frac{1}{2}} c(x = \infty) \qquad (7.127)$$

where $c(x = \infty)$ is the bulk concentration, A is the electrode area, and v is the sweep rate (dE/dt). Comparing this expression for the peak current with the expression for the limiting current i_l in Equation 7.123, we see that the main difference is the appearance of the square-root sweep rate dependency.

A more detailed fast-sweep cyclic voltammogram is shown in Figure 7.68. In the forward direction, voltage window A to B, no electron transfer occurs. Between B and C, the voltage is sufficient for a reduction reaction to take place, and concentrations of reactants and products are given by the Nernst equation at the surface. At potential D, a maximum current is reached as diffusion limitation sets in. Because the solution

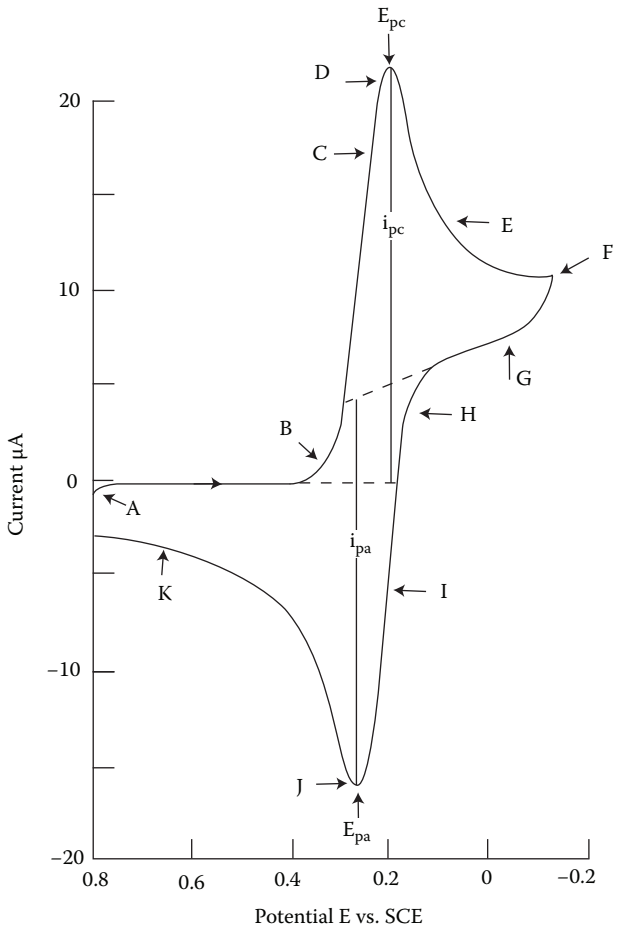


FIGURE 7.68 Fast-sweep cyclic voltammogram (*i-E*); see text.

is not stirred, δ increases with time, resulting in a decrease in current (D to E). At potential F we start the reverse scan, and we observe less current as the potential increases (F to G) until reduction no longer occurs. The reverse reaction (oxidation) then takes off at H to I (if we are dealing with a reversible reaction) and reaches a peak at J. The current decreases again because of the widening of the diffusion layer over time (J to K). A cyclic voltammogram provides a wealth of information. The cathodic (i_{pc}) and anodic peak (i_{pa}) currents at the cathodic (E_{pc}) and anodic peak (E_{pa}) potentials are all important parameters. For a reversible process, $E_{1/2} = E^0$ is the midpoint between E_{pc} and E_{pa}; also, $\Delta(E_{pa} - E_{pc}) = 0.0592/n$, independently of the sweep rate v (where n = number of electrons in reaction). For such a Nernstian process, the peak currents are proportional to the square root of the sweep rate (diffusional control) (see Equation 7.127) and $i_{pc}/i_{pa} = 1$. For a quasireversible reaction, $\Delta(E_{pa} - E_{pc}) > 0.0592/n$, which increases with increasing v. For an irreversible reaction, there is no return wave or the two waves do not overlap.

Higher sensitivities can be reached by increasing the sweep rate v (i.e., 10^{-7}–10^{-8} M; i.e., 10–100 times higher than with the slow sweep technique). This gain in sensitivity is limited by the capacitive, non-Faradaic current, because the capacitive current is linearly proportional to v [$i_C = C_H v$; see Equation 7.43, where v = dV/dt (V was used instead of E in this equation)] and increases faster than the Faradaic peak current. In the next section, we will see that the use of ultrasmall electrodes enables one to measure at higher sweep rates (as high as 20,000 Vs^{-1}) because the Faradaic current on those types of electrodes is relatively larger, allowing for more sensitive measurements.

Nonlinear Diffusion Effects on Ultramicroelectrodes

Electrochemical reactions are influenced by the electrode size relative to the thickness of the diffusion layer, δ. With microstructures of the dimensions of the diffusion layer, macroscale electrochemistry theory breaks down, and we can exploit some unexpected beneficial effects afforded through miniaturization. An ultramicroelectrode (UME) is an electrode having one of its dimensions (called the *critical dimension*) equal to or smaller than 20 μm.

For a small, spatially isolated electrode, with its size reduced to a range comparable with the thickness of the diffusion layer, nonlinear diffusion caused by curvature effects must be taken into account. In Figure 7.69a we illustrate various types of UMEs, including disc shaped (most common are planar or recessed), spherical (drop), band, ring, array, and so on. The diffusional fields for various types of microelectrode shapes and a semi-infinite planar electrode are shown in Figure 7.69b. The analysis shows that as the curvature effects become more pronounced, more diffusion of ions from all

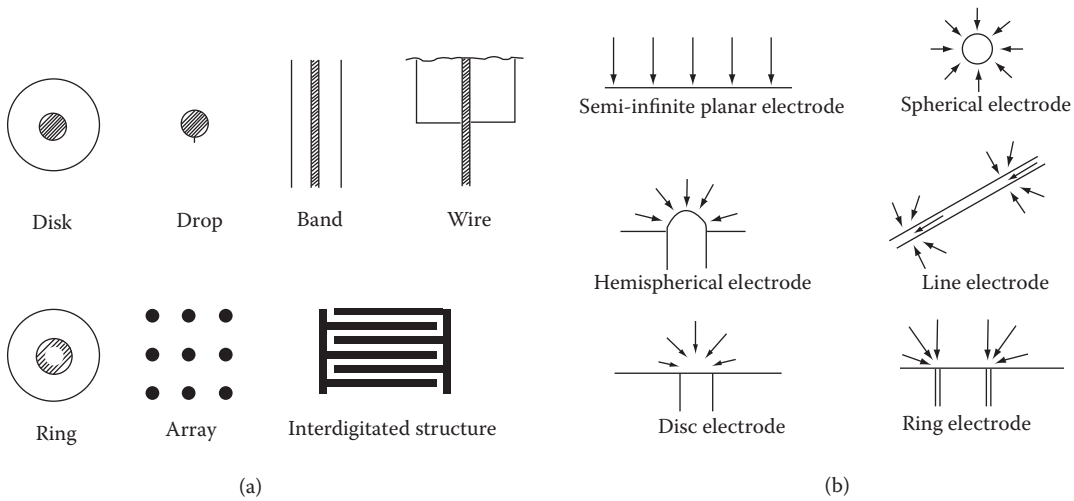


FIGURE 7.69 (a) Examples of types of ultramicroelectrodes (UMEs). (b) Diffusional fields around a semi-infinite planar electrode and various UMEs.

directions takes place, thus increasing the supply of reactants to the electrode.

The diffusion layer thickness arising from linear diffusion is time-dependent and given by [16,18]:

$$\delta = \left(\pi D_0 t\right)^{\frac{1}{2}} \tag{7.128}$$

δ is essentially the distance moved by a molecule during the experiment. Substituting Equation 7.128 into Equation 7.124, we obtain the *Cottrell equation for planar diffusion*:

$$i_1 = nFAc\left(x=\infty\right)\left(\frac{D_0}{\pi t}\right)^{\frac{1}{2}} \tag{7.129}$$

This equation represents the current-versus-time response on an electrode after application of a potential step sufficiently large to cause the surface concentration of electroactive species to reach zero. This equation, at short times after the potential step application, is appropriate regardless of electrode geometry and rate of solution stirring as long as the diffusion layer thickness is much less than the hydrodynamic boundary layer thickness.* Nonlinear diffusion at the edges of microelectrodes results in deviation from the simple Cottrell equation at longer times. The total current time relation with the correction term is given as:

$$i_1 = \frac{nFAD_0^{\frac{1}{2}}c\left(x=\infty\right)}{\pi^{\frac{1}{2}}t^{\frac{1}{2}}} + \frac{nFAD_0c\left(x=\infty\right)}{r} \tag{7.130}$$

At short times, when the diffusion layer is thin compared with radius *r*, Equation 7.129 holds. At longer times and for small electrodes, Equation 7.130 predicts that the correction term can become significant. In practice, a study on ultrasmall electrodes often starts with a linear regression of the measured current versus $t^{-1/2}$ after application of a potential step. The intercept gives the steady-state term. This term, the diffusion-limited current, $i_{l,m}$, at sufficiently long times on UMEs for some important electrode shapes, is given by[18]:

$$\begin{aligned} i_{l,m} &= \pi rnFD_0c_{bulk}\left(x=\infty\right) && \text{DISC} \\ i_{l,m} &= 2\pi rnFD_0c_{bulk}\left(x=\infty\right) && \text{HEMISPERE} \\ i_{l,m} &= 4\pi rnFD_0c_{bulk}\left(x=\infty\right) && \text{SPHERE} \end{aligned} \tag{7.131}$$

In the correction term of Equation 7.130, the electrode surface area *A* is divided by the radius *r*. Hence the principal location of charge transfer appears to be on the outer edge of the electrode. This constitutes a very favorable scaling. Phenomena scaling with the linear dimension to the power of one becomes important in the microdomain. Surface tension exemplifies another scaling law with the power of one as a very dominant force in the microdomain. In the current case, $i_{l,m}$ is proportional to the electrode radius ($\sim r^1$), whereas the background current i_c (associated with the charging current of the Helmholtz capacitance) is proportional to the area ($\sim r^2$); the ratio of the Faradaic current (the charging current) to background currents should increase with decreasing electrode radius (l/*r*). For electrodeposition, this means that smaller features will be plated faster; for sensor applications, this translates into a higher S/N ratio or improved sensitivity. In the previous section, we saw that scanning the potential faster (v = dE/dt) leads to higher sensitivities (see Equation 7.127), limited by the even faster increase in capacitive current. We recognize here that the capacitive contribution of a microelectrode is relatively small, so faster scan rates are possible. The latter makes amperometric sensing with microelectrodes possible at unprecedented sensitivities.

With a single microelectrode, the analytical current remains small, and the analytical gain is relative. However, using an array of microelectrodes (see example in Figure 7.69) with all electrodes connected in parallel and separated by an insulating layer provides an elegant solution to this problem.

* In diffusion the parameter *D* is analogous to the kinematic viscosity ν_k in convection. These parameters are important in determining the thickness of the diffusion layer and the hydrodynamic boundary layer, respectively. For aqueous solutions, the diffusion coefficient is typically 1000 times smaller than the kinematic viscosity ($D = 10^{-5}$ $cm^2\,s^{-1}$ and $\nu_k = 10^{-2}$ $cm^2\,s^{-1}$). It can be shown that the relation between δ, the thickness of a steady-state diffusion layer, and δ_h, the thickness of the stagnant layer on an electrode in a stirred solution, is $\delta = (D/\nu_k)^{1/2}\delta_h$. For $D = 10^{-5}$ $cm^2\,s^{-1}$ and $\nu_k = 10^{-2}$ $cm^2\,s^{-1}$, $\delta = 0.1\,\delta_h$. Thus, the thickness of the diffusion boundary layer, δ, is considerably smaller than that of the hydrodynamic boundary layer. In an unstirred solution, δ is not well defined, and all types of disturbances can affect the transport. To prevent random convective motion from affecting transport to and from the sensing electrode, we want the diffusion layer thickness to be smaller than the hydrodynamic boundary layer and δ_h to be regular.

In arrays of ultrasmall electrodes, the electrodes behave as an equivalent number of individual microelectrodes when the spacing between them is sufficiently large. This configuration enables analytical currents in a higher range compared with the single-electrode case. This gain is temporary because the diffusion layers around the individual microelectrodes expand across the insulating surface at a rate given by Equation 7.128. The time at which the overlap occurs is both a function of electrode size and spacing in the array, resulting in a decrease in the steady-state current. In an analytical experiment, short compared with the time frame of overlap, more sensitive measurements at reasonably high currents become possible. Another gain, besides higher currents and improved S/N ratio through higher mass transport, results from using an array because the signal can be averaged over many electrodes in parallel (an $\sqrt{n}$ improvement in S/N, with n the number of electrodes). An example of an array of UMEs based on carbon nanotubes is shown in Figure 7.70.[19]

Advantages of using UMEs include:

- High mass transfer rates at ultrasmall electrodes make it possible to experiment with shorter time scales (faster kinetics).
- An array of closely spaced ultrasmall electrodes can collect electrogenerated species from adjacent electrodes with very high efficiency.
- At high scan rates, greater than 500 V/s, the data obtained from conventional electrodes are useless because they are impossible to interpret as a result of growing *iR* drop and increasing influence of charging current. With UMEs, because of reduction in size and hence the current, the *RC* and *iR* effect are greatly reduced so that UMEs can be used for fast-scan voltammetry up to 10^5–10^6 V/s.
- Electrochemical measurements in high-resistivity media, including possibly in air, become feasible.[20]
- Microelectrodes can be used as probes for local environment and processes, in which the small physical size of the electrode is important.
- The steady-state diffusion layer thickness is equal to the electrode dimension. Hence, any process with a larger characteristic dimension, such as convection, will not influence the steady-state current. This means that these electrodes can be used independent of the flow characteristics of the medium in which they are used.
- Another important analytical advantage is that greatly reduced background signals afford a greater S/N.

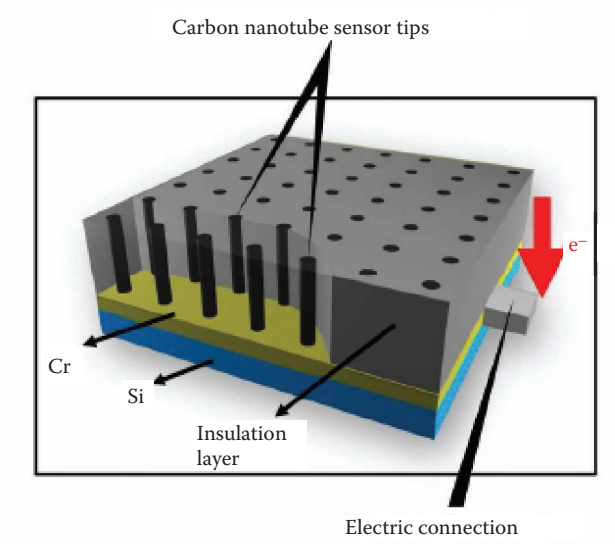


FIGURE 7.70 Arrays of CNT ultramicroelectrodes. (Lin, Y.H, W. Yantasee, F. Lu, J. Wang, M. Musameh, Y. Tu, and Z.F. Ren. 2004. *Dekker encyclopedia of nanoscience and nanotechnology.* Eds. J.A. Schwartz, C. Contescu, and K. Putyera. New York: Taylor & Francis.[19])

For example, microarray electrochemical detectors were applied in liquid chromatography for the detection of carbamate pesticides in river water.[21] The detection limits, in the subnanogram range (50–430 pg), represent a 60-fold improvement over other reported liquid chromatography detectors such as fluorescence detection and electrochemical detection with a single macroelectrode. In this early work, the electrochemical detector array consisted of a mixture of graphite and Kel-F, the so-called *Kelgraf electrode* with a microarray-like structure. Platinum microelectrodes have been microfabricated as detector electrodes for capillary electrophoresis.[22]

Micromachining small reaction chambers in which L, the outer boundary of the electrochemical cell, is made smaller than the diffusion layer thickness, δ, also increases the sensitivity of an

electrochemical sensor. In Equation 7.124, δ is now replaced by a smaller L, resulting in a higher diffusion-limited current i_l. The latter finds its application in the design of thin-layer, channel-type flow cells.

When considering a small metal disc embedded in a photoresist, nonlinear diffusion can be neglected for large electrodes (r large or $i_{l,m}$ small). If r becomes small (i.e., of the dimension of and smaller than δ), the contribution of $i_{l,m}$ to the total current becomes larger than i_l; that is, an enhanced mass transport to small electrodes occurs compared with large area electrodes. As pointed out above, we can expect that nonlinear behavior increases the plating rate of small features over large features.

But another factor influences the deposition rate on a small electrode, i.e., the level or position of the conductor surface with respect to surrounding insulating surfaces. The correction term in Equation 7.130 was derived for an inlaid microdisc electrode for which the electrochemically active substrate coincides with the dielectric insulation level (Figure 7.71a). If the conductor disc is recessed within an insulating medium (Figure 7.71b), the diffusion inside the hole must be considered as well, leading to a slight modification of the correction term $i_{l,m}$:

$$i_{l,m} = \frac{nFAD_0 c(x=\infty)}{r+L} \qquad (7.132)$$

in which L stands for the depth of the recession. Mass transport is only enhanced compared with

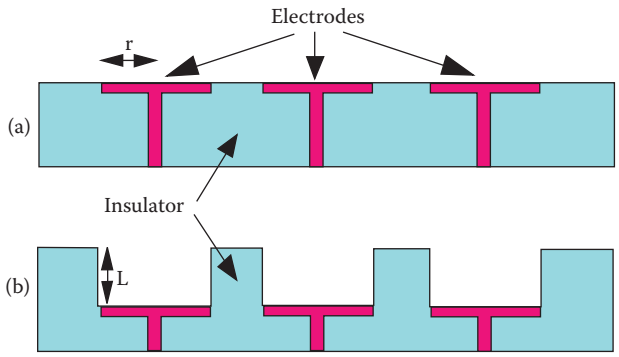


FIGURE 7.71 Nonlinear diffusion on inlaid and recessed electrodes. (a) Inlaid microelectrodes. (b) Recessed microelectrodes.

large area electrodes for small values of L. Van der Putten et al.[23] found that the metal deposition rate at smaller recessed areas (with $r \le L$) was greater than on larger areas because overfilling was obtained faster for smaller recessed features. When plating features in the LIGA process (see Volume II, Chapter 10), we may deal with situations where a metal layer is deeply recessed ($r \ll L$) in a layer of insulating photoresist. As very high-aspect-ratio features (L/r large) can be made with LIGA, the thickness of the diffusion layers is increased artificially by the microstructured photoresist layer because it is more difficult for convective flow to reach the bottom of the deep gaps.

A microelectrode of the same size or smaller than the diffusion layer thickness could be expected to plate faster than a larger electrode because of the extra increment of current resulting from the nonlinear diffusion contribution (Equation 7.130). On the other hand, as derived from Equation 7.132, describing the current to a metal electrode recessed in a resist layer, the nonlinear diffusion contribution increases with decreasing radius r but decreases with increasing resist thickness, L. High-aspect-ratio features consequently will plate slower than low-aspect-ratio features. Moreover, the consumption of hydrogen ions in the high-aspect-ratio features during electroplating causes the pH to increase locally. Because no intense agitation is possible in these crevices, an isolating layer of metal hydroxide might form, preventing further metal deposition. This all contributes to make the deposition rate, important for economical production, much slower than the rates expected from the linear diffusion model for the current density in large, low-aspect-ratio structures.

The treatise of electroplating in micropatterned resists is important for both micromachining and the IC industry. We cover the topic of electroplating in microstructures in Volume II, Chapter 8 on electrochemical-based forming. Additional reading on this topic is suggested (e.g., Masuko et al.,[24] Dukovic,[25] and Leyendecker et al.[26]).

Scaling in Electrochemistry/Summary

Potentiometric devices such as ISEs or ISFETs (see examples in the following section) measure a

voltage and are scaling invariant (voltage is an intrinsic parameter). Amperometric devices, on the other hand, measure a current and are scaling-sensitive (current is an extrinsic parameter). Miniaturization efforts have gone predominantly into the miniaturization of potentiometric devices, although more benefits can be derived from miniaturizing amperometric ones.[16] Scaling of amperometric detector electrodes was addressed in the preceding section. Here we summarize several benefits that derive from miniaturization of amperometric detectors, including higher sensitivity and the possibility of measuring in solutions with higher resistivity. Closely spaced ultramicroelectrodes (UMEs) are highly efficient in collecting electrogenerated species, and the high mass transfer rate makes it possible to experiment with shorter time scales. Nonlinear diffusion effects in amperometric microelectrodes lead to improved sensitivities for sensors with electrode dimensions of a size comparable to the diffusion layer thickness. To compensate for the decrease of the overall absolute current level, an array of microelectrodes is used. Using an array of microelectrodes in a thin, chamber-like configuration could further enhance sensitivity.

Other electrochemical devices such as batteries, as we discuss in Volume III, Chapter 9 on power in miniature devices, scale disadvantageously as the battery capacity is based on its volume.

Example Electroanalytical Devices

Potentiometric Sensors

Ion-Sensitive Field-Effect Transistors, Extended Gate Field-Effect Transistors, and Isn't FET The cable (signal line) connecting an ion-sensitive electrode (ISE) to a high-impedance voltmeter is typically about 1 m long. Because the electrical impedance of a sensor goes up as its size goes down, noise for MEMS- and NEMS-based ISEs with long signal lines is often intolerably high (this is a special problem in hospital settings that constitute an electronically very noisy environment). Integrating the ISE membrane directly onto the gate of a metal oxide semiconductor field-effect transistor (MOSFET) reduces the signal line to practically zero. This arrangement is known as an *ion-sensitive field-effect transistor* (ISFET). Pieter Bergveld invented the ISFET in 1970 at the University of Twente, Enschede, the Netherlands.[27] The operation of ISFETs, also called *CHEMFETS* (chemically sensitive FETs), and the technical and fundamental problems associated with their construction and use, were treated in detail by Madou et al.[16] and will only be summarized here. The structure of the ISFET is essentially a MOS transistor rendered ion-sensitive by eliminating its metal gate electrode to expose the gate insulator either directly to a solution or by putting an ISE membrane on top of the dielectric to interface with the solution (Figure 7.72). In the MOSFET, the metallic gate electrode controls the drain-source current, I_{ds}, through the external voltage, V_{gs}. The metallic gate electrode is insulated from both drain and source by a thin silicon dioxide layer and controls the magnitude of drain-source current electrostatically (see Chapter 4). The only difference between the electrical circuits of a MOSFET and an ISFET is the replacement of the metal gate electrode of the MOS transistor by a series combination of a reference electrode, electrolyte, and a chemically sensitive gate insulator or gate with an ion-selective membrane on top (Figure 7.73). As we saw earlier, insulators do not support any electron exchange at the interface with the solution, but hydroxyl groups at the surface ionize and are charged differently depending on the value of the pH of the surrounding solution. This electrostatic charge influences the drain-source current of the transistor.

The pH sensitivity of an insulator normally cannot be measured because a thick insulator has too high an impedance. A voltmeter, measuring the voltage between an insulator and a reference electrode, would go in overload. Typical dielectric gate materials used as pH sensors are very thin films of Al_2O_3, Ta_2O_5, and Si_3N_4 (even this dielectric has plenty of –OH surface groups in an aqueous medium). With these pH-sensitive thin-film dielectric gate materials immersed in a solution, the hydrogen ion concentration in solution influences the drain-source current, I_{ds}. The reason is that the dielectric is so thin that a substantial electric field across it can influence the magnitude of the drain-source current just as a gate voltage on the metal electrode in a MOSFET would. In an actual ISFET, the drain current value is kept

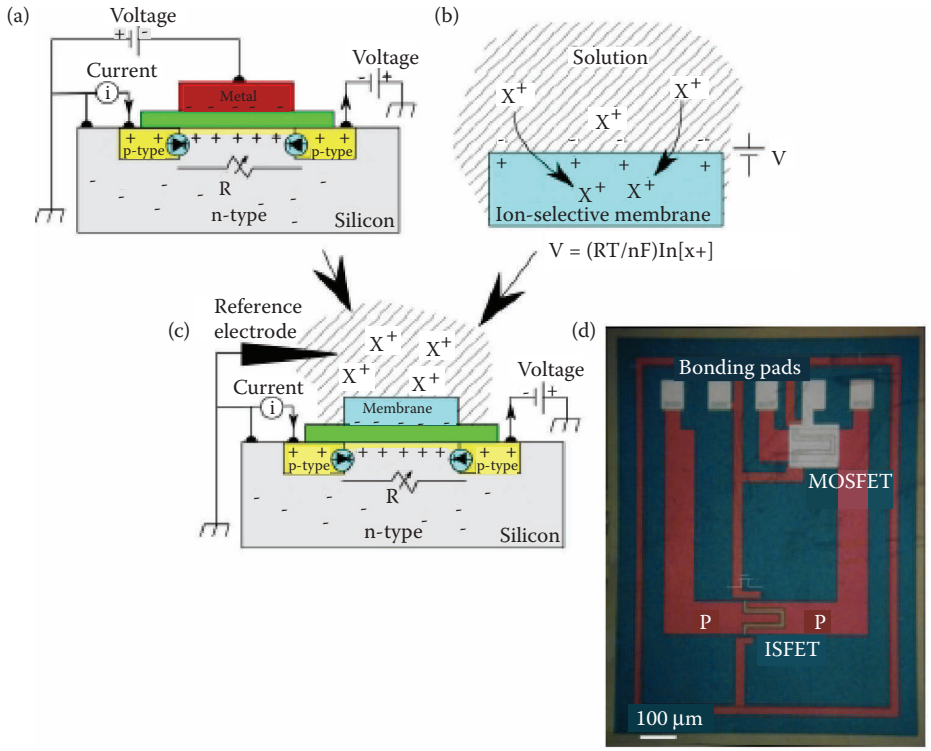


FIGURE 7.72 (a) Classical MOSFET. (b) Ion-selective membrane (ISE) for cation X^+ produces a Nernstian response. (c) Metal gate is replaced by the ion-selective membrane. (d) ISFET device (reference electrode not shown). A differential measurement between the ISFET and an on-board MOSFET helps reduce drift.

constant by means of electronic feedback, and if the pH value of the solution changes, the potential of the gate changes, and, therefore, also the drain current.

Besides using gate dielectrics directly, any of the membranes discussed above may be deposited on top of the gate dielectric.

Three important lessons were learned from working with ISFETs.

1. NEED FOR A REFERENCE ELECTRODE. In the early ISFET days, it was erroneously believed that an ISFET could work without the reference electrode (REF E). Early devices had been found to operate without them. However, it was later discovered that this was an effect caused by leaky SiO_2 gate oxides, and that provided a current return path, which, for reliable operation, should be provided via the REF E.[16] Although a FET device relates a signal through a dielectric via an electrical field, the countercharge of the charge built up in front of the dielectric needs to arrive somehow to the other plate of the capacitor, and this is only possible if a current return path is provided.
2. DEBYE SHIELDING. If a charged group is attached to a long chain (spacer arm) extending from the gate insulator, say, at a distance d into the solution, that charge may be neutralized by counterions in solution, diminishing the surface potential. The higher the ionic strength of the solution, the more the charge will be shielded. This is illustrated in Figure 7.74, where we consider a film with 4′-nitrobenzo-18-crown-6 ether groups on spacer arms. Calculated curves

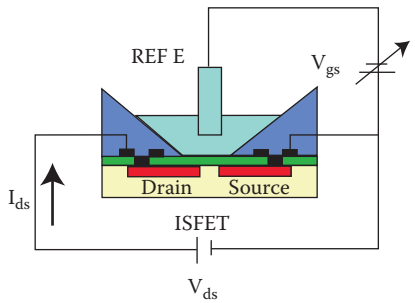


FIGURE 7.73 ISFET operation. Ions do not have to pass through the ion-sensitive membrane. The membrane's control of the drain-source current is solely based on electrostatic effects.

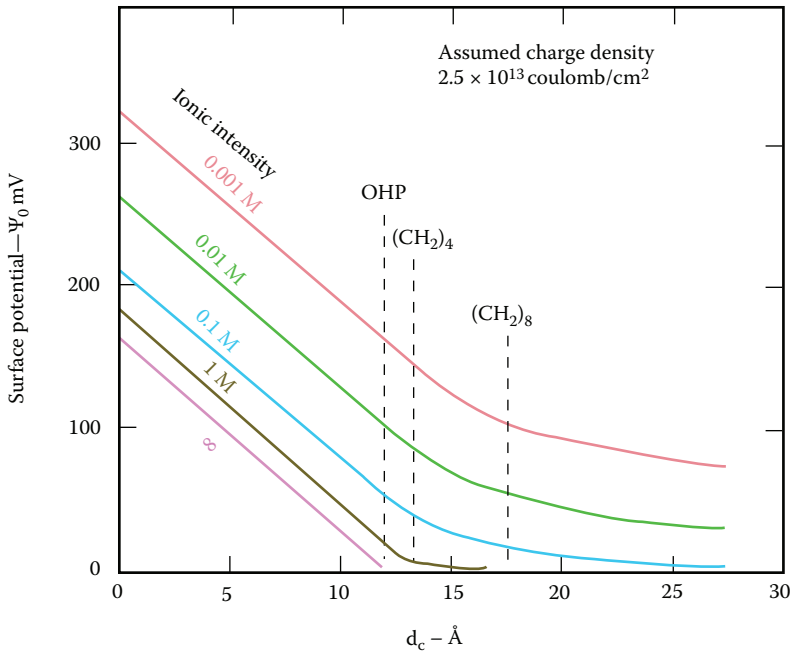


FIGURE 7.74 Calculated surface potential versus chain length d_c. Charges on spacer arms are shielded by ions from solution. This might make the signal of a large charged molecule ionic strength-dependent. (Matsuo, T., and H. Nakajima. 1984. Characteristics of reference electrodes using a polymer gate, ISFET. *Sensors Actuators* 5:293.[28])

of surface potential versus spacer chain length (d_c) by Matsuo and Nakajima[28] are shown. It can be seen that at the highest concentration of the indifferent electrolyte in the solution (∞), the surface potential becomes zero when charged centers extend all the way to the OHP or beyond. This observation is extremely important for the construction of derivatized FET devices. Obviously, sensitivity will be lost if the charge groups are big; sensitivity will depend on the ionic strength of the analyte solution. Attempts to build immunologically sensitive FET devices are hampered by this effect. Moreover, smaller charged molecules can go between the large forest of proteins on the surface and set up a mixed potential at the FET surface. Furthermore, the nonavailability of perfectly polarizable electrodes, nonselective protein interactions with the electrode, and the influence of pH and ionic strength (through Debye shielding) all make the prospect of developing selective and sensitive direct potentiometric or capacitive immunosensors rather dim.

3. IC COMPATIBILITY AND DRIFT. After more than 37 years, ISFETs are still plagued with at least four critical problems. 1) Putting a very high-impedance FET in a conductive solution requires tremendous care in packaging the device; 2) implementing a true reference electrode with a Luggin capillary and a sufficient amount of REF E to survive for an extended period, is very difficult to implement in a CMOS-compatible technology; 3) even a modest-size REF E renders the ISFET large, and the hope for IC miniaturization benefits are lost; and 4) the membranes in an ISFET have no internal electrolyte to fix the Donnan potential on the inside of the membrane. As a consequence, the drift of an ISFET is typically large compared with that of an ISE. A differential measurement between the ISFET and an on-board MOSFET as illustrated in Figure 7.72d does not help reduce this drift source. The on-board reference MOSFET only reduces solid-state-based drift sources.

Most sensor experts have come to appreciate that for chemical sensors, one should only implement on-board electronics if it is absolutely essential for the integrity of the signal. The extended

gate field effect transistor (EGFET) device was conceived to avoid some of the problems associated with the leakage problems of an ISFET.[29] By depositing the ISE membranes a short distance away on conducting lines extending from the FET gates, encapsulation is made easier, and the sensors are light-insensitive. The EGFET seemed like an elegant solution at the time (Figure 7.75), but i-STAT (now Abbott), a commercial entity selling blood electrolyte analysis equipment, eliminated all on-board electronics and abandoned the EGFET approach after extensive research.

For multiuse electrochemical reference electrodes, thick hydrogel films with a large capacity to store reference solution are preferred for long-term stability, and for electrochemical sensor electrodes, thick ion-selective membranes are better at preventing pinholes. Consequently, optimized chemical sensor materials end up much thicker than the common IC electronics layers and are more easily prepared with thick-film technology. An additional reason for the slow commercialization of silicon-based electrochemical sensors is that the currently available electrochemical and chemical sensing technologies are multiuse, proven, and cost-effective. Usually, it is simpler and less expensive to build electrochemistry on a nonsilicon substrate (see Volume III, Figure 1.19).

NOTE: During a 37-year period, we went from a long-signal line (e.g., 1 m long) of a large-sized ISE to a zero signal line in an ISFET. In an EGFET, the signal line was extended back to about 100 μm, but with the latest hybrid ISEs, built on plastic substrates, the signal lengths are again from a couple of millimeters to 1 cm long. In other words, we went from ISEs to ISFETs to something that is not even a FET anymore (Isn't FET!). Hopefully, this lesson will be taken at heart in the development of nano-FETs and nanowires.

Potentiometric Immunosensors

From Immunoassays to Immunosensors Immunoassays, introduced in the late 1950s, constitute a major class of medical diagnostic tests that, along with DNA probes, have become an established tool

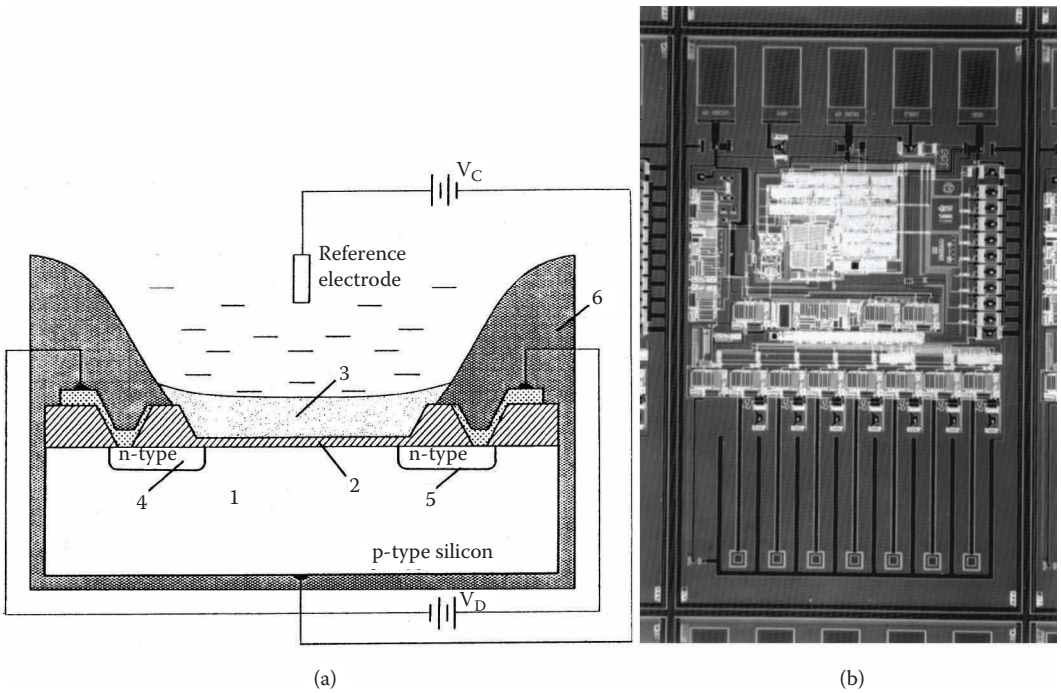


FIGURE 7.75 Integration of chemical-sensitive material and FETs. (a) Chemistry layers on the gate of the FET: ion-sensitive field effect transistor (ISFET). Si substrate (1), insulator (2), chemically sensitive membrane (3), source (4), drain (5), and insulating encapsulant (6). (b) Chemistry layers extended away from the FET: extended gate field effect transistor (EGFET). The seven small squares at the bottom of the photo are the extended gates. The rectangular pads on top are the contact pads.

in analyzing clinically important analytes. Immunoassays are used to detect hormones, vitamins, metabolites, diagnostic markers for therapeutic drug monitoring, and diagnostic procedures for detecting infections. Clinical analytes, as we can learn from Volume III, Figure 10.49, are often present in extremely low concentrations (e.g., 0.1 ng/mL).

The immunoassay principle is based on the use of antibodies (Ab) and a signal-generating irreversible chemical reaction of an antibody with an antigen (Ag, the analyte). The irreversibility of the reaction stems from the large association constants (*Ka*) involved in the antigen-antibody binding; typical values range between 10^5 and 10^9 M^{-1}. The antigen-antibody association constants are composed of large forward (k_1) and small reverse (k_{-1}) rates. These kinetic parameters make antibodies very selective for the analyte of interest but also make the binding irreversible. In fact, "reversible" immunosensors have been developed with lower *Ka*'s to allow for repeated measurements. The ability of natural cells to monitor biochemical events is mimicked in immunoassays. Therefore, these assays constitute the ultimate in selectivity and sensitivity. The utility of antibodies for detecting and isolating antigens can hardly be overstated, given the wide spectrum of extremely powerful antibody applications developed over the years. The use of antibodies is further enhanced by their relative stability in various chemical modification reactions that alter the antibody structure without destroying their capacity to bind to antigens. The two major configurations for immunoassays are 1) solid-phase/heterogeneous assays (often in microtiter plate wells) that require immobilization and separation steps, and 2) solution-phase/homogeneous assays that only involve mixing and measuring.

The binding event for an immunoassay may be sensed directly or via a label; the latter is the more common approach. For example, in an enzyme-linked immunosorbent assay (ELISA), developed in 1971, immunochemical affinity is relied on for selectivity and chemical amplification of an enzyme label for sensitivity. In an ELISA, the enzyme label on the bound Ab-Ag, in the presence of its substrate, generates a reaction product that is measurable (e.g., with GOx as the enzyme and glucose as its substrate, hydrogen peroxide is produced, which may be detected electrochemically). In a typical procedure, titer wells are coated with antibody and known quantities of antigen-containing patient sample, and labeled antigens (e.g., with GOx enzyme) are incubated in the titer wells until the antigen-antibody reaction is complete. Washing removes the unbound antigens, and adding substrate (glucose in the example case) to the bound antigen leads to an electrochemical active reaction product (H_2O_2 in this case). A crucial step is the removal of unbound and nonselectively bound antigens. This is achieved by binding the antibodies to a solid surface, such as a microtiter well surface. Antigens bound to the fixed antibodies remain stuck to the solid surface, whereas washing and decanting removes unbound and nonselective bound antigens. One appreciates that one can either detect the antigen or the antibody in an immunoassay.

The assay illustrated in Figure 7.76 is a competitive electrochemical assay, with native antigen

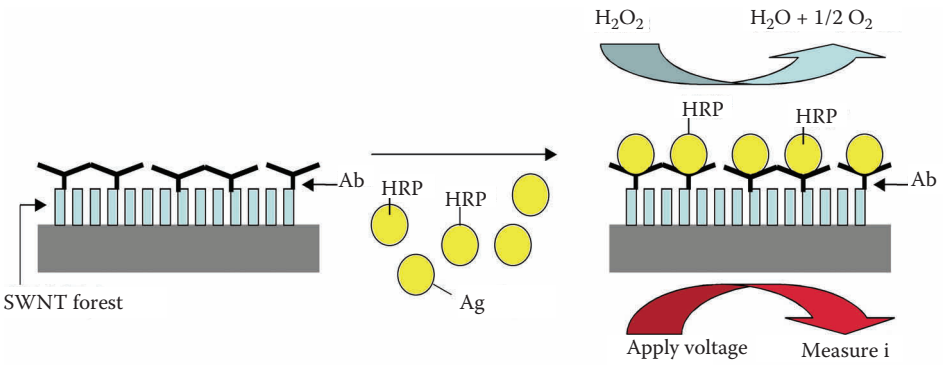


FIGURE 7.76 Enzyme-linked immunosorbent assay (ELISA). Antibody is bound to a forest of single-wall nanotubes. The current is inversely proportional to the amount of nonlabeled antigen (Ag; i.e., native antigen); antigen was prebound on antibody (Ab). Applying a voltage enables one to measure a current proportional to the amount of oxygen.

competing with labeled antigen for reaction sites on a solid surface. The solid surface in the example consists of a forest of single-wall carbon nanotubes (SWNTs) with antibody bound to it. An electrode of SWNTs has the potential of leading to higher sensitivity. The label on the antigen we use in this example of an ELISA is the enzyme horseradish peroxidase (HRP), which converts peroxide to oxygen and water. After competitive binding of the native and labeled antigen on the solid surface and washing off of unbound antigen, the substrate of the HRP enzyme, hydrogen peroxide, is added. The HRP converts hydrogen peroxide to oxygen, which diffuses to the SWNT electrode surface. Current *i* measures the amount of oxygen at the electrode with an applied potential E. The more native species (analyte in the patient's blood), the more effective the competition with labeled species and the lower the resulting signal (less oxygen means less native antigen). Frustratingly, this means that the less native species (analyte), the higher the signal, and vice versa.

In Figure 7.77, an alternative noncompetitive or sandwich assay is illustrated we again use an electrochemical antigen-measuring system, for example). The SWNT forest is coated first with a suitable antibody (Ab1), and a patient sample containing the antigen (Ag; analyte) is then added and incubated until the antigen-antibody reaction is complete. A washing step removes unbound antigen, and a second antibody (Ab2) labeled with enzyme is added and incubated until the antigen binds with this second antibody. A new wash removes the unbound labeled antibody, whereas adding substrate for the enzyme (hydrogen peroxide) leads again to a measurable reaction product.

Antibodies/antigens have also been chemically tagged with fluorescent, magnetic, radioactive (e.g., ^{125}I,^{14}C,^{3}H), and other assorted compounds, called *reporters* or *labels*, as a way of facilitating antigen or antibody detection or isolation under a variety of experimental conditions. Different types of immunoassays include Western blotting, agglutination, precipitation, radioimmunoassay (RIA), enzyme-linked immunosorbent assay (ELISA), bioluminescent immunoassay, homogeneous immunoassay, magnetic particle immunoassay, immunofluorescence, and flow cytometry assay. Typical sensitivities in microgram per milliliter for various immunoassays are listed in Table 7.16.

A major problem introduced through the use of most labeling techniques is that a separation step is the only possible way to differentiate between free and bound labeled species because both contribute to the signal. An assay involving such a separation step is called a *heterogeneous assay*. These assays are simpler to develop but somewhat more complicated to perform. A homogeneous assay, in contrast, requires no separation because the signal-producing reaction is affected in some way by the binding of the antigen to the antibody without interference from the unbound species.

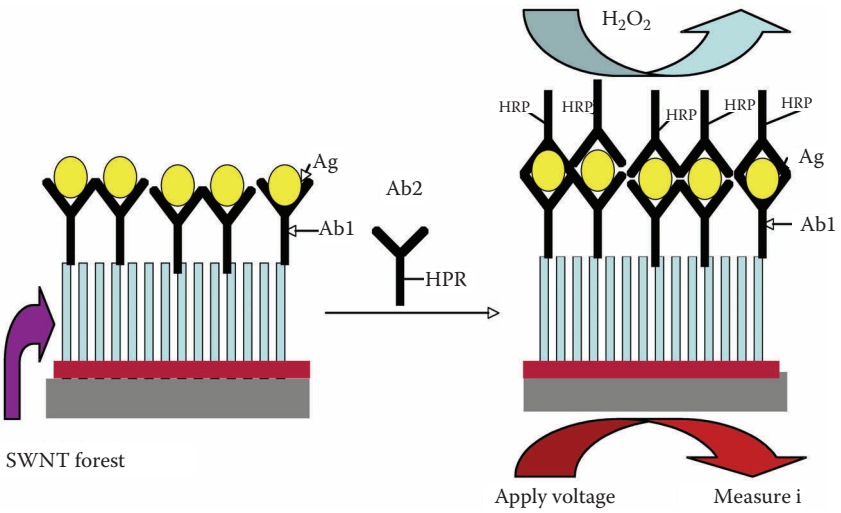


FIGURE 7.77 Noncompetitive assay or sandwich assay.

Table 7.16 Sensitivity of Various Immunoassays

Assay	Sensitivity* (μg antibody/mL)
Precipitation reaction in fluids	20–200
Precipitation reaction in gels	
Mancini radial immunodiffusion	10–50
Ouchterlony double immunodiffusion	20–200
Immunoelectrophoresis	20–200
Rocket electrophoresis	2
Agglutination reactions	
Direct	0.3
Passive agglutination	0.006–0.06
Agglutination inhibition	0.006–0.06
Radioimmunoassay	0.0006–0.006
Enzyme-linked immunosorbent assay (ELISA)	<0.0001–0.01
ELISA using chemiluminescence	<0.0001–0.01†
Immunofluorescence	1.0
Flow cytometry	0.06–0.006

*The sensitivity depends on the affinity of the antibody and the epitope density and distribution.
†Note that the sensitivity of the chemiluminescence-based ELISA assays can be made to match that of a radioimmunoassay.

Ideally, the assay and its accompanying instrumentation should not require a label, and detection thereof, to ascertain the binding of any type of biological affinity pair. It should also be inexpensive, easy to use, and compact. The latter assays are called *direct binding* or *label-less assays*. They continue to be a topic of intense research.

An immunosensor is a sensor monitoring either an antigen or an antibody based on the selective binding of antigen-antibody pairs. The interest in immunosensing is to adapt the above-described assays (Figures 7.76 and 7.77) into a sensor format and to make measurements by simply bringing the sample in contact with the sensor. Our interest here is to explore how micromachining could help this cause. It is especially challenging to develop a homogeneous immunosensor where a washing step is not required. Micromachining might result in a microinstrument, perhaps with some microfluidics flushing the unbound species from the sensor surface. Better yet, micromachining might open the road to label-less detection, as demonstrated with the static cantilever approach for DNA SNP detection in Example 4.1.

Several difficulties still must be overcome in immunosensors, such as assays involving too many steps to qualify as sensors, the deposition of uniform protein coatings, unselective protein adsorption, distinguishing between bound and unbound antigen, and the inherent irreversibility of immunosensors.

Micromachined Potentiometric Immunosensors

We highlight a few developments here to illustrate how new detection schemes and micromachining techniques affect the commercialization of immunosensing. For a more detailed review on immunosensors with commercial potential, we refer to Buerk,[30] Madou et al.,[31] and the March/April 2000 issue of *Fresenius' Journal of Analytical Chemistry*.[32]

Early on, immunosensor work was geared toward direct detection of antigens or antibodies with potentiometric techniques, that is, by simply measuring a voltage. Antigen-antibody binding on the surface of a potentiometric-sensing electrode leads to a new charge distribution. In principle, this redistribution of charges should be measurable. The effect of direct binding on the potential of a sensing electrode was observed on polymeric membrane-coated electrodes, as well as on metal electrodes. However, the signal was found to be small (1–5 mV), making the precision of the measurement poor.[33] Also, ISFETs (see above) have been attempted as one special embodiment of potentiometric devices, and several "immunoFETs" were made in laboratories around the world.[34] Unfortunately, as Janata et al.[35,36] concluded, the potential changes on potentiometric devices are caused by a mixed potential generated from the antigen-antibody binding, as well as other charge transfer reactions occurring at the electrode surface/electrolyte interface. To make a potentiometric sensor work, Janata et al. pointed out that a substrate surface with an infinitely high-charge transfer resistance is needed, as only then will protein charges exclusively account for the potential generated. Such a perfect polarizable surface has yet to be identified (see Figure 7.43). Not even densely packed Langmuir-Blodgett films qualify because they do not block all current across the interface.[37] In Figure 7.78A, an "immunoFET" is schematically shown with a membrane (perhaps

a Langmuir-Blodgett film) and attached antibody molecules.

An alternative, direct-measuring immunosensor is the capacitive affinity sensor (Figure 7.78B). This immunosensor is based on the capacitance change associated with the antigen-antibody binding between an insulated pair of interdigitated electrodes on a dielectric surface. The capacitance change arises because the antibody, a large molecule with a small dielectric constant, displaces water with a high dielectric constant. Again, as with the potentiometric sensor, unless a perfect polarizable interface (i.e., a perfect stacking of only the antigen-antibody complexes at the capacitor surface) is realized, capacitive changes cannot unequivocally be associated with the reaction of interest. Along this line, Interuniversity Microelectronics Centre (IMEC, in Leuven, Belgium) is developing capacitive biosensors that allow for the detection of affinity binding of biomolecules (e.g., antigens,

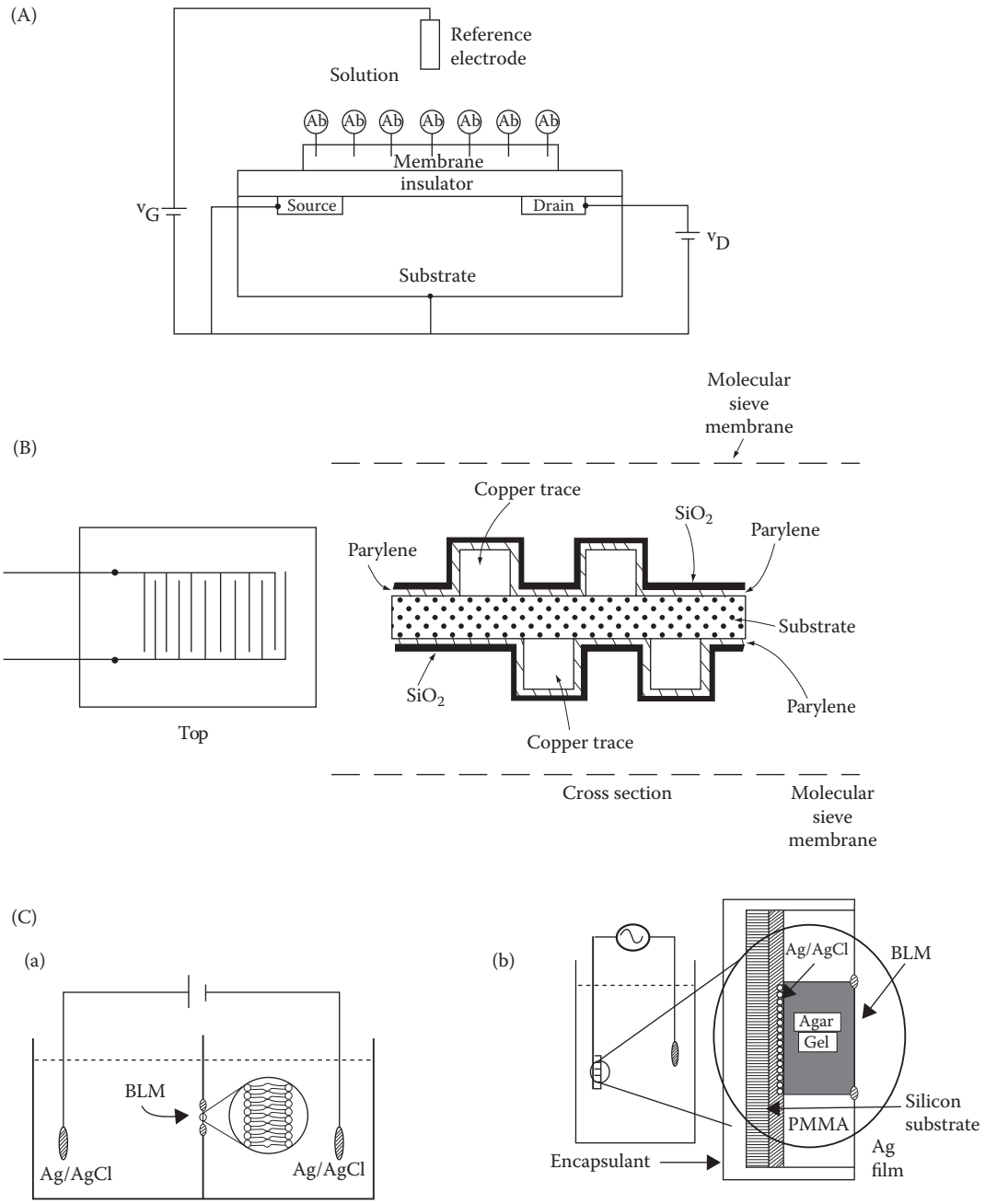


FIGURE 7.78 Schematic diagram of an immunoFET (A). Examples of micromachined, direct immunosensors (B). (C) Lipid membrane suspended in a pinhole (a) and more robust Langmuir-Blodgett lipid layer deposited on a hydrogel (b). See text.

DNA). The presence of DNA sequences or antigen-antibody binding can be detected by looking for impedance changes between closely spaced pairs of electrodes, as sketched in Figure 7.79 (http://www.wtec.org/loyola/mcc/mems_eu/Pages/Chapter-5.html). Nanoscaled interdigitated electrode arrays, as shown here, are made with deep UV lithography. Electrode widths and spacings from 500 nm to 250 nm are produced on a dielectric substrate (0.5 mm × 1 mm). In Figure 7.78Ca, a lipid membrane is suspended in a pinhole between two water reservoirs. Upon exposure to certain analytes, a current across the membrane is observed. This very brittle configuration is rendered a little bit more robust by depositing a thin Langmuir-Blodgett lipid film on a hydrogel as shown in Figure 7.78Cb.

The nonexistence of perfectly polarizable electrodes, nonselective protein interactions with the electrode, and the influence of pH and ionic strength (through Debye shielding) make the prospect of developing selective and sensitive direct potentiometric or capacitive immunosensors rather dim.

Amperometric Sensors

Glucose Sensor The amperometric glucose sensor we are using as a commercial example is a second-generation (see page 571 "Electron Tunneling Transfer Rate k_t," for a discussion of three generations of glucose sensors), disposable glucose sensor strip for diabetes management that relies on an electron acceptor (mediator) to establish electrical communication between the redox center in the glucose oxidase (GOx) enzyme and the electrode. This MediSense/Abbott sensor, the Precision QID, is shown in Figure 7.80. The sensor consists basically of a two-electrode system with counter and reference electrodes combined in a Ag/AgCl electrode. Blood is sufficiently conductive, and the sensor works in the diffusion-limited regime; thus, there is no need for a real reference electrode that is expensive to implement. The working electrode is made of carbon. A second, dummy working electrode has a mediator, but no enzyme for background correction; the mediator used is carboxy-ferrocene. A differential current measurement between the two working electrodes makes for more precise measurement of the additional ferrocene generated in regeneration of the FAD (see Figure 7.60, bottom). All films are silk-screened onto a PVC substrate. Single sensors are less than $0.50 each.

Amperometric Immunosensor A one-step amperometric immunosensor developed by this author and his coworkers is shown in Figure 7.81. The approach combines the fundamentals of enzyme immunoassay, electrochemistry, and gel filtration.[38] All the components necessary for the assay are incorporated into the sensor; only the sample has to be added for analysis. The base of the sensor consists of working and reference electrodes deposited onto an appropriate substrate. To make the components as cheaply as possible, one may use thick-film technology on a

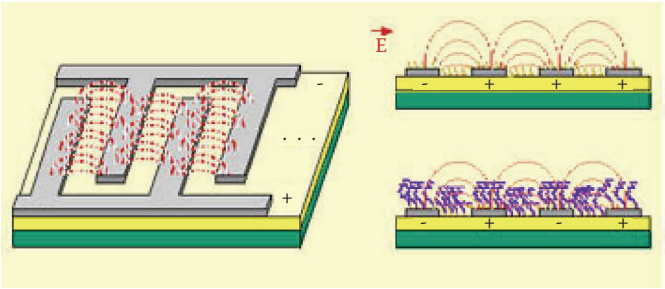
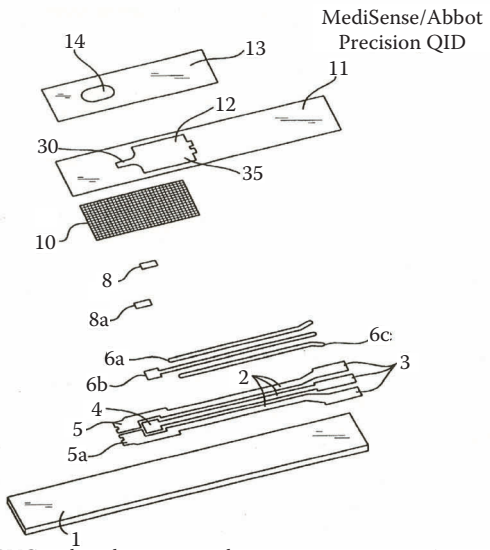


FIGURE 7.79 IMEC capacitive biosensor for DNA and immunosensing. (From IMEC, Leuven, Belgium.)



1. PVC, polycarbonate or polyester
2. Carbon ink
3. Contacts
4. Reference electrode
5. Working electrode
5a. Dummy electrode
6a–c. Ag/AgCl ink
8. Enzyme mediator filler ink
8a. Mediator filler ink
10. Mesh layer
11. Dielectric coating
13. Cover membrane

FIGURE 7.80 Example of a disposable amperometric glucose sensor. MediSense/Abbott Precison QID.

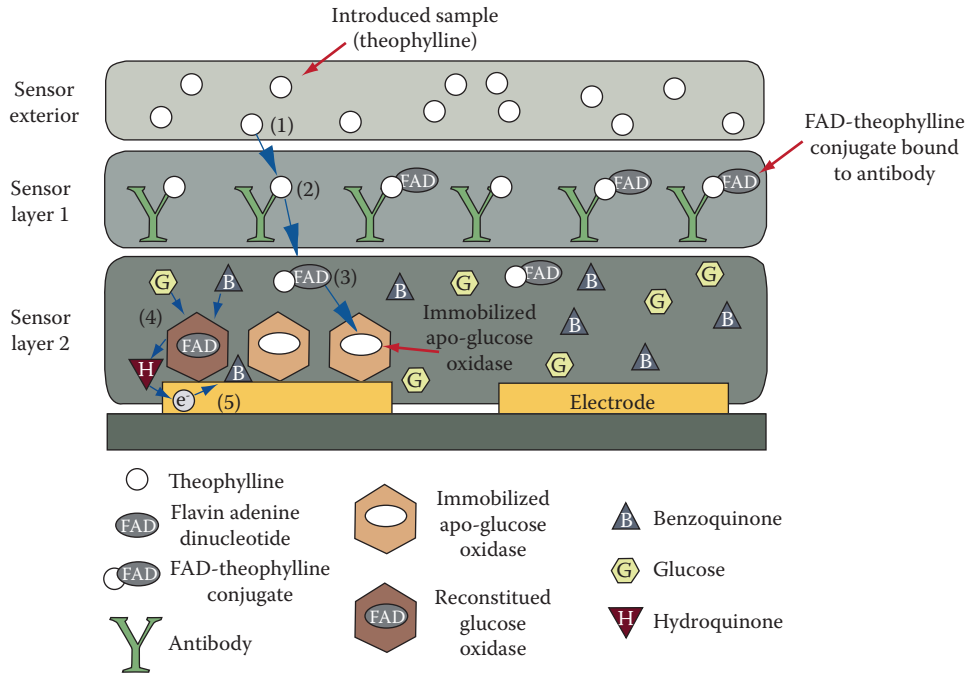


FIGURE 7.81 Layer configuration of immunosensor using theophylline as example reaction scheme. Theophylline sample diffuses into first layer (1). Free theophylline displaces prebound FAD-theophylline conjugate (2). FAD-theophylline diffuses into the second layer and activates apo-glucose oxidase (3). Benzoquinone is consumed and hydroquinone is produced (4). Hydroquinone is oxidized at the metal electrode, leading to an increase in current (5).

plastic substrate. Layered onto the electrode surface is a mixture of an inactive apoenzyme (apo-GOx), an enzyme substrate (glucose in case the enzyme is GOx), and a mediator—reagents necessary to generate an electrical signal. Above the apoenzyme matrix, there is a wettable gel containing the antibody specific for the analyte of interest, topped with a layer of analyte labeled with a prosthetic group or cofactor capable of activating the apoenzyme. A drop of sample placed on the surface of the sensor immediately hydrates the sensor with a fixed volume dictated by the total volume of hydrogel in the device. Sample analyte and cofactor-labeled analyte compete for a limiting amount of immobilized antibody in layer 1. The amount of cofactor-labeled analyte that diffuses to the apoenzyme layer is directly proportional to the concentration of sample analyte. If the sample does not contain antigen, the labeled antigen remains firmly bound to the antibody layer and will not be able to activate the apoenzyme. Activation of apoenzyme by cofactor-labeled analyte results in catalytic breakdown of the enzyme substrate. Reduction or oxidation of one of the reaction products results in a current proportional to the amount of apoenzyme activated, and thus to the concentration of analyte in the sample. The working electrode detects a product of the enzyme reaction with the substrate in an amperometric mode. Using theophylline for analyte and apo-GOx as an inactive enzyme that FAD renders active demonstrates this simple, easy to use, and quantitative immunodiagnostic system. The principle of the new assay is generic and may be applicable to the development of sensors for a variety of analytes with biomedical interest. The detection limit of the FAD/apo-GOx system is well below 10^{-10} M, adequate for the analysis of a majority of analytes. This detection limit can be reduced further by optimization of the system. A further challenge is to find ways of reproducibly layering the different organic materials shown in Figure 7.81. This challenge does not concern Si micromachining but rather involves learning how to manufacture and manipulate hydrogel materials.

Optical Spectroscopy

Quantum Mechanics in Spectroscopy

Optical spectroscopy is the study of the interaction between matter and light. By measuring and

TABLE 7.17 The Energy of the Electromagnetic Radiation Determines the Type of Interaction with Matter

EM Radiation	Interaction Type	Energy/ Mole	Energy/ Photon
γ-rays	Nuclear transitions	10^{7} kJ/mole	10^{-14} J/photon
x-rays	Core e^-	10^{6} kJ/mole	10^{-16} J/photon
Ultraviolet	Valence e^-	10^{3} kJ/mole	10^{-18} J/photon
Visible	Valence e^-	10^{2} kJ/mole	10^{-19} J/photon
Infrared	Molecular vibrations	10^{1} kJ/mole	10^{-20} J/photon
Microwave	Molecular rotations	10^{-2} kJ/mole	10^{-22} J/photon
Radio	Nuclear spin	10^{-4} kJ/mole	10^{-24} J/photon

quantifying that interaction, it becomes an analytical technique that can be used to identify (wavelength) and quantify (intensity) the constituents in a sample. When electromagnetic energy is applied to matter, it can be absorbed, emitted, transmitted, or induce chemical changes (reactions). The precise mode of interaction depends on the energy of the electromagnetic radiation involved, as illustrated in Figure 5.2 and Table 7.17. From quantum mechanics (Chapter 3), we know that particles are restricted to fixed energy levels: molecules can only rotate and vibrate at certain frequencies, and electrons in atoms can only move at certain velocities. Quantum numbers define the electronic, vibrational, and rotational states of a species. These quantum numbers are obtained from the Schrödinger equation, the solution to which is a wavefunction (ψ) describing these states. As a consequence, molecules absorb energy as quanta (photons with energy E = hν) based on criteria specific to each molecular structure. This is what makes the absorption process selective to specific wavelengths of the electromagnetic spectrum ($E = hc/\lambda$, with c the velocity of light at 2.9972×10^{10} cm/s). Molecular rotation involves an entire molecule moving, and the energy change, ΔE, involved is very small (~0.1 kcal/mol). Bond vibrations involve a few nuclei moving, and ΔE is therefore larger (~1–10 kcal/mol), whereas electronic states involve an electron moving out of its bonding orbital, with ΔE the largest among the three (~100 kcal/mol). It is normally the electron spin angular momentum, the orbital angular momentum, and the symmetry properties of the two wavefunctions (initial and final state) that determine whether an energy transition is allowed or forbidden. The feasibility of a transition is given by selection rules in terms of the quantum numbers of the states involved.

UV/VIS Absorption Spectroscopy

Absorption is the most widely used spectroscopic technique because most materials absorb electromagnetic radiation. Absorption of a photon increases the energy of a molecule from a ground state to an excited state (Figure 7.82a). When a molecule emits light, the energy level of the molecule is decreased (Figure 7.82b). The total energy of a molecule is the sum of electronic, vibrational, rotational, translational, and spin orientation energies, some of which are sketched in Figure 7.83. The mechanism of absorption of radiation energy is different in each case; however, the fundamental process is the absorption of a certain amount of energy. Valence electrons are excited in UV/VIS (PHz) spectroscopy, whereas more energetic x-rays excite core electrons in x-ray spectroscopy, and gamma rays (EHz) induce nuclear transitions in Mössbauer spectroscopy. At lower energies, radiofrequencies (MHz) involve the nuclear spin in nuclear magnetic resonance (NMR); microwaves (GHz) involve electron spin and are used in electron spin resonance; and in microwave spectroscopy, molecular rotations are probed. Infrared (THz) excites molecular vibrations and is relied on in IR and Raman spectroscopy.

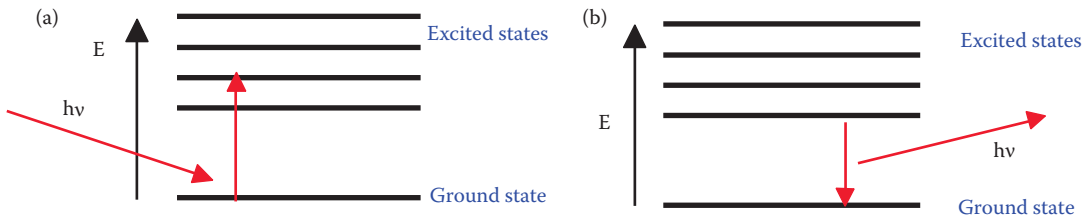


FIGURE 7.82 Absorption of a photon (a) and emission of a photon (b).

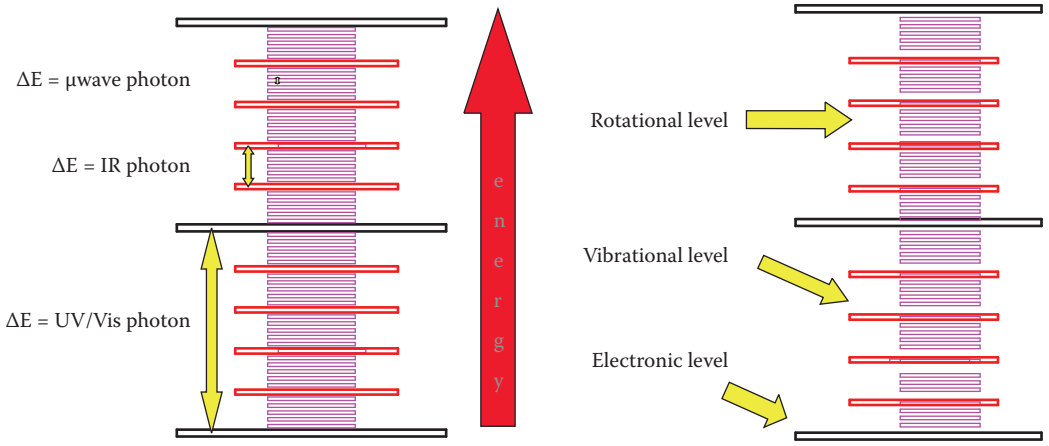


FIGURE 7.83 The total energy of a molecule is the sum of electronic, vibrational, rotational, translational, and spin orientation energies, some of which are sketched here.

Let us focus for a moment on the absorbance, *A*, of visible or UV light (from 160–780 nm) by a solution containing absorbing molecules. Groups of atoms that are responsible for color or pigmentation are called *chromophores*. They come with characteristic optical absorptions and usually contain alternating single and double bonds (such as in aromatic rings). The transmittance of light, *T*, through the solution is given as $T = P/P_0$, with P_0 the incoming light intensity (i.e., *P* at x = 0) and *P* the light leaving the absorbing medium (e.g., at x = b) (see Figure 7.84). Transmittance *T* has values between 0 and 1 and is independent of P_0. The absorbance may also be expressed as $A = \log(P_0/P) = -\log T$. Absorbance or optical density is given by Beer's law (also Beer-Lambert law):

$$A = \varepsilon bc \tag{7.133}$$

where ε is the molar absorptivity or absorption coefficient (wavelength-dependent but also a function of solvent, ionic strength, temperature, and pressure): absorbance of a solution when b = 1 cm and c = 1 M. The molar absorptivity or molar extinction coefficient has units of $M^{-1}\,cm^{-1}$ if *b* is in centimeters and *c* is M. The extinction coefficient of a chemical species at a given wavelength is a measure of how strongly the species absorbs and refers to a single wavelength, usually the absorption maximum. Additionally, *c* is the molar concentration, and *b* is the distance into the absorbing medium (optical path length typically 1 cm).

Equation 7.133 can be derived as follows. In a solution element *dx*, the intensity of the incoming light, *P*, is decreased by *dP* because of absorption. With *dP* given by:

$$dP = -Pc\beta dx \tag{7.134}$$

where c = molar concentration
P = light intensity
β = probability of a photon being absorbed

We can rewrite Equation 7.134 as $dP/P = -c\beta dx$ or:

$$\int_{P_0}^{P} \frac{dP}{P} = -c\beta \int_0^b dx \quad \text{or} \quad \ln P - \ln P_0 = -c\beta x \tag{7.135}$$

where P_0 is the value of *P* at x = 0. In other words:

$$\ln \frac{P_0}{P} = c\beta b = 2.303 \log \frac{P_0}{P} \tag{7.136}$$

The latter expression can be rewritten in the familiar format:

$$A = \log \frac{P_0}{P} = \frac{\beta}{2.303} cb = \varepsilon cb \tag{7.137}$$

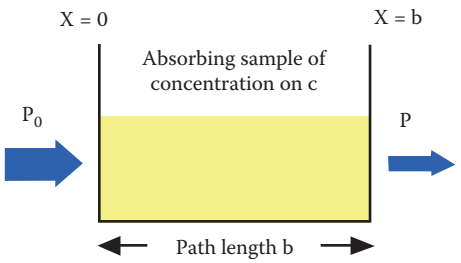


FIGURE 7.84 Absorbance of visible light in a solution of concentration c.

The lowest energy transition (and the most often observed in UV/VIS spectroscopy) is typically that of an electron going from the highest occupied molecular orbit (HOMO) to the lowest occupied molecular orbit (LUMO). For any bond (pair of electrons) in a molecule, the molecular orbitals are a mixture of the two contributing atomic orbitals; for every bonding orbital "created" from this mixing (σ, π), there is a corresponding antibonding orbital of symmetrically higher energy (σ*, π*) (see Chapter 3 and Figure 7.85). The lowest energy-occupying orbitals are typically the σ orbitals; likewise, the corresponding antibonding σ* orbitals are of the highest energy. The π orbitals are of somewhat higher energy than σ orbitals, and their complementary antibonding orbitals are somewhat lower in energy than σ*. Unshared electron pairs lie at the energy of the original atomic orbital, shown as n in Figure 7.85. Most often this energy is higher than π or σ (because no bond is formed, there is no lowering of the energy for *n*).

Electronic absorption spectroscopy of molecules usually measures $\pi \rightarrow \pi^*$ and $n \rightarrow \pi^*$ transitions. Even if the incoming light has the correct energy (i.e., the frequency is exactly right) for an optical transition, not much may happen if a selection rule is broken, i.e., the specific transition is forbidden. However, selection rules are rarely absolute, and in practice even forbidden transitions can still occur (especially in solution), but such transitions will usually have low intensities. For example, $n \rightarrow \pi^*$ electronic transitions are "forbidden" (so they are inefficient), whereas $\pi \rightarrow \pi^*$ electronic transitions are "allowed" (so they are very efficient). The *n* orbitals do not overlap well with the π* orbital, so the probability of this excitation is small.

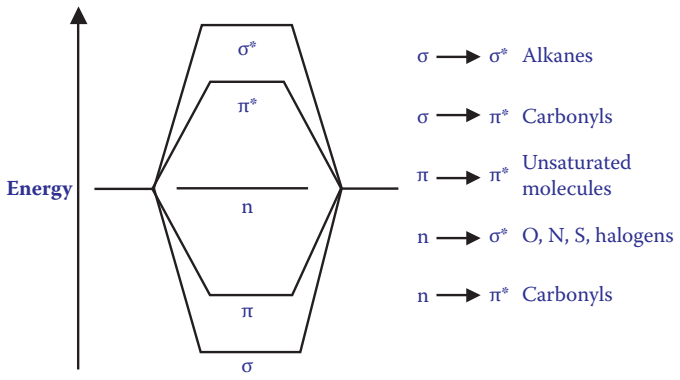


FIGURE 7.85 Electronic transitions in UV spectroscopy with example molecules.

Molar absorptivities, ε, are very large for strongly absorbing chromophores, with ε > 10,000 M^{-1} cm^{-1}; a weak chromophore is one with ε = 10–100 M^{-1} cm^{-1}. The magnitude of ε reflects both the size of the chromophore and the probability that light of a given wavelength will be absorbed when it strikes the chromophore. A general equation stating this relationship may be written as follows: $\varepsilon = 0.87 \times 10^{20}$ *Ta*, where *T* is the transition probability (0–1) and *a* is the chromophore area in square centimeters. Chromophores are typically about 10 Å long; therefore, a transition of T = 1 will have an ε of 10^5.

Absorption measurements are widely applicable but offer only medium sensitivity in the 10^{-4}–10^{-5} M range. Optimization of the measurement procedure can extend the range to 10^{-6} M or even 10^{-7} M. Selectivity is only moderate, and separation before measurement is often required.

Standard absorption spectrometers are shown schematically in Figure 7.86. Light sources are tungsten-halogen (visible) and hydrogen or deuterium discharge (UV) lamps; a typical monochromator is a Czerny-Turner, and detectors are photomultiplier tubes (PMT) or diode arrays (which are less sensitive than PMTs but allow for faster spectral scanning).

Light intensity passing through the sample is measured directly in the path of the incoming light. In the system depicted in Figure 7.86a (a so-called

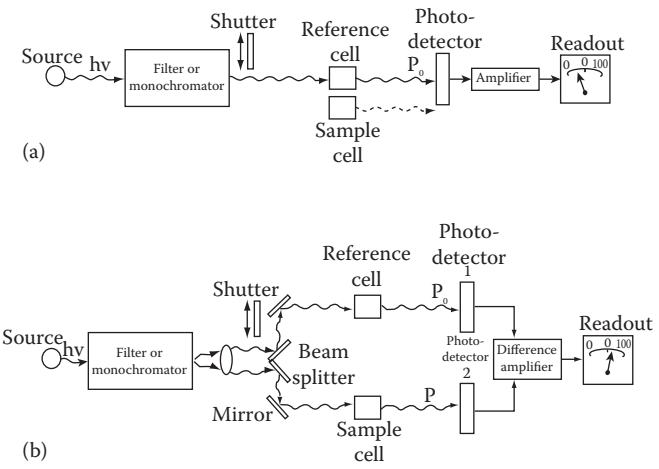


FIGURE 7.86 Two types of optical absorption spectrometers. In the system depicted in (a), the user has to switch sample and reference cell. The system in (b), using a beam splitter, accomplishes the referencing automatically.

single-beam instrument), the user has to switch between sample and reference cell manually. Spectra of absorbing species and solvent are additive. Therefore, one takes the spectrum of a blank solution and subtracts this spectrum from that of the sample solution. The measurement setup in Figure 7.86b is a double-beam system, where a beam splitter accomplishes this referencing automatically. After passing through a monochromator, a light beam splits into two beams; one goes to the sample and the other one to the reference. The beams pass to each photodetector at alternate times by virtue of a chopper. Double-beam instruments compensate for all but very short-term electrical fluctuations because measuring P_0 and P is made almost simultaneously. When many wavelengths must be measured, double beam is a time saver. A single beam is simpler and easier to maintain and is used for quantitative analysis at a single wavelength because it is cheaper.

A fiberoptic-based absorption measurement instrument is very useful when the sample cannot be placed in a cuvette, for example, in a flowing solution or in a 96-well plate reader.

IR and Raman Spectroscopy

The above analysis focused on UV and visible light absorption spectroscopy. A few words about IR absorption and Raman spectroscopy follow. As discussed in Chapter 5, IR absorption and Raman are complementary techniques based on vibrational transitions. They are not used much for quantitative work (their detection limits are not very good) and are most useful for the identification of molecules. As such, they are more often used in studies of structure and conformation of complex molecules. In Raman spectroscopy one measures the frequency (wavelength) change of the incident radiation through inelastic scattering with molecules in a wide range of wavelengths (not only those corresponding to vibrational levels). Scattered light under these circumstances reveals satellite lines below and above the Rayleigh scattering peak at the incident frequency. Raman scattering depends on the polarizability of the scattering molecules (see Chapter 5).

In IR absorption spectroscopy one measures the absorption of the IR radiation at any specific wavelength that matches (in energy) a vibrational transition of the molecule. If the selection rules for optical transitions were all the same for IR and Raman, the peak intensity may be different, but the energy shifts observed in Raman spectra should be identical to the energies of its IR absorption. However, IR absorption selection rules require that a vibrational mode of the molecule have a change in dipole moment or charge distribution associated with it. For example, homonuclear molecules, N_2, Cl_2, and H_2, are not IR active, but they are Raman active because the polarizability of the bond varies periodically in phase with the stretching vibrations. Thus, IR and Raman differ with regard to which vibrational energy transitions are active as a result of selection rules. Some important advantages of Raman over IR are as follows. First, sample preparation for Raman is easy: water is an excellent solvent for Raman, but it is a serious interference for IR absorption. Also, Raman spectra can be measured with normal transparent sample holders like glass or quartz, whereas IR measurement requires KBr or other atmospherically unstable window materials. Moreover, because Raman can also involve rotational or electronic transitions of the molecules from which the scattering occurs, a much wider range of energy sources can be used (to interact with the electronic, vibrational, and rotational energies of molecules as calculated in Chapter 3). Raman intensity increases with the fourth power of the frequency of the source (see Chapter 5, Equation 5.261), and a high Raman peak intensity can be obtained by using a short wavelength laser source.

Photodecomposition of samples may be avoided by using red or near-infrared irradiation sources, such as a diode laser (782 or 830 nm) or a neodymium-doped/yttrium aluminium garnet (Nd-YAG) laser (1064 nm). Higher-temperature studies are possible in Raman spectrometry because IR radiation at higher temperature does not affect the measurement; this would interfere with IR spectroscopy, but Raman spectrometry measures the shift (energy difference). On the other hand, IR is better than Raman for measuring molecules with a highly asymmetrical charge distribution and has a lower detection limit than Raman [except when comparing with surface-enhanced Raman spectroscopy (SERS); see Chapter 5]. Also, Raman must use much more

intense sources than IR to get a measurable signal, and IR instruments are also much cheaper.

Many variations of powerful Raman spectroscopy have been developed, such as surface-enhanced Raman spectroscopy (SERS), where surface plasmons of Ag or Au surfaces or nanoparticles are excited by a laser. The resulting electric fields cause nearby molecules to become Raman active, resulting in an amplification of their detection by a factor up to 10^{11}! This effect was originally observed by Fleischmann[39] and was explained by Van Dyune in 1977.[40] Surface plasmon resonance (SPR) and SERS were discussed in Chapter 5.

Scaling in Absorption Spectroscopy and Cavity Ring-Down Spectroscopy

From Beer's law, optical absorption of any type of wavelength through a material scales poorly in the microdomain because it depends on the path length b (see Equation 7.133). Therefore, miniaturization of the absorption path length always decreases the sensitivity! This is unfortunate because besides being the most generic optical spectroscopic method (all molecules absorb electromagnetic radiation at some frequency), the intrinsic molecular cross-sections are typically largest for absorption (compared with any other optical technique, such as fluorescence).

To miniaturize an absorption spectrometer without losing sensitivity, the optical absorption path can be folded and put on a small footprint by bouncing the light from a set of mirrors. Although an improvement, the reflectivity of the mirrors limits the maximum number of bounces typically to about 100, thereby limiting the ultimate total optical path length, and thus the sensitivity of the miniaturized instrument. The best alternative absorption method, originally devised for use in the visible, UV, and infrared spectra, was developed during the 1980s by O'Keefe[41] at Stanford University and later by Saykally[42] at the University of California, Berkeley and Meijer[43] at the University of Nijmegen. This new method is the so-called *cavity ring-down spectroscopy* (CRDS), a highly sensitive method for absorption spectroscopy using pulsed lasers. It can be viewed as a multipass technique affording 10–1000 times longer path lengths than traditional multipass reflection cells. The problem in conventional multipass approaches lies in the requirement for the light beam to traverse distinct paths for each pass. In CRDS one does not take measure of the absorbed signal strength for a given distinct sample path, but instead measures the time rate of absorption of a sample located within a closed optical cavity.

In CRDS, the time required for a laser pulse to decay inside an optical cavity, also termed an *optical resonator*, formed from two highly reflective dielectric plano-concave mirrors, is measured.

A schematic is shown in Figure 7.87. In this setup, pulsed laser radiation enters the cavity through the first mirror, M1, and makes a number of round trips (n) between the mirrors.

The mirrors are ultrasmooth surfaces with ~0.05 nm rms surface roughness and may feature a reflectivity R = 0.99999 in the visible part of the spectrum. The intensity of the laser pulse trapped between the two mirrors decays exponentially with time at a rate determined by the round-trip loss experienced by the laser pulse (Figure 7.88).

The round-trip time is controlled by the cavity length L, and at each bounce on the second mirror M2, some light leaks from the cavity and is detected at a photomultiplier (PMT). For a 1-m cavity, the round-trip time is 6.7 ns. The intensity of the light pulse in the cavity decays exponentially

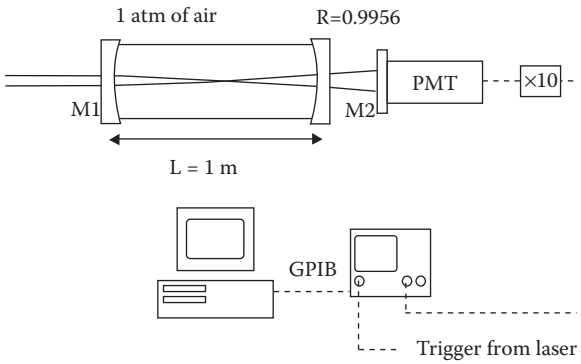


FIGURE 7.87 CRDS, a typical experimental setup. Pulsed narrow band radiation from a tunable (dye) laser system is introduced in the ring-down cavity formed by two plano-concave mirrors (M1 and M2). The mirrors are coated for optimum reflectivity in the desired wavelength range. To ensure that all transverse modes are detected with equal probability, the photomultiplier tube (PMT) that is used to measure the ring-down transients is placed directly behind the output mirror (M2).

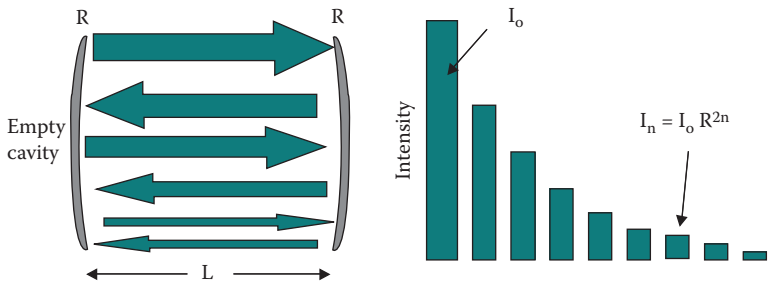


FIGURE 7.88 Decay of light intensity of the laser pulse as a function of the number of round-trips (*n*).

and is characterized by a ring-down lifetime, τ, or the time it takes the intensity to decrease to 1/*e* of its initial value. For mirror reflectivities of 99.99%, the ring-down time is on the order of 30 μs, during which time the pulse has made 5000 round trips for a total path length of 10 km. The decay curve is fit to a single exponential function to determine the ring-down time constant. In the case of an empty absorption cell (i.e., ε = 0), the decay time or ring-down time is approximately given by:

$$\tau_0 = \frac{L}{c\,|\ln(R)|} \cong \frac{L}{c(1-R)} \qquad (7.138)$$

where L = absorption cell length (longest distance between the mirrors)
R = reflectivity of the optical cavity mirrors
c = speed of light

With the absorption cell filled, one obtains:

$$\tau = \frac{L}{c(1-R+\varepsilon L)} \qquad (7.139)$$

with ε equal to the Beer's law absorption coefficient of a sample in the cavity.

Two intensity decay curves for an empty and filled optical resonance cavity are compared in Figure 7.89.

From Equations 7.138 and 7.139 we can derive the absorption coefficient as:

$$\varepsilon = \frac{1}{c}\left(\frac{1}{\tau} - \frac{1}{\tau_0}\right) \qquad (7.140)$$

Typically the ring-down time can be determined with a relative accuracy of 10^{-3}. With a mirror reflectivity loss of $(1 - R) = 10^{-4}$, this implies that line-integrated absorption coefficients of better than 10^{-7} (absorption coefficients less than 10^{-9} cm^{-1} in a 1-m cavity) can be measured. In comparison, conventional absorption experiments can only measure absorption coefficients of 10^{-6} cm^{-1}. The effective sample path lengths in CRDS may be in the tens of kilometers, and the method is in principle intensity-independent. Clever setups that exploit variations in CRDS (e.g., measuring monolayers instead of bulk gas adsorption) are being investigated; see, for example, Pipino et al.[44,45] CRDS has been performed in a large spectral region, from the UV (200 nm) to the infrared (10 mm), almost without gaps. A large variety of molecules have been studied, including radicals, ions, clusters, and even molecules in the solid phase. It is a type of spectroscopy that can be applied in many environments, e.g., open air, cells, jets, supersonic expansions, discharges, and flames. Thus far, no MEMS research group has tackled a micromachined CRDS. This is not too surprising given the difficulty of microfabricating highly reflective dielectric plano-concave mirrors. Lithography-based MEMS are typically very good at making projected shapes, but they are not so good at making truly 3D surfaces as required here.

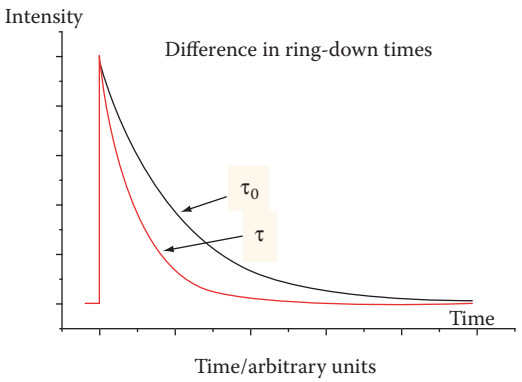


FIGURE 7.89 Difference in ring-down times for filled and empty optical resonance cavity (τ_0 is empty and τ is filled).

Luminescence

Introduction

Luminescence is the emission of light by an atom or molecule. In the case of photoluminescence, the first step in the process is the absorption of a photon, but in chemiluminescence, the light emission is caused by a chemical reaction in which chemical energy is converted into light energy. Bioluminescence is a type of chemiluminescence where the production and emission of light in a living organism are the results of a chemical reaction. We also make a distinction between stimulated emission and spontaneous emission. In stimulated emission, an excited species interacts with the incoming oscillating electric field and gives up its energy to the incident radiation. Stimulated emission, as we saw in Chapter 5, is an essential part of laser action. In spontaneous emission, the excited species gives up its energy of its own accord. Fluorescence, phosphorescence, and chemiluminescence are all spontaneous emission processes.

Photoluminescence

Fluorescence and Jablonski Diagrams Photoluminescent processes are divided into fluorescence and phosphorescence. The difference between fluorescence and phosphorescence is best illustrated in an electronic state diagram or Jablonski diagram as shown in Figure 7.90. Electronic transitions responsible for absorption spectra give rise to singlet and triplet states, and hence create fluorescence and phosphorescence, respectively. In the singlet state, electrons spin in the ground and excited states are paired, creating a net spin of zero. In the triplet state, the electron spins in the ground and excited states are not paired, so there is a net spin (Figure 7.91). Fluorescence does not involve change in spin (the transition is spin allowed) and is short lived (10^{-9}–10^{-6} s).

On the other hand, phosphorescence involves a change in electron spin and is long lived (10^{-3}–100 s) (Figure 7.92). Although it is forbidden, it still happens but with a low probability. The lifetime of an excited molecule depends on how the molecule disposes of the extra energy. Because of the uncertainty principle, the more rapidly the energy is changing, the less precisely we can define the energy (see Chapter 3). Therefore, long-lived excited states have narrower absorption peaks, and short-lived excited states have broader absorption peaks.

The blue lines in Figure 7.90 are radiative processes: absorption (1), fluorescence (2), and phosphorescence (3). The red lines are nonradiative processes: vibrational relaxation or deactivation (4), intersystem crossing to triplet state from a singlet state (5), and internal conversion (6). When a fluorophore, a small molecule, or a part of a larger molecule that can be excited by light to emit fluorescence absorbs a photon from an incandescent lamp or a laser, an excited electronic singlet state might result. This excited state exists for a finite time (typically 10^{-10}–10^{-9} s), during which the fluorophore undergoes conformational changes and is subject to interactions with its molecular environment. These processes have two important consequences. First,

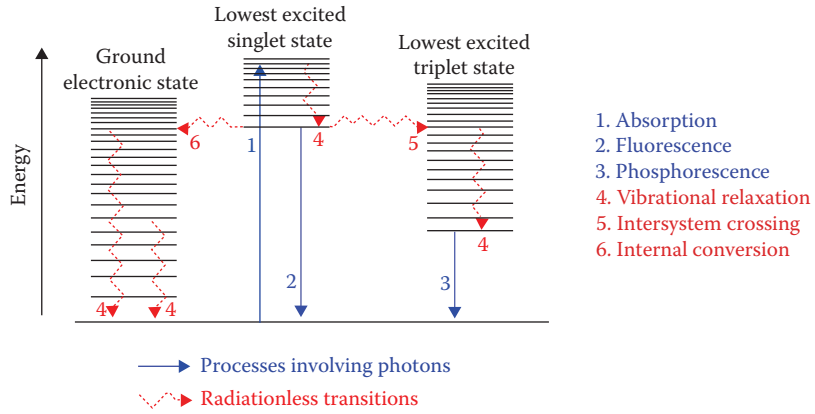


FIGURE 7.90 Electronic state diagram or Jablonski diagram.

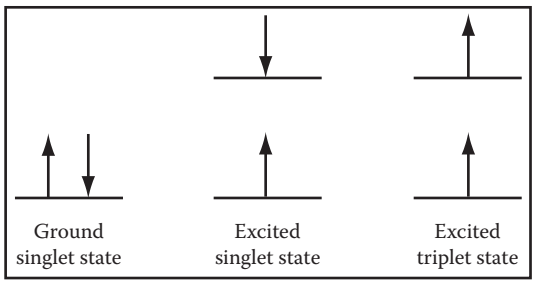


FIGURE 7.91 Singlet ground state and excited singlet and triplet states.

the energy of the excited singlet state is partially dissipated, yielding a more relaxed singlet state from which fluorescence originates. Second, not all the molecules initially excited by absorption return to the ground state by fluorescence emission (i.e., there is a certain limited yield to the process). Other processes such as collisional quenching, Förster resonance energy transfer (FRET) (see 7.147), and intersystem crossing (Figure 7.90, red line 5) may also depopulate the excited singlet state.

George Stokes, professor of mathematics and physics at Cambridge University, interpreted the light-emitting phenomena described here. He formulated the law that fluorescent light is of longer wavelength than the exciting light in 1852 (Stokes law or the Stokes shift), coining the term "fluorescence" instead of the earlier-used term "internal dispersion" in 1853. Stokes law is illustrated in Figure 7.93.

A photon that is absorbed by a molecule does not merely kick that molecule into a higher energetic state, but the molecule also acquires some vibrational energy. It was Franck-Condon who recognized that the photon absorption process is instantaneous, with the rearrangement of almost inertia-free electrons (10^{-16} s) being extremely fast (10^{-16} s) and nuclear motions much slower (10^{-13} s). Nuclear motions are negligible during the time required for an electronic excitation. Because the nucleus does not move during the excitation, the most probable electronic transitions are vertical transitions. In Figure 7.93, the energy diagrams show the potential energy of a molecule as a function of the distance between the nuclei, and according to the Franck-Condon principle, excitation is represented by a vertical arrow. This vertical arrow hits the upper curve not at its lowest point (corresponding to a nonvibrating state), but somewhere higher. After absorbing a photon, the molecule finds itself in a nonequilibrium state and begins to vibrate, lowering the potential energy of the molecule in a radiation-free manner. When the molecule fluoresces, the light originates from near the bottom of the upper potential curve and follows a vertical arrow down until it strikes the lower potential curve. The emission arrow does not

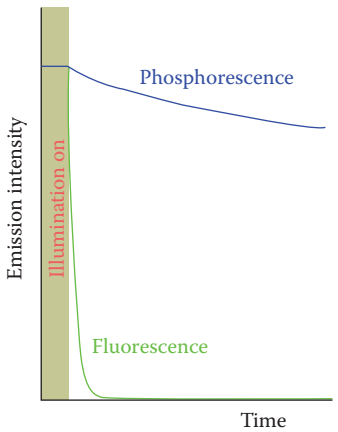


FIGURE 7.92 Decay of phosphorescence and fluorescence.

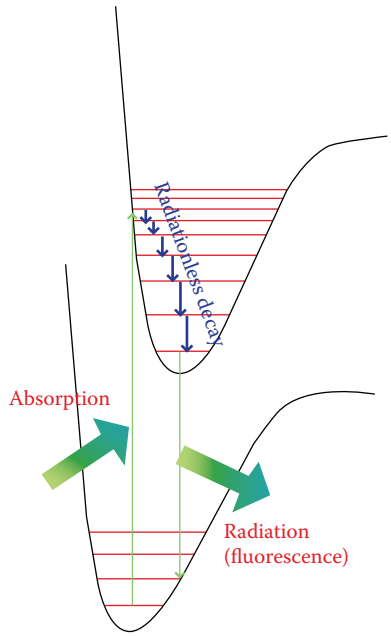


FIGURE 7.93 Stokes law illustrated: fluorescence is at a longer wavelength than absorption. These energy diagrams show the potential energy of a molecule as a function of the distance between the nuclei and, according to the Franck-Condon principle, excitation is represented by a vertical arrow.

strike the lower curve at its deepest point, so again, some excitation energy gets converted into vibrational energy. Because of energy dissipation during the excited state lifetime, the energy of the emitted photon is lower and therefore is of longer wavelength than the excitation photon. This difference in energy or wavelength is called the *Stokes shift*. The Stokes shift is fundamental to the sensitivity of fluorescence techniques because it allows emission photons to be detected against a low background, isolated from excitation photons. There is approximate mirror symmetry between the fluorescence and absorption bands when the shapes of the potential curves in the ground state and the excited state are similar.

The fluorescence quantum yield, which is the ratio of the number of fluorescence photons emitted to the number of photons absorbed, is a measure of the relative extent to which these processes occur. Mathematically, the fluorescence quantum yield is expressed as:

$$\Phi = \frac{\text{Number of emitted photons}}{\text{Number of absorbed photons}} \qquad (7.141)$$

For a highly fluorescent molecule, such as fluorescein, $\phi = 1$, and for a nonfluorescent molecule, $\phi = 0$. The quantum yield can also be defined in terms of various rate constants in the Jablonski diagram. This is shown for a simple system in Figure 7.94, where k_r is the radiative pathway and k_{nr} is the nonradiative pathway:

$$\phi = \frac{k_r}{k_r + k_{nr}} \leq 1 \qquad (7.142)$$

The fluorescence lifetime, τ, is the time delay between the absorbance and the emission and is given by:

$$\tau = \frac{1}{k_r + k_{nr}} \qquad (7.143)$$

Whereas emission of fluorescence occurs in all directions, only the photons emitted at a 90° angle to the incident radiation are monitored to avoid interference from transmitted photons. This is shown schematically in Figure 7.95. An excitation monochromator selects λ_{ex}, the λ of light that the molecule absorbs, and the emission monochromator selects λ_{em}, one of the λ of light emitted by the molecule. Fluorescence detection sensitivity may be severely compromised by background signals, which may originate from endogenous components used in the sample or sample holder (referred to as *autofluorescence*, perhaps of the plastic substrate) or from unbound or nonspecifically fluorescent bound probes (referred to as *reagent background*).

There are many types of fluorescent measuring techniques used in biotechnology. Fluorescence microscopes resolve fluorescence as a function of spatial coordinates in two or three dimensions for microscopic objects (<0.1-mm diameter). Flow cytometers measure fluorescence per cell in a flowing stream, allowing subpopulations within a large sample to be identified, quantitated, and sorted. Flow cytometry involves the use of a beam of laser light projected

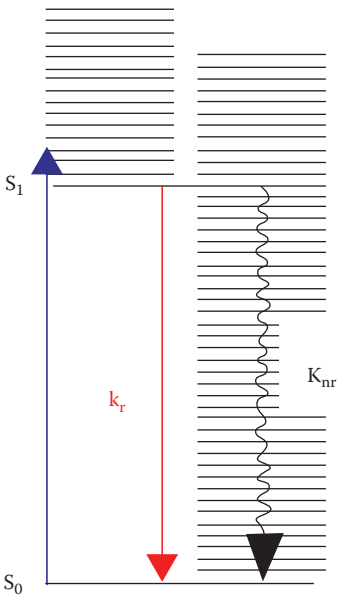


FIGURE 7.94 Simplified Jablonski diagram with rate constants for radiative (k_r) and nonradiative processes (k_{nr}).

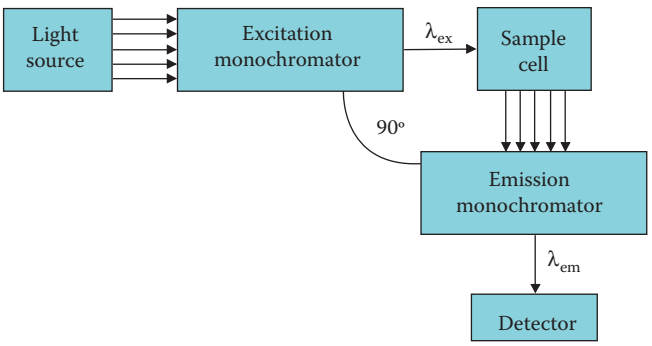


FIGURE 7.95 Simplified fluorescence measurement setup.

through a liquid stream that contains cells or other particles, which, when struck by the focused light, give off signals that are picked up by detectors. These signals are then converted for computer storage and data analysis, and can provide information about various cellular properties. Spectrofluorometers and microplate readers (Figure 7.96) measure the average properties of bulk samples (microliter to milliliter).

Fluorescence scanners resolve fluorescence as a function of spatial coordinates in two dimensions for macroscopic objects such as electrophoresis gels, blots, and chromatograms. In contemporary instruments, the excitation source is usually the 488-nm spectral line of the argon-ion laser. All these mentioned fluorescence instruments are attractive candidates for miniaturization.

Photoluminescence in Analytical Chemistry and Its Scaling Luminescence intensity F of an analytical sample is given by:

$$F = kP_0\varepsilon bc \tag{7.144}$$

where k = constant
P_0 = incident radiant power (W/m^2, also called *intensity*)
c = concentration of the analyte
ε = absorptivity of the sample
b = path length of the sample holder

Fluorescence intensity is quantitatively dependent on the same parameters as absorbance, but a major difference is that luminescence is also dependent on the number of incident photons (P_0). The expression in Equation 7.144 can be derived as follows.

FIGURE 7.96 A typical microplate (96 wells) and microplate reader: the MRX II microplate reader from Dynex.

The fluorescence emission intensity F is given by $F = K'(P_0 - P)$, with P_0 the power of the incident radiation and P the power of the radiation after traversing through a path length b of the medium. The constant K' depends on the quantum efficiency of the fluorescence and the light collection efficiency. In the case of absorption, one measures transmission (T) and determines A as a ratio of P_0 and P, whereas in the case of F, the absolute number of photons is counted. This situation provides fluorescence spectroscopy with a much lower detection limit, and is pictured in Figure 7.97.

We can write $F = K'(P_0 - P)$ as:

$$F = K'\left(P_0 - P_0 10^{-\varepsilon bc}\right) \text{ or } F = K'P_0\left(1 - 10^{-\varepsilon bc}\right) \tag{7.145}$$

Expanding the above equation as a Taylor's series, one obtains:

$$F = K'P_0\left(2.303\varepsilon bc - \frac{(2.303\varepsilon bc)^2}{2!} + \frac{(2.303\varepsilon bc)^3}{3!} + \right) \tag{7.146}$$

If $\varepsilon bc \le 0.05$, then $10^{-\varepsilon bc} \sim 1\text{–}2.3\varepsilon bc$, in which case Equation 7.145 reduces to Equation 7.144 because $F = K'\ P_0\ 2.3\varepsilon bc$ or $F = kP_0\varepsilon bc$.

According to Equation 7.144, fluorescence scales with path length in the same way absorption does. Thus, in theory, both scale poorly into the microdomain, but in the case of fluorescence, we can turn up the number of incoming photons to increase the signal. Therefore, lasers are typically used for excitation because of their higher photon flux, but one needs to be careful about sample decomposition (photobleaching). The lower limit of detection with fluorescence is frequently 10 times better than with absorption measurements, and it is also more specific and less susceptible to interferences because fewer materials absorb and also emit light. Furthermore, fluorescent molecules have different lifetimes, so that the emission from a shorter-lifetime fluorescent interferent might be left to decay before the species of interest is measured (letting the background die out). The concentration detection range is very wide with a reach of three to six decades. With fluorescence, 2–10% accuracy is common because many issues affect the quantum efficiency, ϕ.

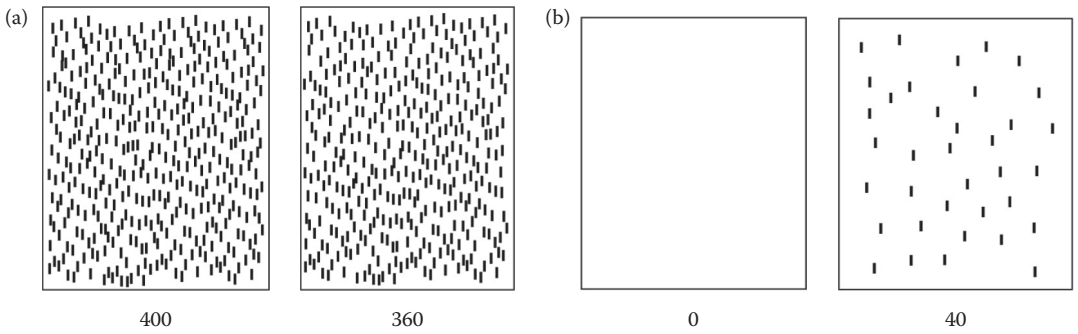


FIGURE 7.97 In absorbance spectroscopy the number of incoming and outgoing photons at the same wavelength are compared (a); in fluorescence the number of emitted photons are counted (b).

Fluorophores, Quantum Dots, and Fluorescent Proteins

Fluorophores A fluorophore is a small molecule, or a part of a larger molecule, that can be excited by light to emit fluorescence by absorbing a photon from an incandescent lamp or a laser. William Perkin, an English chemist, synthesized a coal-tar dye, aniline purple (the first synthetic dye, also known as *mauve* and *mauveine*), in 1856. His breakthrough attracted the attention of numerous synthetic chemists, and a variety of dyes were synthesized soon afterward. Perkin is acknowledged as the founder of the synthetic dye industry. Adolph von Baeyer, a German chemist, synthesized the well-known fluorescent dye fluorescein in 1871 (see Figure 7.98), and Paul Ehrlich, a German bacteriologist, used the fluorescent dye uranin (sodium salt fluorescein) to track the pathway of secretion of aqueous humor in the eye in 1882. The latter represents the first case of the use of in vivo fluorescence in animal physiology. A fluorescent probe is a fluorophore designed to bind to a specific region of a biological structure. Fluorescent probes can be used to tag proteins, nucleotides, lipids, oligosaccharides, and other biological molecules. These fluorescence-tagged molecules may then be used, for example, to localize and monitor interesting biological processes in living cells.

The molar extinction coefficient ε (see Equation 7.133) is a direct measure of a dye's ability to absorb light. This capability is clearly important in determining the amount of light a molecule can generate via fluorescence emission. Most fluorophores in common use have molar extinction coefficients at their wavelength of maximal absorption, ranging between 5,000 and 200,000 $cm^{-1}\ M^{-1}$. The Stokes shift is the difference in wavelength between the fluorescence excitation maximum and the fluorescence emission maximum. Interestingly, dyes with large Stokes shifts tend to have relatively small extinction coefficients, whereas dyes with small Stokes shifts tend to have relatively large extinction coefficients. The fluorescence quantum yield is a measure of the efficiency with which the excited molecule is able to convert absorbed light into emitted light. It is defined as the fraction of absorbed photons that are converted into fluorescence emission (see Equation 7.141). Typical quantum yields for commonly used fluorophores range from 0.05–1.0. The wavelength of absorption of a fluorophore is related to the size of the chromophores, with smaller chromophores leading to higher energy (shorter wavelength). As we just learned, the intensity of a fluorescence measurement is related to the probability of the event and the wavelength of the absorbed light.

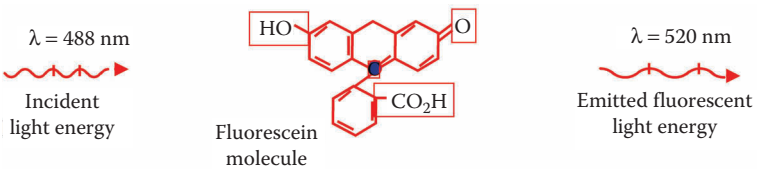


FIGURE 7.98 Fluorescein molecule.

Fluorescence is favored in molecules with structural rigidity. The quantum yields for fluorene and biphenyl, for example (see Figure 7.99), are 1 and 0.2, respectively. The increased rigidity of fluorene stabilizes the π-π* singlet state, leading to higher quantum yield. Chelation also can lead to increased fluorescence. For polyatomic molecules in solution, the discrete electronic transitions are replaced by rather broad energy spectra called the *fluorescence excitation spectrum* and *fluorescence emission spectrum*, respectively. Quantum yield of fluorescence of most molecules decreases with increasing temperature as a result of collisional deactivation of the singlet state. Solvents containing heavy atoms, such as those containing halogens, also decrease fluorescence. Heavy atoms promote intersystem crossing to the triplet state. This decreases fluorescence quantum yield but increases phosphorescence quantum yield (see below). Unless the fluorophore is irreversibly destroyed in the excited state (an important phenomenon known as *photobleaching*), the same fluorophore can be repeatedly excited and detected. Photobleaching is an irreversible photochemical process that causes a molecule to lose its fluorescent property. Countermeasures for photobleaching include the use of shorter illumination, reducing the excitation intensity, and the use of "antifade" reagents. The latter sometimes act by reducing oxygen concentration to prevent formation of singlet oxygen (see below, Reaction 7.6). However, photobleaching has also led to the development of a useful analytical method called *Fluorescence Recovery After Photobleaching* (FRAP). The FRAP method is used to measure the ability of a molecule to move around in a cell over time. A fluorophore is covalently attached to the migrant molecule, and using an epifluorescence microscope, light is focused onto a small subset of the fluorescent molecules. After initial imaging of the amount of fluorescence, a very intense light pulse is flashed onto these same molecules, causing photobleaching. When these molecules are monitored again, the bleached area will now appear dark. If the surrounding fluorescent molecules that were not photobleached are able to diffuse, they will mix with the photobleached molecules, and the dark area will slowly become bright as fluorescent molecules move into it.

Early investigations showed that many substances autofluoresce when irradiated with UV light. In the 1930s, Haitinger and others developed the technique of secondary fluorescence using fluorochrome stains to stain specific tissue components, bacteria, or other pathogens that do not autofluoresce (using acridine orange). In the early 1950s, Coons and Kaplan developed immunofluorescence using fluorescein-tagged antibodies to localize antigens in tissues. They labeled antipneumococcal antibodies with anthracine, allowing them to detect both the organism and the antibody in tissue using UV-excited, blue fluorescence. Two commonly used fluorescent dyes that are covalently bound to antibodies are fluorescein, which emits an intense green fluorescence when excited with blue light (Figure 7.98), and rhodamine, which emits a deep red fluorescence when excited with green-yellow light. Some typical fluorescent dyes are shown in Figure 7.100.

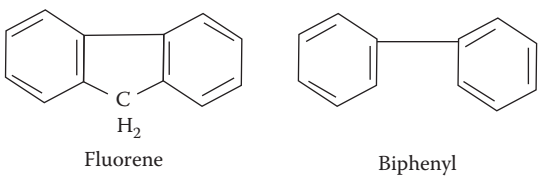


FIGURE 7.99 Fluorene and biphenyl. The fluorescence quantum yields for fluorine and biphenyl are 1 and 0.2, respectively. Fluorene has the higher rigidity and thus the higher quantum yield.

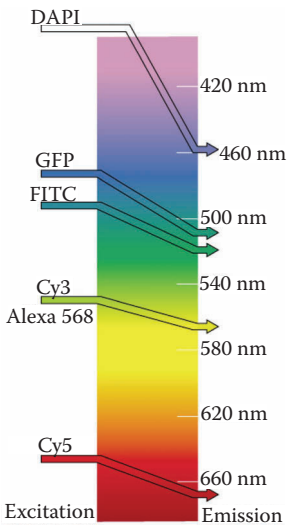


FIGURE 7.100 Typical fluorescent dyes. CY3, indotrimethinecyanines; CY5, indopentamethinecyanines; DAPI, 4′,6-diamidino-2-phenylindole; FITC, fluorescein isothiocyanate; GFP, green fluorescent protein.

Because of their particular structural nature, some fluorescent probes can bind noncovalently to special biomolecules in cells in either a specific or a nonspecific fashion. The binding can dramatically affect the fluorescence quantum yield and wavelength, which is the basis of DNA and RNA probes, protein probes, lipid probes, etc. In Table 7.18, we show some of the fluorophores that are commonly used for lighting up specific cell organelles.

Today there are fluorescent probes for almost any analyte. The binding of hydrogen or metal ions changes the electronic structure of a number of fluorophores, thus affecting the absorption properties and/or the fluorescence properties of the probes. For example, various fluorescein-labeled proteins have quantum yields in the range of 0.2–0.7 at pH 8, but drop off rapidly with decreasing pH, making a fluorescent pH probe. Certain lipophilic ester derivatives of fluorescent probes are cell membrane permeable. Cellular esterases convert these compounds into the free-acid forms of the probes, which are trapped inside the cell and are able to chelate calcium, making for good intercellular calcium probes. Fluorogenic enzyme substrates are fluorescent probes that are converted by specific enzymes into products that have either increased fluorescence or shifted spectra. These are useful probes for monitoring enzyme activities.

In fluorescence quenching, excited molecules relax to ground states via nonradiative pathways instead of fluorescence emission. Nonradiative processes include vibrational relaxation, collision with quenching molecules, and intersystem crossing. Molecular oxygen quenches by increasing the probability of intersystem crossing:

$$M^*(\text{singlet}) + O_2(\text{triplet}) \rightarrow M^*(\text{triplet}) + O_2(\text{singlet})$$

Reaction 7.6

Other typical quenchers are aliphatic and aromatic amines. Polar solvents, such as water, quench fluorescence by orienting around the excited state dipoles. Emission transfer to nonfluorescent molecules, formation of a nonfluorescent complex, aggregation forming a nonfluorescent aggregate, and emission transfer to a fluorescent molecule with an excitation wavelength equal to the emission wavelength of the first fluorescent molecule are all mechanisms that lead to quenching of the fluorescent light signal.

Förster resonance energy transfer (FRET), named after the German scientist Theodor Förster, is a radiationless energy transfer mechanism, where dipole-dipole resonance energy transfers between two chromophores. A donor chromophore in its excited state can transfer energy by a nonradiative long-range dipole-dipole coupling mechanism to an acceptor chromophore in close proximity (typically <100 Å). When both molecules are fluorescent, the term *fluorescence resonance energy transfer* is often used, although the energy is not transferred by fluorescence (there is no emission of a photon during this transfer). The phenomenon is easily detected because instead of the donor fluorescence, the emission of the acceptor is observed. The absorption spectrum of the acceptor must overlap with the fluorescence emission spectrum of the donor as shown in Figure 7.101. Donor and acceptor

Table 7.18 Specific Organelle Probes

Probe	Site	Excitation	Emission
BODIPY	Golgi	505	511
NBD	Golgi	488	525
DPH	Lipid	350	420
TMA-DPH	Lipid	350	420
Rhodamine 123	Mitochondria	488	525
DiO	Lipid	488	500
dil-Cn-(5)	Lipid	550	565
diO-Cn-(3)	Lipid	488	500

BODIPY, borate-dipyrromethene complexes; NBD, nitrobenzoxadiazole; DPH, diphenylhexatriene; TMA, trimethylammonium.

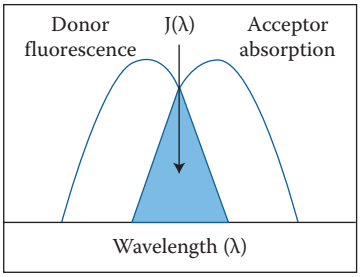


FIGURE 7.101 For fluorescence resonance energy transfer, the absorption spectrum of the acceptor must overlap the fluorescence emission spectrum of the donor. The larger the overlap J(λ), the stronger the FRET.

transition dipole orientations must be approximately parallel. As the intensity of the fluorescence of the acceptor fluorophore increases, the intensity of the fluorescence of the donor fluorophore decreases. Because the interaction between the two fluorophores is a typical dipole-dipole interaction, FRET is dependent on the inverse sixth power of the intermolecular separation (see Equations 7.13 and 7.147), making it useful over distances comparable with the dimensions of biological molecules. Summarizing, the FRET efficiency is determined by three parameters:

1. The distance between the donor and the acceptor
2. The spectral overlap of the donor emission spectrum and the acceptor absorption spectrum
3. The relative orientation of the donor emission dipole moment and the acceptor absorption dipole moment

The FRET efficiency E is defined by the Förster equation:

$$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6} \qquad (7.147)$$

From this expression, the efficiency E depends on the donor-to-acceptor separation distance, r, with an inverse sixth order law as a result of the dipole-dipole coupling mechanism, with R_0 being the Förster distance of this pair of donor and acceptor at which the FRET efficiency is 50%. The donor-to-acceptor separation distance r can be from 0–100 Å.

FRET is an important technique for investigating a wide variety of biological phenomena that produce changes in molecular proximity.

A nice illustration of the FRET principle is found in the use of molecular beacons shown in Figure 7.102. The molecular beacon is a single-stranded nucleic acid molecule with a stem-and-loop structure. The loop portion of the molecular beacon "hairpin" is complementary to a target DNA sequence. The hairpin stem is formed by two sequences that are complementary to each other. In hairpin formation, the quencher, attached to one end of the probe, quenches the fluorophore attached to the other end of the probe, and no fluorescence is emitted. On hybridization of the molecular beacon to its target, the stem dissociates, causing the quencher and the fluorophore to move away from each other, resulting in the restoration of fluorescence.

The rate of fluorescence emission depends on the amount of time that the molecule remains within the excited state (say, excited state with lifetime τ_f). Optical saturation occurs when the rate of excitation exceeds the reciprocal of τ_f. Molecules that remain in the excited beam for extended periods have higher probability of interstate crossings and thus phosphorescence (see below). Usually, increasing dye concentration can be the most effective means of increasing signal when energy is not the limiting factor (i.e., using laser-based confocal systems) and until the dye concentration reaches the limit of the inner filter effect.

Quantum Dots Quantum dots, also known as *nanocrystal semiconductors*, are crystals composed of materials from periodic groups II–VI, III–V, or IV–VI

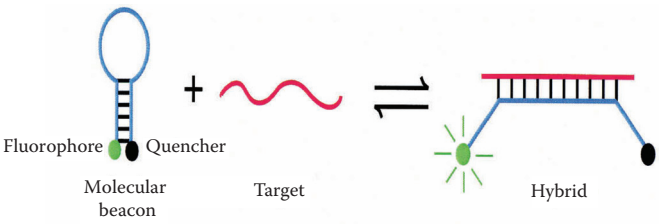


FIGURE 7.102 Schematic presentation of a molecular beacon. In the hairpin loop structure, the quencher (black circle) forms a nonfluorescent complex with the fluorophore (green circle). On hybridization of the molecular beacon to a complementary sequence, the fluorophore and quencher are separated, restoring the fluorescence.

(e.g., CdS, CdSe, PbS, PbSe, PbTd, CuCl). These nanocrystals are used as crystalline fluorophores and can be conjugated to biomolecules as an alternative to the above-reviewed organic dyes. As we saw in Chapter 3, quantum dots (QDs) form a unique class of semiconductors because of their small size, ranging from 2–10 nm (10–50 atoms) in diameter. At these sizes, material properties become size-dependent. In Figure 7.103, we illustrate how smaller dot size leads to emitted light at shorter wavelengths.

With QDs, by adjusting the size of the dot or its composition, the emission wavelength is tunable from blue to the near infrared. CdS and ZnSe dots of decreasing size emit from the blue light region to the near-UV light, different-sized CdSe dots emit light across the visible spectrum; and InP, and InAs QDs emit in the far-red and near-infrared. Elongated QDs (quantum rods) show linearly polarized emission, whereas the fluorescence emission from spherical CdSe dots is circularly polarized or nonpolarized.[46] QDs have a broad absorption spectrum (any light source "bluer" than the dot of interest works) and yield narrow emission. For example, the emission spectra of single ZnS-capped CdSe QDs are as narrow as 13 nm (full-width at half-maximum; FWHM) at room temperature.[47] Symmetric, narrow emission spectra minimize overlap of adjacent colors. Some of the manufacturing techniques of QDs, such as wet chemistry and template chemistry (zeolite, alumina templates), are detailed in Volume III, Chapter 3. See also earlier in this chapter where the Derjaguin, Landau, Verwey, and Overbeek theory (DLVO) was introduced to explain colloid stability. We will learn how crystal size can be controlled by the growth conditions. Before 1993, QDs were often prepared in aqueous solution with added stabilizing agents (e.g., thioglycerol or polyphosphate) to avoid colloid precipitation (see colloid stability, page 519). This procedure yielded low-quality QDs with poor fluorescence efficiencies and large size variations [relative standard deviation (RSD) > 15%]. In 1993, Bawendi and coworkers[48] synthesized better luminescent CdSe QDs using a high-temperature organometallic procedure instead. The resulting nanocrystals had nearly perfect crystal structures and exhibited much narrower size variations RSD < 5%), but the fluorescence quantum yields were still relatively low (~10%). At that time, even if one could make QDs in a narrow size range, they still came with two major problems: poor fluorescence and hydrophobicity, with the latter preventing their use in biological experiments. Today, these problems have been overcome. The addition of semiconductor caps such as ZnS or CdS were found to dramatically increase the fluorescence quantum yield to 45% or higher. A quantum dot often used in this regard is shown in Figure 7.104, a configuration with a CdSe core and a ZnSe shell. The core determines the nanocrystal color, and the shell, of a higher bandgap material (ZnS), dramatically enhances not only the brightness but also the chemical stability. The ZnS protects

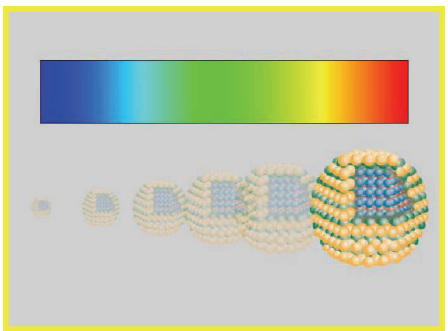


FIGURE 7.103 Quantum dots. Quantum dots change color with size because additional energy is required to "confine" the semiconductor excitation to a smaller volume.

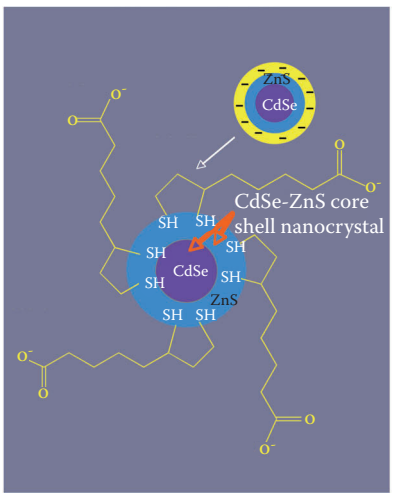


FIGURE 7.104 Nanocrystal with CdSe core and a ZnSe shell. The surface is derived with dihydrogenlipoic acid (DHLA) groups that terminate with negatively charged carboxyl groups. The addition of this coat makes the particles water soluble. To conjugate quantum dots to biomolecules, avidin or protein-G with a positively charged tail may be conjugated to the negatively charged DHLA coat.

the core from oxidation and prevents the CdSe from leeching into the surrounding solution.

Compared with traditional organic dyes such as rhodamine 6G and fluorescein, CdSe nanocrystals show similar or lower quantum yields at room temperature. The lower quantum yields are compensated by their larger absorption cross-sections and much reduced photobleaching rates. Bawendi and coworkers[48] estimated that the molar extinction coefficients of CdSe QDs are about 10^5–10^6 M^{-1} cm^{-1}, depending on the particle size and the excitation wavelength. Chan and Nie[47] estimated that single ZnS-capped CdSe QDs are ~20 times brighter and ~100–200 times more stable than single rhodamine 6G molecules (see also comparison of dyes and QDs in Table 7.19). CdSe cores with ZnS shells are the best QDs for bioapplications with an easily reproducible chemistry and a wide range of emissions.

QDs are generally negatively charged, owing to the colloid stabilizer molecules adsorbed on their surface.[49] These stablizers include hydrophobic tri-*n*-octylphosphine oxide (TOPO), which strongly coordinates to the surface metal atoms, whereas tri-*n*-octylphosphine (TOP) and tributylphosphine (TBP) are weakly bound. These surface stabilizers prevent flocculation in organic solvents and must be replaced with surface stabilizers for use of the same particles in aqueous environments. A few different methods for making nanocrystals water soluble are available. Methods include derivatizing the surface with mercaptoacetic acid or dithiothreitol, encapsulating in phospholipid micelles, derivatizing the surface with dihydrogenlipoic acid (DHLA; groups that terminate with negatively charged carboxyl groups, as shown in Figure 7.104), and coating them with amine-modified polyacrylic acid. In yet another approach, Alivisatos et al.[50] use a silica/siloxane coating for creating water-soluble, ZnS-capped CdSe QDs. During their process, 3-(mercaptopropyl) trimethoxysilane (MPS) is directly adsorbed onto the nanocrystals, and the TOPO molecules are displaced. A silica/siloxane shell is formed on the surface by the introduction of a base and hydrolysis of the silanol groups. The addition of these coats makes the particles water soluble, and therefore more useful for biological applications.

The water-soluble QDs are now ready to be conjugated to all sorts of biomolecules. The surface area of a single QD is large enough to bind multiple biomolecules. Two to five protein molecules and 50 or more small molecules (such as oligonucleotides or peptides) may be conjugated to a single 4-nm QD. To attach QDs to biomolecules in a simple

Table 7.19 Comparison of Organic Dyes and Quantum Dots

Organic Dye	Quantum Dot
Broad absorption and emission profiles spectrum	Broad absorption but narrow emission. Sharper spectrum (FWHM ~25–40 nm)
Low photobleaching threshold. Fades quickly ~100 ps	High resistance to photobleaching and photo and chemical degradation ~5–40 ns
Unstable	Stable output over time (100 times more stable)
One dye excited at a time	Multicolor imaging
Low quantum yield (usually smaller than quantum dots)	High quantum yield. Fluorescent yield is 20 times larger
	Can be made water soluble
	Requires much less sample preparation
Large Stokes shifts come with low molar extinction coefficients	Excited at one wavelength for multiple colors
Can diffuse through cell walls	QDs cannot diffuse through cell walls
Smaller	Bulkier
	Large Stokes shifts
	Wide range (UV-IR)
	High molar extinction coefficient (~10–100 times organic)

electrostatic fashion, avidin or protein-G with a positively charged tail may be conjugated to the negatively charged DHLA coat of a QD. Other scenarios for QD to biomolecule conjugation are illustrated in Figure 7.105. In general, one deals with reactive functional groups, including primary amines, carboxylic acids, alcohols, and thiols. Bioconjugated QDs are finding more and more applications, including DNA hybridization, multiplexed optical encoding, and high-throughput analysis of genes and proteins, immunoassays, and in vivo cellular imaging. Some important advantages of QDs for the latter application are the extremely high photostability of QDs, which allows real-time monitoring or tracking of intracellular processes over long periods (minutes to hours); the ability to use multicolor nanocrystals to simultaneously image multiple targets; and the fact that nanocrystals are believed to be less toxic to the cell than organic dyes. This will allow, for example, viral particles to be followed in vivo, drug molecules to be analyzed in biological systems, and tumor cells to be tracked in real time.

Micro- or nanobeads filled with QDs, as shown in Figure 7.105e, constitute one of the most fascinating new applications of QDs. In this configuration, polystyrene or latex beads are filled with QDs of various combinations of color and intensity (by using different concentrations of the same QD). Specific capturing molecules such as peptides, proteins, and oligonucleotides are then covalently linked to the beads and thus are encoded by the beads' spectroscopic signature, like a barcode on a product. The use of six colors and 10 intensity levels could, in theory, encode 1 million protein or nucleic acid sequences. A single light source is sufficient for reading all the QD-encoded beads. To determine whether an unknown analyte is captured, conventional assay methodologies may then be applied.[51] Because large numbers of barcode variations are possible, each perhaps specific to a different DNA sequence, the application of QD-encoded beads to DNA microarray assays is obvious. Traditional DNA microarrays are composed of short DNA oligomers attached to an inert substrate (glass slide) and typically contain a grid of 10^5–10^6 features (spots), each with a different DNA molecule (a DNA probe) at a known location in the grid. Fluorescently labeled target DNA or RNA hybridizes to these complementary probes. The hybridized array is scanned with a laser to check for a fluorescent signal for each spot, and because we know where each DNA probe is located, we know which ones hybridize successfully. In the case of QD-encoded beads,

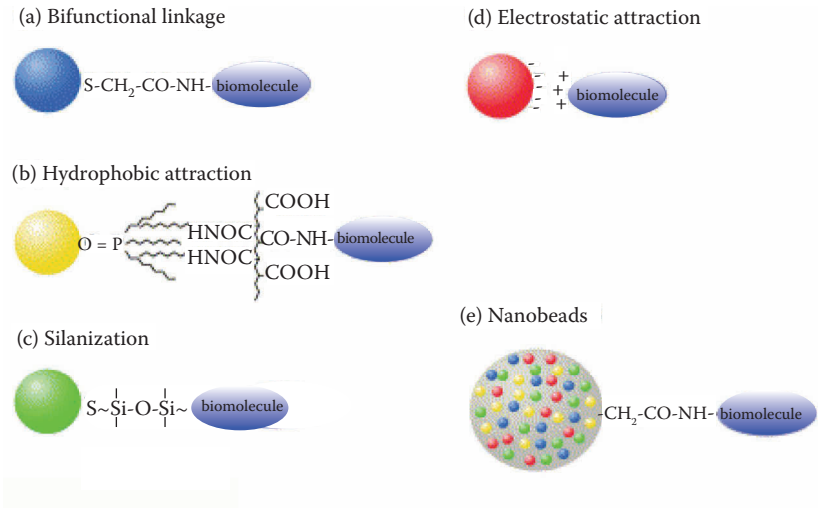


FIGURE 7.105 Quantum dot-biomolecule conjugation schemes. (a) Use of a bifunctional ligand such as mercaptoacetic acid for linking QDs to biomolecules. (b) Tri-*n*-octylphosphine oxide (TOPO)-capped QDs bound to a modified acrylic acid polymer by hydrophobic forces. (c) QD solubilization and bioconjugation using a mercaptosilane compound. (d) Positively charged biomolecules are linked to negatively charged QDs by electrostatic attraction. (e) Incorporation of QDs in microbeads and nanobeads. (From Chan, W. C. W., D. J. Maxwell, X. Gao, R. E. Bailey, M. Han, and S. Nie. 2002. Luminescent quantum dots for multiplexed biological detection. *Curr Opin Biotechnol* 13:40.[46])

multiplexed assays are simply created by mixing the appropriate beads in a solution rather than placing them on a solid surface. We do not need a grid location anymore because each particle comes with its own barcode. One can envision these 3D liquid arrays (Figure 7.106) as the successors of the current generation of gene chips (e.g., see the Nanogen DNA chip in Volume II, Figure 8.51). These new liquid "bar-coding" technologies have the potential to offer significant advantages over current planar chip devices, such as faster binding kinetics (3D vs. 2D), decreased hybridization time, more massive parallel processing, and ease of adding new targets. These novel multiplexed multicolor assays are being developed for massive parallel biosensing and analytical detection.

An entirely different bar-coding principle was developed by Natan and coworkers[52] and uses metallic nanobars. These authors report the development of metallic nanobarcodes for multiplexed bioanalysis. Fluorescence is still used for target detection, but the barcodes are read by measuring the reflectivity differences between different metals. See also the BeadArray™ from Illumina further below in Figure 7.120.

It is still unclear how similar technologies, such as encoded nanobars and QD-tagged microbeads and encodable radiofrequency devices, will compete in genomic, proteomic, and clinical diagnostic applications.

From the above, it appears that QDs could eventually constitute a big improvement over fluorescent dyes. Unlike fluorescent dyes, which tend to decompose and lose their ability to fluoresce, QDs maintain their integrity, withstanding many more cycles of excitation and light emission (they do not bleach easily!). Some of the disadvantages associated with QDs are the possibility of their irreversible aggregation and their relatively bulky size (they cannot diffuse through a cell membrane) compared with most organic fluorophores. In Table 7.19, we compare organic dyes with QDs.

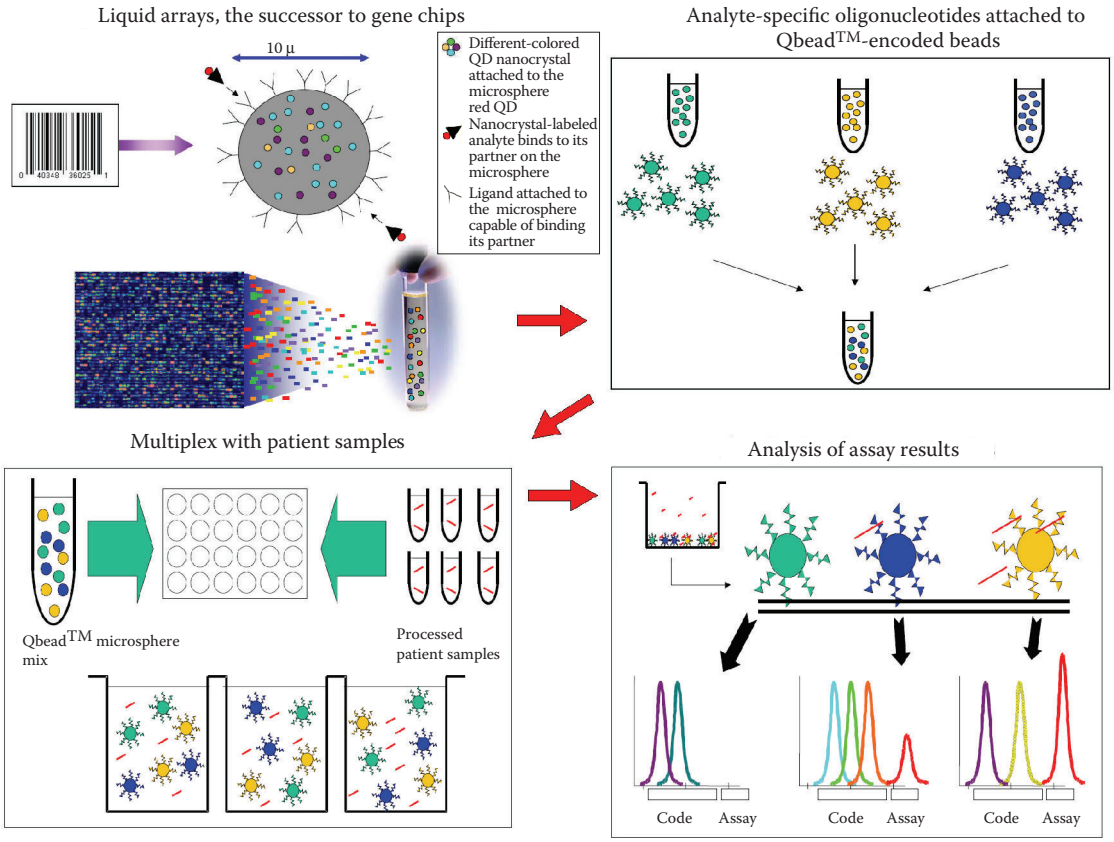


FIGURE 7.106 Spectral barcodes. How to use liquid arrays. The Qbead™ technology is based on the use of QD particles to create spectral barcodes that enable high levels of multiplexing for genetic analysis. As few as 10 assays per well or potentially as many as millions of assays per well may be encoded. Quantum Dot Corporation now known as Invitrogen.

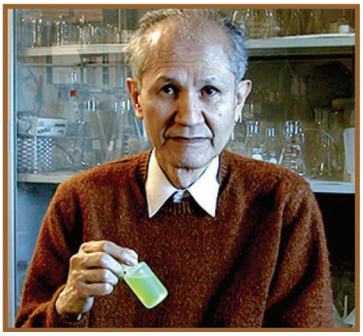


FIGURE 7.107 Dr. Osamu Shimomura.

Fluorescent Proteins As an example of a fluorescent protein, we will focus on green fluorescent protein (GFP). GFP is an intensely fluorescent probe, with a quantum efficiency of approximately 80% and a molar extinction coefficient ε of 2.2×10^4 cm^{-1} M^{-1}. The GFP from the molecular system of the bioluminescent jellyfish *Aequorea victoria* is making an impact on biotechnology today. It was Osamu Shimomura who first noticed GFP in 1962 (Figure 7.107) (http://en.wikipedia.org/wiki/Green_fluorescent_protein). At first, a mere footnote in a scientific paper about a small, bioluminescent jellyfish called *Aequorea aequorea*, or *A. victoria* as the natives from the Bay of Puget Sound call it, the study of this jellyfish's glow became Shimomura's life's work.

The gene encoding GFP was first isolated from the photocytes in the umbrella of the jellyfish *A. victoria*, which was then cloned and introduced into cells of other species. The structure of the GFP molecule, along with a picture of the *A. victoria* jellyfish, is shown in Figure 7.108. This extremely stable protein is made up of 238 amino acids. Its fluorescent properties are unaffected by prolonged treatment with 6 M guanidine HCl, 8 M urea, or 2-day treatment with various proteases such as trypsin, chymotrypsin, or papain. The excitation maxima of GFP are at 395 nm (A state) and 470 nm (B state), and its peak emission is at 509 nm.

The structure of the GFP protein is a marvel of molecular engineering. Looking like a glowing green lantern, it consists of an outer coat, which is a symmetrical large barrel made up of 11 strands of β-pleated sheets. The barrel even has a top and bottom lid made up by shorter protein helices. A much longer helix is centered inside this barrel, and the chromophore itself is nestled in the middle segment of this helix. The cylinder has a diameter of 30 Å and a length of 40 Å. The GFP molecule emits green light when activated by blue or UV light. In vivo, in the jellyfish, there is a radiationless energy transfer from aequorin (AEQ), a bioluminescent protein (see below) present in the same jellyfish that emits at 469 nm to excite GFP, which in turn then emits at 509 nm. This is possible because of the overlap between the emission spectrum of AEQ and the excitation spectrum of GFP [see bioluminescence resonance energy transfer (BRET) in Figure 7.116].

The mechanism of luminescence emission in GFP involves the formation of an internal chromophore, p-hydroxybenzylidene-imidazolone, by an unusual cyclization of three neighboring amino acid residues (Ser-Tyr-Gly at the 65–67 positions of the primary

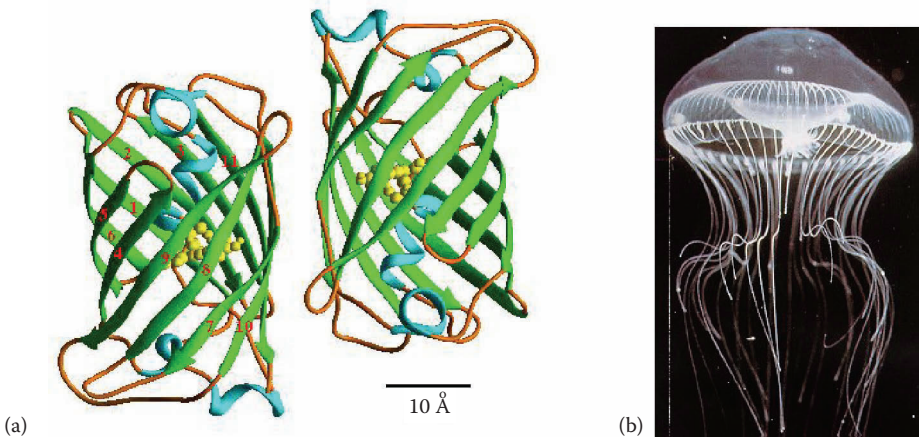


FIGURE 7.108 (a) The overall shape of the GFP protein. Eleven strands of sheet (green) form the walls of the cylinder; short segments of helices (blue) cap the top and bottom of the β barrel (can). (b) The *Aequorea victoria* jellyfish.

sequence). While the wild-type GFP emits in the green region, by introducing amino acid changes in the region close to the chromophore, all types of GFP color variants can be made. This enables the production of GFP variants (mutants) with tailor-made optical properties. Miyawaki et al.[53] have combined the protein calmodulin and the peptide M13 (binds to calmodulin on exposure to Ca^{2+}) with GFP and produced an optical sensing probe for Ca^{2+} with a detection limit in the 100 to 10 billionth of a mol per liter. The sensitivity range depends on the calmodulin mutant used. The spatial resolution of this fluorescent probe is good enough to allow one to compare concentrations of Ca^{2+} between different compartments of one cell.

In 1994, Chalfie et al.[54] coupled the gene for GFP to the gene of another protein of interest. This technique is known as *gene fusion*. The resulting gene fusion, or protein chimera, is then transferred to an organism (e.g., bacterial, yeast, or mammalian cell) so that the latter may produce a protein of interest (protein chimera) that combines the characteristics of both partner protein molecules.[54] After transfer of this target gene into the organism, one can measure the success of the gene transfer simply by shining UV light on the transformed cells. If the cells produce green light, it indicates that the transfer has occurred. GFP has been used to produce animals [e.g., fruit flies, zebrafish, mice, a dog (http://www.ekac.org/transgenic.html), and a bunny (http://www.ekac.org/gfpbunny.html)], as well as plants that light up brightly (see Figure 7.109). This fluorescence can be generated either in the whole animal or plant, or just in some desired part of the organism, such as the cells in a tumor.

FIGURE 7.109 Alba, the fluorescent bunny. (From http://www.ekac.org/gfpbunny.html.)

Further developments have made it possible for the GFP to blink intermitently.[55] This was the first demonstration of light switching of molecules induced at room temperature and of monitoring this switching on a molecule-by-molecule basis.[56] Applications in optoelectronic devices and computing are obvious future targets for GFP. Another protein that is often mentioned in the context of molecular computing and optical switching, bacteriorhodopsin,* requires the measurement of absorbance of a large number of molecules; thus, it appears less efficient for these applications. For more GFP literature, see http://www.imb-jena.de/www_sbx/gottfried/gfp/gfp_literature.html. Books on GFP are listed at http://pantheon.cis.yale.edu/~wfm5/gfp_gateway.html.

Chemiluminescence and Bioluminescence

Chemiluminescence Emission of light by matter (luminescence) has always been known to humans: lightning, the aurora borealis, and light emission by bacteria in the sea or by decaying organic matter are common natural phenomena. Scientific investigation of the luminescence phenomena began with a discovery by Vincenzo Casciarolo, a Bolognian shoemaker and alchemist who in 1603 prepared, serendipitously, an artificial phosphor known as the *Bolognian stone* (or Bolognian phosphor), which glows after exposure to light. Chemi- and bioluminescence measurements with luminometers have become extremely popular in recent years. A luminometer consists of a sample chamber, a detector [usually a photomultiplier tube (PMT)], a signal processing method, and a signal output display. In chemiluminescence, the light emission (sometimes called *cold light*†) is caused by a chemical reaction in which chemical energy is converted to light energy.

* Bacteriorhodopsin (BR) is a photon-driven ion pump. BR is a seven-helical transmembrane protein with a retinal cofactor. It is found in the bacterium *Halobacterium salinarium*, where it converts light to a proton gradient, which in turn is used by a second membrane protein called *ATP synthase* to generate chemical energy in the form of ATP. ATP is then used by the cell to drive a multitude of vital processes. Bacteriorhodopsin movies can be found at http://www.utexas.edu/depts/pharmacology/gonzales/brhod.html.

† To create light, another form of energy must be supplied. The two common ways are incandescence and luminescence. Incandescence refers to the emission of light from heat energy, e.g., the tungsten filament of an ordinary incandescent light bulb glows "white hot."

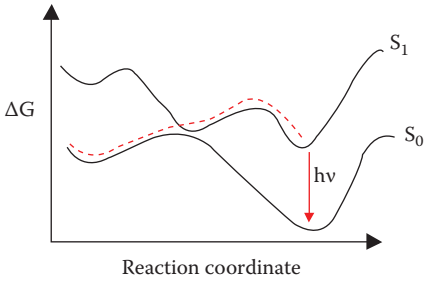
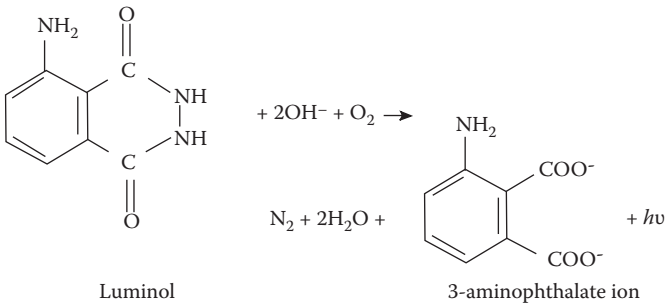


FIGURE 7.110 Chemiluminescence occurs when the ground state of one molecule crosses with the excited state of a rearranged or fragmented molecule.

Chemiluminescence occurs when the ground state of one molecule crosses with the excited state of a rearranged or fragmented molecule, as sketched in Figure 7.110.

These reactions can occur in the gas and liquid states and can last from less than a second to more than a day. Reagents containing the group —CO—NH—NHR react with oxygen, hydrogen peroxide, and many other oxidizing compounds to yield chemiluminescent products. A typical example is that of luminol, given as:



Reaction 7.7

When luminol is placed in an alkaline solution with hydrogen peroxide and a catalyst such as iron, manganese, copper, nickel, or cobalt, the luminol is oxidized to 3-aminophthalate ion in the excited state, causing fluorescence. The luminol creates light via oxidation because two oxygen atoms easily replace the two nitrogen atoms. As this reaction occurs, nitrogen gas is discharged, leaving the luminol in an excited state and releasing additional energy as light. Amino acids, fructose, glycerols, thiols, and serum albumin can also react with luminol to produce an intense light. No excitation source is needed to produce the glow, and a PMT is used to measure and detect low amounts of light in a simple luminometer.

The most important and well-known use of luminol is in the field of forensic science. Blood, which is slightly alkaline, contains cells, water, enzymes, proteins, and hemoglobin. Hemoglobin reacts with the luminol, allowing for the detection of very small amounts of blood. Once an area is suspected to have blood (even if the area has been cleaned) luminol can be applied. The lights are turned off, and after a few seconds a glow appears (Figure 7.111). Many other organic compounds can be oxidized using enzymes such as uricase and oxygen to yield hydrogen peroxide:

$$\text{uric acid} + O_2 \xrightarrow{\text{uricase}} \text{allantoin} + H_2O_2 \qquad \text{Reaction 7.8}$$

The hydrogen peroxide can then be detected with luminol (Reaction 7.7). Examples of compounds that are often detected in this manner are glucose, cholestereol, choline, uric acid, amino acids, aldehydes, and lactate. Biosensors with enzymes immobilized on optical fibers have been developed for the in situ detection of many compounds using this principle.

Another well-known chemiluminescence system is Cyalume from American Cyanamid. In this case a simplified format of the reaction is given as:

$$\text{phenyl oxalate ester} + H_2O_2 \rightarrow \text{intermediate}$$
$$(I) + \text{products I} + \text{flurophor}(F) \rightarrow F^* + \text{products}$$
$$F^* \rightarrow F + h\nu$$

Reaction 7.9

The same reaction is reproduced in Figure 7.112. Here, the phenyloxalate ester is shown, and the reactive intermediate (*I*) is dioxetane. The excited molecule can also transfer energy to another molecule, which then emits light. Different fluorophores (*F*)

FIGURE 7.111 Blood detection using luminol.

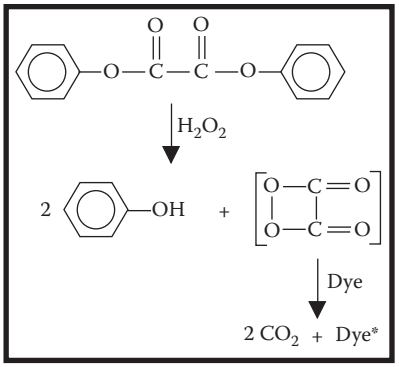


FIGURE 7.112 Oxalate ester in contact with H_2O_2 makes the reactive intermediate dioxetane (*I*). This intermediate reacts with a variety of dyes to give different colors.

reacting with intermediate *I* give different colors used in "glow-in- the-dark" light sticks (Figure 7.113). For red light, rhodamine B is used; the common green comes from 9,10-*bis*(phenylethynyl) anthracene; 9,10-diphenylanthracene results in blue.

A "glow-in-the-dark" light stick is a plastic tube with a glass vial inside. To activate such a light stick, one bends the plastic stick, which breaks the glass vial containing the hydrogen peroxide. This allows the hydrogen peroxide that was inside the glass to mix with the chemicals in the plastic tube. Once these substances contact each other, a reaction sets in that might last as long as 12 hours. When mixed with hydrogen peroxide and a dye, phenyl oxalate gives a quantum yield of 5%, which is not as efficient as a firefly (see below) but still very useful.

One of the widest analytical implementations of chemiluminescence is in the detection of NO or O_3.

FIGURE 7.113 Oxetane intermediate reacts with a variety of dyes to give different colors in light sticks. Plastic tubes have a glass vial inside that breaks on bending and lets the reagents react.

Together they react to produce electronically excited NO_2^*, which emits light throughout the red and near-infrared spectral regions (600–2800 nm):

$$NO + O_3 \rightarrow NO_2^* + O_2$$
$$NO_2^* \rightarrow NO_2 + h\nu \qquad \text{Reaction 7.10}$$

Using an excess of either component, one can detect trace amounts of the other. For example, ozone from an electrogenerator can be drawn together with an atmospheric sample into a reactor, where chemiluminescence is monitored by a photomultiplier in a luminometer. Linearity is obtained for NO concentrations from 1 ppb to 1%. This is the method of choice for measuring NO concentrations from ground level to altitudes up to 20 km. Atmospheric ozone is detected by chemiluminescence, resulting from the reaction of ozone with rhodamine B adsorbed on a silica gel. This reaction has a linear response in the range of 1–400 ppb. Ozone is also detected by the chemiluminescence from its reaction with ethylene.

Bioluminescence Bioluminescence is a type of chemiluminescence where the production and emission of *cold light* in a living organism is the result of a chemical reaction. Bioluminescence is caused by enzymes and other proteins present in the organism. Because it is a biochemically driven reaction/phenomenon, there is no need for excitation sources. These proteins can be used as markers or reporters of other biochemical processes in biomedical research. It is found in many marine animals, both invertebrate (e.g., some cnidarians, crustaceans, squid) and vertebrate (some fishes), in some terrestrial animals (e.g., fireflies, some centipedes), and in some fungi and bacteria (see Table 7.20).

The two insets in the heading of this chapter—a shrimp vomiting a bioluminescent substance to keep a predator at bay and a firefly—are examples of bioluminescence. The anglerfish lives more than a mile deep in the ocean. On her forehead, the female has a "fishing rod" tipped with a lighted "artificial worm" that she dangles over her mouth to attract her next meal. Deep-sea predators, like the anglerfish, need their own light to attract both prey and

TABLE 7.20 Organisms in Nature That Are Bioluminescent

Group	Features of Luminous Displays
Bacteria	Organisms glow constantly; system is autoinduced
Fungi	Mushrooms and mycelia produce constant dim glow
Dinoflagellates	Flagellated algae flash when disturbed
Coelenterates	Jellyfish, sea pansies, and comb jellies emit flashes
Annelids	Marine worms and earthworms exude luminescence
Mollusks	Squid and clams exude luminous clouds and have photophores
Crustacea	Shrimp, copepods, ostracodes; exude luminescence; also have photophores
Insects	Fireflies (beetles) emit flashes; flies (Diptera) glow
Echinoderms	Brittle stars emit trains of rapid flashes
Fish	Many bony and cartilaginous fish are luminous; some use symbiotic bacteria; others are self-luminous; some have photophores

mates. Fireflies also use their flashes to attract mates. The pattern differs from species to species. In one species, the females sometimes mimic the pattern used by females of another species. When the males of the second species respond to these "femmes fatales," they are eaten! The organism itself, or certain symbiotic bacteria living in the organism, can control the light emission. In a firefly, about 100% of the ATP energy is converted into light. By comparison, a normal electric light bulb gives off only 10% of its energy as light, whereas the rest is wasted as heat. Molecular details vary from organism to organism, but each involves a luciferin, a light-emitting substrate; a luciferase, an enzyme that catalyzes the reaction; ATP, which is the source of the energy; and molecular oxygen, O_2:

$$\text{luciferin} + \text{luciferase} + \text{ATP} \rightarrow$$
$$\text{luciferyl adenylate-luciferase} + \text{pyrophosphate}$$
$$\text{luciferyl adenylate-luciferase} + O_2 \rightarrow$$
$$\text{oxyluciferin} + \text{luciferase} + \text{AMP} + h\nu$$

Reaction 7.11

For a simplified reaction scheme, see Figure 7.114. Firefly luciferase was originally prepared by extraction of firefly tails and purified by crystallization. Luciferases and luciferins from different bioluminescent organisms might differ greatly. Today genetic cloning of various luciferases allows for its recombinant production. Genetic engineering of the luciferases and luciferins can extend both the performance and capabilities of bioluminescence.

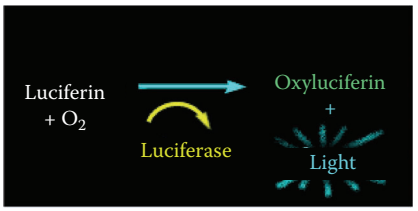


FIGURE 7.114 Luciferin is the substrate and luciferase the enzyme that catalyzes the reaction; ATP (not shown) is the source of energy. Molecular oxygen O_2 is also needed.

The more ATP available, the brighter the light. In fact, firefly luciferin and luciferase are commercially available for measuring the amount of ATP in biological materials (e.g., Millipore's Milliflex Rapid). A well-designed luminometer can detect as little as 0.6 pg of ATP or 0.1 fg of luciferase with six decades of linearity.

Sometimes the luciferin and luciferase (as well as a cofactor such as oxygen) are bound together in a single unit called a *photoprotein*. This molecule can be triggered to produce light when a particular type of ion is added to the system (frequently calcium ions) (see Figure 7.115). This mechanism can be used for a very sensitive Ca^{2+} detector. An example of such a photoprotein is aequorin.

Aequorin, like green fluorescent protein (GFP), is a photoprotein native to the jellyfish *A. victoria* (see Figure 7.108). Aequorin is composed of two distinct units, apoaequorin and an organic chromophore known as *coelenterazine*, a molecule belonging to the luciferin family. In the presence of molecular oxygen, the two components of aequorin reconstitute spontaneously, forming the functional protein.

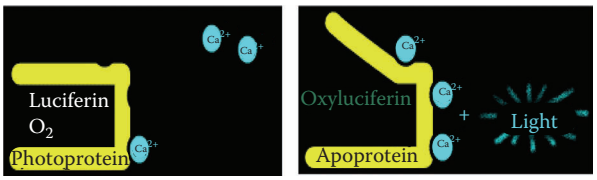


FIGURE 7.115 Luciferin, luciferase, and oxygen are bound together in a photoprotein.

Aequorin emits a flash of blue light (469 nm) that lasts 3–10 s on exposure to calcium ions, which produces a direct, ultrasensitive signal that can be measured with a luminometer:

$$\text{Aequorin} + 2Ca^{2+} \rightarrow$$
$$\text{blue fluorescent protein: } 2Ca^{2+} + CO_2 + h\nu$$

Reaction 7.12

When conjugated to other proteins, such as antibodies or streptavidin, or to small molecules such as biotin, the aequorin conjugates can be used as detection labels in bioluminescence-based assays (BIAs). Adding a calcium salt solution and measuring the amount of light produced during the bioluminescent reaction determines the amount of label present. Because of this calcium-dependent light emission, the aequorin complex has been used as an intracellular Ca^{2+} indicator.

The jellyfish *A. victoria* contains two photoactive proteins. In bioluminescence resonance energy transfer (BRET), a radiationless energy transfer takes place from aequorin, which emits at 469 nm (blue), to the GFP (see Figure 7.116), which then emits at 509 nm, i.e., the green light of the jellyfish (Figure 7.108).

Aequorin can be microinjected into cells or introduced into cells by complementary DNA (cDNA) transfection (see Volume III, Chapter 2 on nature as an engineering guide). An additional tool is provided by the targeting of aequorin to specific compartments of the cell (see Brini et al., 1999)[65] by introducing in the cDNA codes for targeting the protein to specific localities. Similar to GFP, AEQ-cDNA attached to a promoter provides a useful reporter gene.

Scaling in Chemiluminescence and Bioluminescence Earlier, we saw that fluorescence and absorption both scale with path length b, so that both techniques scale badly when miniaturizing the sample path. In the case of fluorescence, the situation is a bit better than for absorption because one can at least turn up the number of incoming photons to increase the signal intensity (see Equations 7.137 and 7.144). In chemiluminescence there is no path length, b, dependence, and intensity, I_{CL}, is only a function of the chemiluminescence efficiency, Φ, and the rate of production of the activated complex (intermediate), c^*:

$$I_{CL} = \phi \frac{dc^*}{dt} \tag{7.148}$$

Luminometers are considerably simpler than fluorescence instrumentation; they need no excitation source—the chemical reaction provides the input energy—and because emitters are usually single wavelength, they also need no wavelength selection device (filter or spectrophotometer). Essentially one only needs a light detector, a sample holder, and a mixing system (e.g., to add the Ca^{2+}; see Reaction 7.12). The latter could be further simplified through the use of a microfluidic system (e.g., see the CD-based fluidic system described in Chapter 6 and Volume III, Chapter 5). Spectral overlap between acceptor and donor excitation is of no concern, and photobleaching by an excitation source is avoided. Chemiluminescent

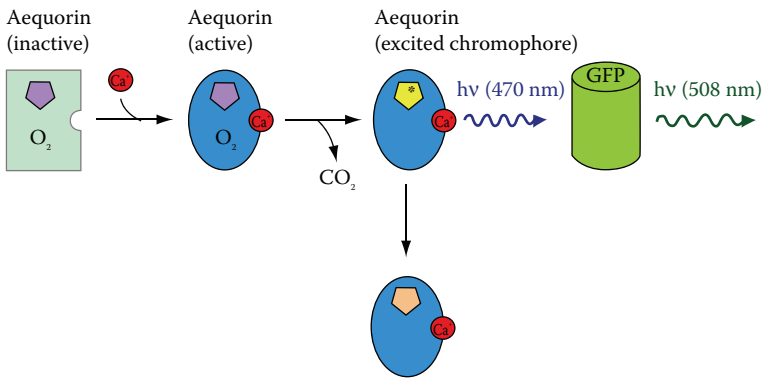


FIGURE 7.116 Bioluminescence resonance energy transfer (BRET) in *Aequorea victoria* from aequorin to GFP.

reactions can be carried out in front of the detector for higher photon collection. Perhaps the most important advantages of chemiluminescence are its high specificity—very few reactions result in light emission—and its very high sensitivity because there is no stray light. Chemiluminescence may be up to 1,000 times more sensitive than fluorometry and 100,000 times more sensitive than absorption spectroscopy. In some instances single molecules have been detected. Bioluminescence is just an extremely efficient subclass of chemiluminescence, and it provides better reliability because the enzyme is naturally evolved to protect the photon emitter. Disadvantages of chemiluminescence include the limited number of donor acceptor pairs available, i.e., Renilla luciferase, bacterial luciferase, firefly luciferase, and aequorin, and the fact that the emission spectra of bioluminescent proteins are broad.

Phosphorescence

Phosphorescence is the emission of photons with a change in spin multiplicity. Because the process violates the spin selection rule, it should not be allowed, but selection rules are broken, and phosphorescence is commonly observed. More specifically, phosphorescence, as we saw above, occurs when the molecule returns to the electronic ground state from the excited triplet state by emission of a photon (i.e., a radiative transition between states of different multiplicities) (see Figures 7.90 and 7.91). Because the transition is spin forbidden and is often slow (i.e., long lifetime, with τ in the 10^{-4}–10-s range), it is also referred to as *delayed luminescence* or *afterglow*. It occurs at much longer wavelength compared with fluorescence. Phosphorescence is susceptible to O_2 and collisions with solvent molecules. Triplet states are rapidly deactivated under these conditions. Because of the long radiative lifetime of phosphorescent transitions, internal conversion and other radiationless transfers of energy compete so effectively with phosphorescence. For many molecules it is only observed at low temperatures or in rigid frozen media (e.g., solvent mixtures at 77 K) and in the absence of oxygen. As illustrated in Figure 7.117, photon absorption occurs between electronic levels with the same spin multiplicity.

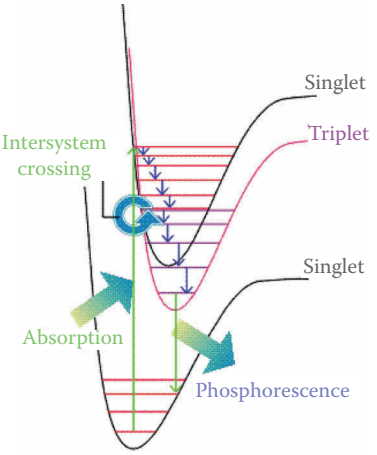


FIGURE 7.117 The phosphorescence process.

The radiationless transition between states of different multiplicity, shown here, is known as *intersystem crossing* (ISC). This may be followed by phosphorescence (weak emission from a long-lived state) to the ground electronic state. Fluorescence is measured much more often than phosphorescence because the lifetime of fluorescence (10^{-8}–10^{-4} s) is shorter than the lifetime of phosphorescence (10^{-4}–10^{2} s), allowing other processes to occur before a molecule has a chance to phosphoresce, thus making fluorescence more likely than phosphorescence.

Example Optoanalytical Devices

Introduction

To be able to compare electrochemical with optical sensors, we explore some examples of optical sensors here, just as we did for electrochemical sensors earlier in this chapter. Common types of optical sensors include fiberoptic, planar waveguide, evanescent wave, interferometric, and surface plasmon resonance (SPR). Under fiberoptic sensors, one further distinguishes between extrinsic and intrinsic sensors and direct and indirect ones. In an extrinsic optic fiber sensor, the fiber is only used as a transmission channel to take light to and from the sensing elements. In an intrinsic sensor, the fiber itself acts as a sensing element; a portion of the optical fiber cladding is removed and replaced with a chemically selective layer. The sensor is then placed directly into the media to be analyzed. Interaction of the analyte with the chemically selective layer creates a change

in absorbance, reflectance, fluorescence, or light polarization. A direct fiberoptic sensor measures the intrinsic optical properties of the analyte, and an indirect sensor measures the absorbance or fluorescence of an immobilized indicator dye or label. Planar waveguides are media in which the optical propagation is confined to a plane instead of a cylindrical fiber. Typical materials used for waveguides include plastic $\lambda > 450$ nm, glass $\lambda > 350$ nm, fused silica $\lambda < 350$ nm, and germanium crystals $\lambda > 1{,}000$ nm. The essential feature in interferometric sensing is the modulation of interference behavior by analyte binding. For example, the illumination of transparent thin-film coatings on reflective surfaces produces the so-called *Fabry-Pérot fringes*—alternating high- and low-intensity features—in which the resulting reflection spectrum and binding of an analyte to such film will shift these interference fringes. Evanescent wave sensors use the interaction with the electromagnetic field that extends a very short distance away from the surface of the light-guiding medium. The interaction of the evanescent wave with the sensing layer is used to get information on the analyte binding. It can involve IR (vibrational), UV-visible (electronic) absorption band, or Raman spectroscopic transitions. Surface plasmon resonance operates on the same principle as evanescent wave sensing, except that the planar waveguide is replaced by a metal/dielectric interface. Surface plasmons are electromagnetic waves that propagate between a metal and a dielectric material.

We encountered several of these types of optical sensors already in Chapter 5, where we introduced evanescent waveguide-type and SPR sensors. Additional optical sensors we review here are a fiberoptic-based continuous immunosensor, an interferometric immunosensor, a fiberoptic-based nanobead array for detection and analysis of DNA sequences, and PEBBLEs (**p**robes **e**ncapsulated **b**y **b**iologically **l**ocalized **e**mbedding) nanosensors. PEBBLEs is a general term that describes a family of matrices and nanofabrication techniques used to miniaturize optode technologies, thus filling a niche that lies between micro-optodes and free molecular probes.

Continuous Immunosensors—Optodes

A fiberoptical chemical sensor (FOCS) is often called an *optode* (also optrode), a word derived from optical electrode. An optode measures changes in the optical properties of a coating or membrane caused by the analyte. The sensors generally have a fiber for the light source and a return fiber to a spectrometer. If the sensor uses a phase change, it has a reference fiber and a sensing fiber. If the analyte is present, the optical parameter of the sensing fiber is changed. This creates a phase difference between the light traveling through the two fibers. If intensity is measured, instead of phase difference, the optode has an important advantage over electrochemical sensors in that no reference is needed.

The sensor consists of three main parts: a light source, optode, and detector. The main part of the sensor, the so-called optode, contains an appropriate indicator that changes its optical properties in dependence on the analyte. In basic fiberoptic sensor configurations, the active sensing region of the optical fiber may be located either on the end facet of the fiber (end-fiber configuration), on the side of the fiber (side-fiber configuration), or as an interrupting porous section inserted in the fiber (interrupted-fiber configuration). The end-fiber configuration is commonly applied for absorbance, luminescence, or fluorescence sensors, using the fiber as a light guide to and from the sample. The side-fiber configuration is usually applied for sensing schemes based on evanescent field measurements using the fiber as a waveguide and as the active transducer in sections with removed cladding. The interrupted-fiber configuration incorporates the active sensing region directly in the fiber by introducing a porous section. Thus, direct absorbance measurements are facilitated along with a large surface area. The fiberoptic immunosensor sensor shown in Figure 7.118 is an end-fiber configuration.

A key issue with immunosensors is their seemingly inherent irreversibility. Barnard and Walt.[57] introduced an ingenious way to make an immunoassay continuous by using the controlled release of labeled antibodies to maintain sensitivity. They used fluorescein-labeled antibody (F-Ab) and Texas Red-labeled immunoglobulin G antigen (TR-Ag) in a competitive immunoassay based on Förster

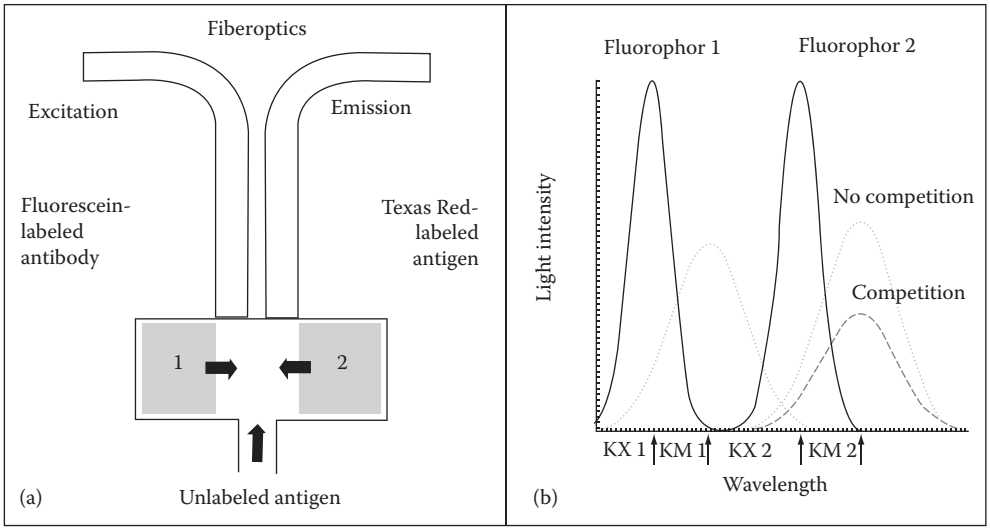


FIGURE 7.118 A new approach to immunosensing. A continuous immunomeasurement is enabled by implementing a polymer-release mechanism. Labeled antibody and antigen are released from polymer reservoirs, and fluorescence is measured by fiberoptics (a). The fluorescence emission from the second fluorophore (Texas Red) is reduced when unlabeled antigen competes with the reaction (b). (From Barnard, S. M., and D. R. Walt. 1991. Chemical sensors based on controlled release polymer systems. *Science* 251:927. With permission.[57])

resonance energy transfer (FRET) (see Figure 7.101). When the antibody (Ab) and antigen (Ag) bind, the two fluorophor labels are very close to one another, allowing nonradiative transfer of energy. When the fluorescein molecule is excited by light, it donates energy to the Texas Red fluorophor, enhancing its fluorescent light emission. To circumvent the irreversibility of the immunoassay, these researchers used a controlled-release delivery system capable of sustaining a constant release of fresh immunochemicals as shown in Figure 7.118. Specifically, they used ethylene-vinyl acetate as the controlled-release delivery system of reagents into the reaction chamber, i.e., the sensing region of the optical fiber. Unlabeled Ag diffused into the reaction chamber and competed with the Ag-TR for the available binding sites on the F-Ab, and bound species diffused out of the reaction chamber while a constant release of reagents maintained a constant concentration of fresh reagents in the reaction chamber. The fluorescence changes induced by the competition reaction were used to calculate the amount of unlabeled antigen in the sample solution. In cases where response time is not very critical but continuous monitoring is of major importance, this technique seems promising. A micromachined version of the sensor, shown in Figure 7.118, seems imminently feasible. This could involve micromachining a reaction chamber, release polymer reservoirs, and a waste dump, all connected with flow channels, and a window in the reaction chamber for the fiberoptics. Plastic injection molding or photoformed glass would be good approaches here.

Interferometric Immunosensor

The interferometric immunosensor the author and collaborators developed at SRI International is based on Langmuir-Blodgett technology.[58] As early as 1936, Irving Langmuir and Katherine Blodgett developed, in a most ingenious way, a method to measure film thickness comprising a small fraction of a wavelength of light. It consisted of building multilayers of barium stearate on a slide using a series of steps in a staircase fashion, as shown in Figure 7.119a. The method they used is known today as the Langmuir-Blodgett deposition method (see Volume II, Figure 8.13 in for a typical setup). Since then, much progress has been made in automating this deposition technology. The difference in thickness of two adjacent steps, as shown in Figure 7.119a, corresponds to a double layer of stearate (approximately 48 Å) and forms an interference-based color gauge that tracks thickness increases. When a film of unknown thickness is coated on the

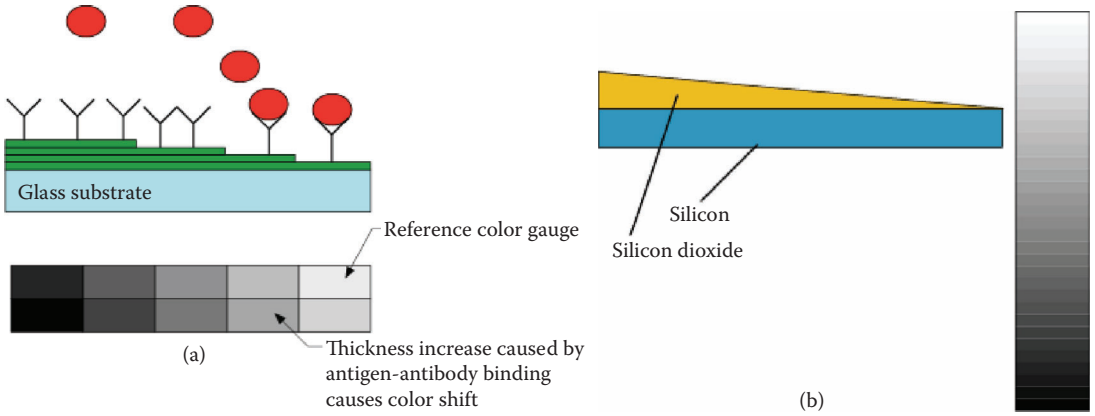


FIGURE 7.119 Optical immunosensors. (a) Staircase made by Langmuir-Blodgett technology (additive technique) or etching (substractive technique); (b) wedge made by pulling sample out of an etchant (substractive technique).

steps and is compared with an uncoated slide with steps, the eye can measure differences of thickness by means of color comparison down to about 10 Å. Greater sensitivity is obtained by illuminating films with sodium light rather than white light. Differences of thickness equal to 5 Å have been made plainly visible to the eye this way. This means that adsorbed layers of molecules or atoms having an optical thickness of 5 Å can be made visible to the eye without the aid of optical apparatus and can be measured with a probable error of about 2.5 Å. Some preliminary experiments were carried out with this simple step technique in the Madou laboratory while he was at SRI International laboratory at SRI International. When albumin was adsorbed on one-half of the staircase slide, the slide then was exposed to antialbumin. A thickness increase was observed via a color shift. What does this mean from the biosensor point of view? For many years, ellipsometers have been used to detect the deposition of antibody onto antigen preadsorbed on a flat reflecting surface. In this way, any molecule with a molecular weight greater than 60,000 (e.g., viruses, antibodies, proteins) could be detected. The ellipsometer is a very involved and expensive instrument (see Volume III, Chapter 6), whereas the system described above can perform the same task in a simple and inexpensive way. However, as pointed out earlier, the Langmuir-Blodgett deposition is a difficult manufacturing proposition. Fortunately, steps can also be built using etching techniques and using inorganic materials with optimum refractive indices for maximum interference effect.[58] Later, we found yet a better way to make interference slides: a substrate pulled at a uniform rate from an etchant solution forms a continuous optical wedge, making for a more elegant and faster manufacturing option than either Langmuir-Blodgett or staircase etching (Figure 7.119b). We believe the wedge approach is an excellent illustration of how micromachining could benefit immunosensor development by making inexpensive and disposable optical transducers possible. Whereas the wedge approach is more appropriate for the detection of larger molecules, the detection of smaller molecules is accomplished more easily with the electrochemical approach.

Bead Arrays for Detection and Analysis of DNA Sequences

Conventional DNA microarrays are manufactured by spotting or synthesizing probes at known locations onto 2D substrates, as we review in Volume II, Chapter 8 on chemical-, photochemical-, and electrochemical-based forming. Above, we discussed 3D liquid arrays (Figure 7.106) as an alternative for these planar gene chips with the idea of conjugating biomolecules to beads that are filled with various combinations of quantum dot (QD) types at various concentrations that can be used as "barcodes." Here we consider yet another option for DNA assaying, one that is already commercialized by Illumina (a San Diego biotech company) and is vying against Affymetrix, the industry pioneer, and other rivals, including Agilent Technologies, GE Healthcare, NimbleGen Systems, ParAllele BioScience, and Perlegen Sciences.

Illumina's BeadArray™ technology is based on the random self-assembly of a bead pool onto a patterned substrate.[59] The patterned substrate may be an etched flat surface (glass or Si) or a bundle of fiberoptic fibers with etched cavities to accommodate one bead per fiber. Illumina's fiber bundle-based chip-making process is illustrated in Figure 7.120.

This elegant design starts by making a pool of hundreds of thousands of tiny glass beads (3-µm diameter), each batch coated with one (among 1500) type of oligonucleotide, DNA fragments that will seek and latch onto a specific sequence of a DNA sample. Each bead has only one type of oligonucleotide attached to it but consists of hundreds of thousands of copies. Next, the batches of beads are mixed to form a mixed "pool" (see Figure 7.120, panel 1). When a bundle of 50,000 optical fibers is dipped into this pool (Figure 7.120, panel 2), a single bead adheres to the end of each fiber (Figure 7.120, panel 3). The beads occupy the wells in a random distribution, and each bead is represented by, on average, about 30 instances within the array. An array of 8 × 12, or 96 fiber bundles (i.e., 96 independent arrays) with 50,000 light-conducting strands per bundle, is constructed to be compatible with standard microtiter plates used to carry out sample preparation and enzymatic steps before exposure of the beads in the array to the sample (Figure 7.120, panel 4). As a part of the array manufacturing process at the company, a decoding process is implemented to map the precise location of a specific bead type on the array.[59]

The decoding strategy involves sequential hybridization of differentially labeled probes to an address site on the array.[59] The differential labeling uses three states: carboxyfluorescein (FAM), labeled green; cyanine 3 (Cy3), labeled red; and not labeled. During any given cycle of the process, a bead is green, red, or blank. Dye-labeled oligonucleotides hybridize to the address sequence of each probe type in the hybridization or "decoding" process (Figure 7.121). Labeled decode oligonucleotides are hybridized to the arrays at high concentrations, which allows for rapid (several minutes) hybridizations, followed by washing to remove nonspecific signal and background. If one assigns a number to each state—0 to blank, 1 to green, and 2 to red—then each cycle of the process generates a trinary digit. If we look at one hypothetical probe, then the first round may be red, which generates 2 for the first digit. The second round is green and generates 1, so the number is now 21. The third round is red, so the number is now 212. Each round just adds a new digit to the

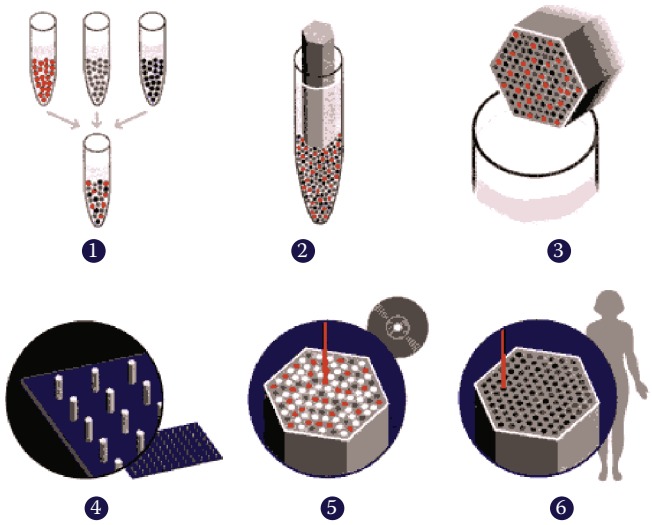


FIGURE 7.120 Illumina's BeadArray technology is based on the random self-assembly of a bead pool on fiber bundles; see text. (From Gunderson, K. L., S. Kruglyak, M. S. Graige, F. Garcia, B. G. Kermani, C. Zhao, D. Che, T. Dickinson, E. Wickham, J. Bierle, et al. 2004. Decoding randomly ordered DNA arrays. *Genome Res* 14:870.[59])

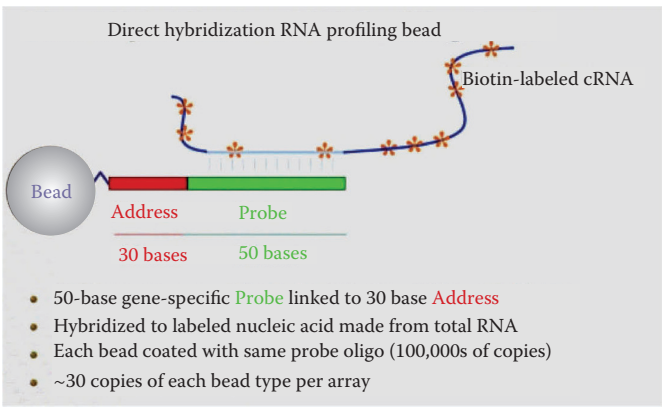


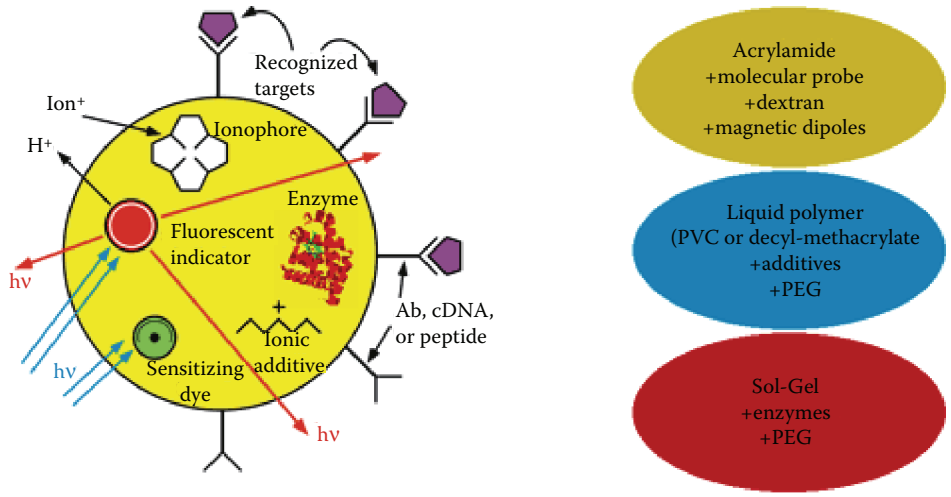
FIGURE 7.121 Gene-specific probes are attached to beads, which are then assembled into arrays. For simplicity, the diagram only shows one oligomer attached to the bead; actual beads have hundreds of thousands of copies attached per bead. The 50-base gene-specific probe is attached to a 30-base address. (From Gunderson, K. L., S. Kruglyak, M. S. Graige, F. Garcia, B. G. Kermani, C. Zhao, D. Che, T. Dickinson, E. Wickham, J. Bierle, et al. 2004. Decoding randomly ordered DNA arrays. *Genome Res* 14:870.[59])

number. This continues until there are sufficient digits to uniquely identify each probe. One digit can uniquely identify three probes (0, 1, or 2). Two digits can uniquely identify nine probes (00, 01, 02, 10, 11, 12, 21, 22, or 23). Three digits can identify 27 probes (we will leave anything further as an exercise for the reader!). Rehybridization continues until there are sufficient data to unambiguously determine the identity of each bead. The physical location of each bead type is thus mapped by using a decoding algorithm[59] that tracks oligonucleotides successfully hybridizing to a specific address. With this compiled information, the physical location of each bead is mapped, and this unique map of all the oligo on the chip is stored on a CD-ROM that is shipped to the customer with the chip. The decoding process also serves as a quality control step for each feature of each array. The address sequences are designed to have no significant homology to genomic sequences within the species queried. Customers put all this to work by dipping the array into a specially treated DNA sample. As the sample wets the beads, each oligo looks for the unique DNA sequence it is built to detect. If the sequence is present, the oligo chemically binds to it and then lights up when a finely focused laser is beamed through the array; if the sequence is absent, the oligo stays dark. The resultant array is compared with the unique map on the CD-ROM to determine which oligonucleotide-bead combination is in each well. The decoded arrays may be used for a number of applications, including gene expression analysis and genotyping.

A high degree of miniaturization can be achieved by exploiting the intrinsic size of the beads and patterned substrate. For example, the density of a randomly assembled 300-nm diameter bead array is more than 40,000 times higher than a typical spotted microarray.[60]

PEBBLE Sensors

With most electrochemical electrodes and optodes, when it comes to penetrating individual live cells, even the introduction of a submicrometer sensor tip causes biological damage. Individual molecular probes (free sensing dyes), on the other hand, are physically small enough to enter biological cells without damage. Unfortunately, they often suffer from chemical interference between the probe and the cellular components. To solve this problem, Kopelman[61] developed PEBBLE (**p**robes **e**ncapsulated **b**y **b**iologically **l**ocalized **e**mbedding) nanosensors. These are nanoscale spherical devices consisting of sensor molecules entrapped in a chemically inert matrix, as shown in Figure 7.122a. PEBBLE is a general term that describes a family of matrices and nanofabrication techniques used to



(a) (b)

FIGURE 7.122 (a) PEBBLE (**p**robes **e**ncapsulated **b**y **b**iologically **l**ocalized **e**mbedding) nanosensors. (b) Three different matrices are commonly used depending on the sensor materials. (http://www.umich.edu/~koplab/index.htm.)

miniaturize optode technologies, thus filling a niche that lies between micro-optodes and free molecular probes (naked indicator dye molecules). PEBBLEs are designed specifically for minimally invasive analyte monitoring in viable, single cells with applications for real-time analysis of drug, toxin, and environmental effects on cell function. The main classes of PEBBLE nanosensors are based on matrices of polyacrylamide hydrogel, sol gel silica, and cross-linked decyl methacrylate (Figure 7.122b).[62] These matrices have been used to fabricate PEBBLE nanosensors for H^+, Ca^{2+}, K^+, Na^+, Mg^{2+}, Zn^{2+}, Cl^-, O_2, NO, glucose, etc. that range from 30–600 nm in size.

The protective coating of a PEBBLE eliminates interferences, such as protein binding and/or membrane/organelle sequestration, which might alter dye response. Conversely, the nanosensor matrix also provides protection for the cellular contents, enabling dyes that might usually be toxic to cells to be used for intracellular sensing. In addition, the inclusion of reference dyes allows quantitative, ratiometric fluorescence techniques to be used. A number of delivery methods of PEBBLE nanosensors to single cells have been tried, including gene gun, picoinjection, liposomal delivery, and sequestration (phagocytosis and pinocytosis) into macrophages. The gene gun can best be thought of as a shotgun for nanoparticles. The gene gun can be used to deliver one to thousands of PEBBLEs per cell into a large number of cells very quickly (http://www.umich.edu/~koplab/research2/analytical/NanoScaleAnalysis.html). Cell viability is excellent: 98% viability compared with control cells. Commercially available liposomes can also be used to deliver PEBBLEs to cells. In this case, liposomes are prepared in a solution of PEBBLEs and then placed in the cell culture, where they fuse with the cell membranes and empty their contents (the PEBBLE-containing solution) into the cell. Dourado and Kopelman[63] detailed the specific advantages of these types of nanoscale dimension optical sensors. They argue that the absolute detection limit decreases with the PEBBLE radius r as r^3 (good!), and the response time is reduced as r^2 (good!). The signal-to-noise (S/N) ratio, on the other hand, decreases with r (bad!), but at least not with r^3 (luckily!) under standard working conditions. Other features that improve as sensors get smaller include required sample volume, invasiveness, spatial resolution, dissipation of heat in sensor and/or sample, toxicity, and materials cost. Features that typically worsen include fluorophore leaching and photodamage to sensor and/or sample.

Comparison of Optical versus Electrochemical Sensors

In this chapter, we gave examples of both optical and electrochemical immunosensors. In Table 7.21, pros and cons of both approaches are summarized. Several features make electrochemical techniques attractive in immunosensing (especially of small molecules). Because accuracy is achieved

Table 7.21 Comparison of Optical and Electrochemical Immunosensors

	Optical	Electrochemical
Instrument cost	Often expensive	Always inexpensive
Sensor cost	Fair	Low
Turbid solutions	Problematic	Usually no problem
Dynamic range	Limited	Very wide
Selectivity	Good	Fair
Sensitivity	Very good	Good
Simplicity of method	Often simple	Simple
Color	Sometimes problematic	No problem
Species size	Better for larger molecules	Better for smaller molecules
Electromagnetic interference (EMI)	No	Yes
Resistance to radiation and corrosion	Yes	No
Cross-talk	No	Yes
Ambient light	Problem (must be modulated)	No problem
Response curve	Sigmoidal	Nernstian (potentiometric) or linear

independent of color and turbidity of the solution, rapid measurements are often obtained with less tedious sample preparation. In principle, analysis of a whole range of analytes in matrices as complex as whole blood can be done. Electrochemical probes are easy to use as analytical tools for measuring ions and neutral molecules, typically in a range of 1 M to 1 μM. Examples of even wider dynamic ranges also exist. Coupled with chemical amplification (enzyme or liposome), lower detection limits of picomoles per liter have been attained. In addition, the sensor is inexpensive. Using optical techniques, one partner of the immunochemical complementary pair is typically immobilized onto a surface with some optical characteristics modified when binding takes place. In optical sensing, the signal is not subject to electrical interference, no reference electrode is required, and simultaneous measurement of two or more analytes is possible by measuring at two or more wavelengths. Whereas classical optical assays rely mostly on absorption spectroscopy, luminescence in the form of fluorescence or phosphorescence (depending on the excited state being single or triplet) is more important for immunosensors. Electrochemical detection tends to be better for small molecules, whereas optical techniques are better suited for large molecule detection.

From a more global perspective, comparing electrochemical with optical methods, we start by pointing out again that electrochemical methods are not affected by color and turbidity of the solution, are very low cost, and have the potential to do elemental speciation (e.g., Fe^{2+} vs. Fe^{3+}).

Acknowledgments

Special thanks to Dr. Mike Tierney, Dr. Sapna Deo, Dr. Jim Zoval, Robin Gorkin, and Jonathan Siegrist.

Questions

7.1: Show how the Butler-Volmer expression leads to an ohmic resistance at low over-potentials.
Thanks to Dhondup Pemba, UC Irvine

7.2: What is a Jablonski diagram?
Thanks to Elisa Morganti, UC Irvine

7.3: Calculate the decay time or ring-down time in cavity ring-down spectroscopy (CRDS).
Thanks to Arezoo Motavalizadeh Ardekani, UC Irvine

7.4: How does the electric double layer (EDL) arise and what are the four theoretical models that attempt to describe the structure of the EDL?
Thanks to Derek So, UC Irvine

7.5: Are ISFETs potentiometric or amperometric devices? Explain the difference between the two types of devices. What does MOSFET stand for and what are the differences between the electrical circuits of an ISFET and a MOSFET?
Thanks to Han Wu, UC Irvine

7.6: In infrared absorption spectroscopy one measures the absorption of the infrared radiation at specific wavelengths that match the energy of vibrational transitions of a molecule. Raman spectroscopy is a complementary technique also based on vibrational transitions. In what ways does Raman spectroscopy differ from infrared (IR) spectroscopy and what advantages does it offer over IR spectroscopy?
Thanks to Kiavash Koko, UC Irvine

7.7: Explain how the Fermi level of a metal electrode regulates electron transfer from the metal electrode to the redox molecules in solution
Thanks to Neil Olson, UC Irvine

7.8: What are two typical optical absorption spectrometers? Explain how they work. Compare these two types.
Thanks to Omid Rohani, UC Irvine

7.9: How do absorbance spectroscopy and fluorescence differ experimentally, and consequently what benefits does fluorescence have to offer?
Thanks to Christopher Clawson, UC Irvine

7.10: Match the following terms with the appropriate definitions:

(1) Fluorophore	(a) Crystals composed of materials from periodic groups II-VI, III-V, or IV-VI and can be conjugated to biomolecules

(2) Stokes shift	(b) Method used to measure the ability of a molecule to move around in a cell over time
(3) FRAP	(c) Intense fluorescent probe used in many biological studies
(4) Quantum dots	(d) Difference in wavelength between fluorescence excitation maximum and the fluorescence emission maximum
(5) Green Fluorescent Protein	(e) Molecule that can be excited by light to emit fluorescence

Thanks to Youssef Farhat, UC Irvine

7.11: Given the concentration of Cu^{2+} in a first solution, design an amperometric sensor to measure the concentration of any other unknown Cu^{2+} solution.
Thanks to Chuan Zhang, UC Irvine

7.12: What is the most commonly used type of spectroscopic technique? How does one representative technique of this type work? Does it scale well in the micro domain compared to other spectroscopic techniques?
Thanks to Irena Aleksieva, UC Irvine

7.13: What are the advantages of using ultra-micro electrodes (UMEs)?
Thanks to Megan Jacks, UC Irvine

7.14: The Ag/AgCl reference electrode is an electrode of the first or second kind? How is the potential of this electrode fixed constant?
Thanks to Mary Amasia, UC Irvine

7.15: Derive absorbance using $dP = -Pc\beta dx$, where c = molar concentration, P = light intensity, β = probability of a photon being absorbed.
Thanks to Kim Bui, UC Irvine

7.16: What does it mean that the standard reduction potential is an intensive property?
Thanks to Kevin Ko, UC Irvine

7.17: Why is the miniaturization of a reference electrode such a challenge?
Thanks to Michelle Tu, UC Irvine

7.18: Given the following galvanic cell and half reactions:
$Au^{3+} + 3e^- \rightarrow Au \qquad E^0 = 1.50$
$Tl^+ + e^- \rightarrow Tl \qquad E^0 = -0.34$
Determine the following:
(a) Overall cell reaction and E^0_{cell}
(b) $\Delta G°$ for the cell reaction at 25°C
(c) Calculate E_{cell} at 25°C when $[Au^{3+}] = 1.0 \times 10^{-2}$ M and $[Tl^+] = 1.0 \times 10^{-4}$
Thanks to Andrew Sumarto, UC Irvine

7.19: There are two carbon electrodes arranged in an electrochemical cell. On one hydrogen is evolving and on the other Ag is being deposited. 1 h after passing current, 95 cm^3 H_2 (g) (19°C, 99.19 KPa) is collected on the hydrogen electrode, while 0.8368g Ag is deposited on the Ag electrode. Use the given information to calculate the current through the circuit.
Thanks to Xi Wang, UC Irvine

7.20: Match each term with the proper definition:
Thanks to Muhammad Uneib, UC Irvine

7.21: What's the ratio of the van der Waals forces between two atoms at a distance of 1 nm and the same atoms at a distance of 3 nm?
Thanks to Elaheh Shekaramiz, UC Irvine

7.22: Under what conditions do oxide semiconductors or nonoxide semiconductors with a native oxide fail to function as pH sensors?
Thanks to Geethika Alapati, UC Irvine

7.23: Explain the differences between electrodes of the first kind and electrodes of the second kind.
Thanks to Nelson Jen, UC Irvine

7.24: Draw the equivalent electrical circuit of a metal/electrolyte interface with the electrode at a potential so that a redox reaction occurs.

7.25: Calculate the potential of a battery with a Zn bar in a 0.5 M Zn^{2+} solution and Cu bar in a 2 M Cu^{2+} solution.

7.26: Show in a cyclic voltammogram the transition from kinetic control to diffusion control and why it really happens.

7.27: Derive how the capacitive charging of a metal electrode depends on the potential sweep rate.

7.28: Calculate the potential difference between two Pt electrodes immersed in two different solutions connected via a salt bridge. The first solution has 1 M KCl and 10^{-1} M Fe^{2+} and the second has 1 M KCl and 10^{-5} M Fe^{3+}.

7.29: What are the requirements of a good reference electrode? What is a double junction reference electrode, and in what cases is it useful?

7.30: Define sensitivity, detection limit, and selectivity in an analytical method.

7.31: What is the effect of the following parameters on the quantum yield of fluorescence?
(a) Raising the temperature.
(b) Adding a heavy atom to the solvent.
(c) Lowering solvent viscosity.
(d) What is the physical process that makes phosphorescence rare?

7.32: In the last few years, applications of quantum dots have become increasingly popular among analytical chemists. Quantum dots are now commercially available for a variety of applications, and many examples of their uses can be found in the literature. Briefly describe what quantum dots are and describe their applications.

7.33: Briefly state the difference between fluorescence and phosphorescence.

7.34: In potentiometry:
(a) What is a "junction potential"?
(b) What does the acronym ISE stand for?
(c) What are the components of ISEs?
(d) Give an example of an ISE.

7.35: How does Raman spectroscopy work?
Thanks to Elisa Morganti, UC Irvine

7.36: In a fluorescence measurement setup, which emitted photons are monitored? Why?
Thanks to Omid Rohani, UC Irvine

7.37: Define photobleaching. Identify an analytical technique that uses photobleaching and describe what kind of information it can provide about the membrane of a cell.
Thanks to Kiavash Koko, UC Irvine

7.38: What problem of conventional multipass approaches does cavity ring-down spectroscopy (CRDS) overcome, and how does it do this?
Thanks to Christopher Clawson, UC Irvine

7.39: What are fluorescence, chemiluminescence, and bioluminescence spectroscopy? Give an example of each technique. Why are chemiluminescence and bioluminescence spectroscopy preferred over fluorescence spectroscopy?
Thanks to Irena Aleksieva, UC Irvine

References

1. Rigby, M., E. B. Smith, W. A. Wakeham, and G. C. Maitland. 1986. *The forces between molecules.* Oxford, UK: Oxford University Press.
2. Stone, A. J. 1997. *The theory of intermolecular forces.* Oxford, UK: Oxford University Press.
3. Barron, A. E., and H. Blanch. 1995. DNA separations by slab gel and capillary electrophoresis: theory and practice. *Separ Purif Meth* 24:1–118.
4. Ohshima, H., T. W. Healy, and L. R. White. 1982. Accurate analytic expressions for the surface charge density/surface potential relationship and double-layer potential distribution for a spherical colloidal particle. *J Colloid Interface Sci* 90:17–26.
5. Casimir, H. B. G., and D. Polder. 1948. The influence of retardation on the London-van der Waals forces. *Phys Rev* 73:360–72.
6. Dzyaloshinskii, I. E., E. M. Lifshitz, and L. P. Pitaevskii. 1961. General theory of van der Waals' forces. *Soviet Physics Uspekhi* 4:153–76.
7. Sukenik, C. I., M. G. Boshier, D. Cho, V. Sandoghdar, and E. A. Hinds. 1993. Measurement of the Casimir-Polder force. *Phys Rev Lett* 70:560.
8. Yao, S., M. Wang, and M. Madou. 2001. A pH electrode based on melt-oxidized iridium oxide. *J Electrochem Soc* 148:H29–H36.
9. Morrison, S. R. 1980. *Electrochemistry at semiconductor and oxidixed metal electrodes.* New York and London: Plenum Press.
10. Marcus, R. A., and N. Sutin. 1985. Electron transfers in chemistry and biology. *Biochim Biophys Acta* 811:265–322.
11. Degani, Y., and A. Heller. 1989. Electrical communication between redox centers of glucose oxidase and electrodes via electrostatically and covalently bound redox polymers. *J Am Chem Soc* 111:2357–58.
12. Heller, A. 1990. Electrical wiring of redox enzymes. *Acc Chem Res* 23:128–34.
13. Gregg, B. A., and A. Heller. 1991. Redox polymer films containing enzymes. 2. Glucose oxidase containing enzyme elctrodes. *J Phys Chem* 95:5976–80.
14. Heller, A. 1992. Electrical connection of enzyme redox centers to electrodes. *J Phys Chem* 96:3579–87.
15. Mao, F., N. Mano, and A. Heller. 2003. Long tethers binding redox centers to polymer backbones enhance electron transport in enzyme "wiring" hydrogeis. *J Am Chem Soc* 125:4951–57.
16. Madou, M. J., and S. R. Morrison. 1989. *Chemical sensing with solid state devices.* New York: Academic Press.

17. Camman, K. 1979. *Working with ion-selective electrodes.* New York: Springer-Verlag.
18. Fleischmann, M., S. Pons, D. R. Rolison, and P. P. Schmidt. 1987. *Utah conference on ultramicroelectrodes*.
19. Lin, Y. H., W. Yantasee, F. Lu, J. Wang, M. Musameh, Y. Tu, and Z. F. Ren. 2004. *Dekker encyclopedia of nanoscience and nanotechnology.* Eds. J. A. Schwartz, C. Contescu, and K. Putyera. New York: Taylor & Francis.
20. Madou, M. J., and S. R. Morrison. 1991. *Sixth International Conference on Solid-State Sensors and Actuators* (Transducers '91). San Francisco.
21. Anderson, J. L., K. K. Whiten, J. D. Brewster, T.-Y. Ou, and W. K. Nonidez. 1985. Microarray electrochemical flow detectors at high applied potentials and liquid chromatography with electrochemical detection of carbamate pesticides in river water. *Anal Chem* 57:1366–73.
22. Reay, R. J., R. Dadoo, C. W. Storment, R. N. Zare, and G. T. A. Kovacs. 1994. *Technical digest: 1994 solid-state sensor and actuator workshop.* Hilton Head Island, SC.
23. van der Putten, A. M. T., and J. W. G. de Bakker. 1993. Geometrical effects in the electroless metallization of fine metal patterns. *J Electrochem Soc* 140:2221–28.
24. Masuko, N., T. Osaka, and Y. Ito, eds. 1996. *Electrochemical technology: innovation and new developments.* Tokyo and Amsterdam: Kodansha and Gordon and Breach Science Publishers.
25. Dukovic, J. O. 1993. Feature-scale simulation of resist patterned electrodeposition. *IBM J Res Dev* 37:125–40.
26. Leyendecker, K., W. Bacher, W. Stark, and A. Thommes. 1994. New microelectrodes for the investigation of the electroforming of LIGA microstructures. *Electrochim Acta* 39:1139–43.
27. Bergveld, P. 1970. Development of an ion-sensitive solid-state device for neurophysiological measurements. *IEEE Trans Biomed Eng* 17:70–71.
28. Matsuo, T., and H. Nakajima. 1984. Characteristics of reference electrodes using a polymer gate, ISFET. *Sensors Actuators* 5:293.
29. Lauks, I., J. Van der Spiegel, W. Sansen, and M. Steyaert. 1985. *Third International Conference on Solid-State Sensors and Actuators (Transducers '85)*. Philadelphia.
30. Buerk, D. G. 1993. *Biosensors theory and applications.* Lancaster, PA: Technomic Publishing.
31. Madou, M. J., and J. Joseph. 1993. Immunosensors with commercial potential. *Immunomethods* 3:134–52.
32. Fresenius, W., and I. Luderwald. 2000. *Fresenius' journal of analytical chemistry.* New York: Springer.
33. Aizawa, M., S. Kato, and S. Suzuki. 1977. Immunoresponsive membrane. I. Membrane potential change associated with an immunochemical reaction between membrane-bound antigen and free anti-body. *J Membr Sci* 2:125–32.
34. Janata, J. 1975. An immunoelectrode. *J Am Chem Soc* 97:2915–16.
35. Janata, J., and G. F. Blackburn. 1984. Immunochemical potentiometric sensors. *Ann N Y Acad Sci* 428:286–92.
36. Collins, S., and J. Janata. 1982. A critical evaluation of the mechanism of potential response of antigen polymer membranes to corresponding antiserum. *Anal Chim Acta* 136:93–99.
37. Anzai, J.-I., and T. Osa. 1990. Langmuir-Blodgett membranes in chemical sensor applications. *Selective Electrode Rev* 12:3–34.
38. Joseph, J., A. Jina, and M. Madou. 1991. *Annual meeting of the AIChE.* Anaheim, CA.
39. Fleischmann, M., P. J. Hendra, and A. J. McQuillan. 1974. Raman spectra of pyridine adsorbed at a silver electrode. *Chem Phys Lett* 26:163–66.
40. Jeanmaire, D. L., and R. P. V. Duyne. 1977. Surface raman spectroelectrochemistry. Part I. Heterocyclic, aromatic and aliphatic amines adsorbed on the anodize silver electrode. *J Electroanal Chem* 84:1–20.
41. O'Keefe, A., and D. A. G. Deacon. 1988. Cavity ring-down optical spectrometer for absorption measurements using pulsed laser sources. *Rev Sci Instrum* 59:2544–51.
42. Paul, J. B., and R. J. Saykally. 1995. Cavity ringdown laser absorption spectroscopy of the jet-cooled aluminum dimer. *Chem Phys Lett* 242:395–400.
43. Engeln, R., G. Berden, R. Peeters, and G. Meijer. 1998. Cavity enhanced absorption and cavity enhanced magnetic rotation spectroscopy. *Rev Sci Instrum* 69:3763–69.
44. Pipino, A. C. R. 1999. Ultra-sensitive surface spectroscopy with a miniature optical resonator. *Phys Rev Lett* 83: 3093.
45. Pipino, A. C. R. 2000. Monolithic, folded resonator for evansecent wave cavity ring-down spectroscopy. *Appl Opt* March:20.
46. Chan, W. C. W., D. J. Maxwell, X. Gao, R. E. Bailey, M. Han, and S. Nie. 2002. Luminescent quantum dots for multiplexed biological detection. *Curr Opin Biotechnol* 13:40–46.
47. Chan, W. C. W., and N. Sie. 1998. Quantum dot bioconjugates for ultrasensitive detection. *Science* 281:2016–18.
48. Murray, C. B., D. J. Norris, and M. G. Bawendi. 1993. Synthesis and characterization of nearly monodisperse CdE(E=S,Se,Te) semiconductor nanocrystallites. *J Am Chem Soc* 115:8706–15.
49. O'Neil, M., J. Marohn, and G. McLendon. 1990. Dynamics of electron-hole pair recombination in semiconductor clusters. *J Phys Chem* 94:4356–63.
50. Gerion, D., F. Pinaud, S. C. Williams, W. J. Parak, D. Zanchet, S. Weiss, and A. P. Alivisatos. 2001. Synthesis and properties of biocompatible water-soluble silica-coated CdSe/ZnS semiconductor quantum dots. *J Phys Chem B* 105:8861–71.
51. Han, M., X. Gao, J. Z. Su, and S. M. Nie. 2001. Quantum-dot-tagged microbeads for mltiplexed optical coding of biomolecules. *Nat Biotechnol* 19:631–35.
52. Nicewarner-Pena, S. R., R. G. Freeman, B. D. Reiss, L. He, D. J. Pena, I. D. Walton, R. Cromer, C. D. Keating, and M. J. Natan. 2001. Submicrometer metallic barcodes. *Science* 294:137–41.
53. Miyawaki, A., J. Liopis, R. Heim, J. M. McCaffery, J. A. Adams, M. Ikura, and R. Y. Tsien. 1997. Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature* 388:882–87.
54. Chalfie, M., Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher. 1994. Green fluorescent protein as a marker for gene expression. *Science* 263:802–05.
55. Dickson, R. M., A. B. Cubitt, R. Y. Tsien, and W. E. Moerner. 1997. On/off blinking and switching behavior of single green fluorescent proteins. *Nature* 388:355.
56. Garcia-Parajo, M. F., G. M. J. Segers-Nolton, J. A. Veerman, J. Greve, and N. F. van Hulst. 2000. Real-time light-driven dynamics of the fluorescence emission in single green fluorescent protein. *Biophys J* 78:384A.
57. Barnard, S. M., and D. R. Walt. 1991. Chemical sensors based on controlled release polymer systems. *Science* 251:927–29.
58. Joseph, J., and K. Itoh. 1992. Etching interference slides. US Patent 5,169,599.

59. Gunderson, K. L., S. Kruglyak, M. S. Graige, F. Garcia, B. G. Kermani, C. Zhao, D. Che, T. Dickinson, E. Wickham, J. Bierle, et al. 2004. Decoding randomly ordered DNA arrays. *Genome Res* 14:870–77.
60. Michael, K. L., L. C. Taylor, S. L. Schultz, and D. R. Walt. 1998. Randomly ordered addressable high-density optical sensor arrays. *Anal Chem* 70:1242–48.
61. Xu, H., J. W. Aylott, and R. Kopelman. 2002. Fluorescent nano-PEBBLE sensors for the real-time measurement of glucose inside living cells. *Analyst* 127:1471–77.
62. Barker, S. L. R., H. A. Clark, and R. Kopelman. 2002. *Biomedical diagnostics science and technology.* Eds. W. T. Law, N. Akmall, and A. M. Usmani. New York: Marcel Dekker.
63. Kopelman, R., and S. Dourado. 1996. *Chemical, biochemical, and environmental fiber sensors VIII*. Ed. R. A. Lieberman. Bellingham, WA: SPIE-International Society for Optical Engineering.
64. Bergström, L. 1997. *Adv Colloid Interface Sci* 70:125.
65. Brini, M., P. Pinton, T. Pozzan, and R. Rizzuto. 1999. Targeted recombinant Aequorins: Tools for monitoring [Ca21] in the various compartments of a living cell. *Microscopy Research and Technique* 46:380–89.

Index

A

Abbe's limit
  grating spacing, 305
  grating spatial frequency, 305
  image formation, 304
  interference pattern, 304
  numerical aperture, 305
  optical grating, 304
  paraxial approximation, 305
  resolution, 305–306
  spatial coherence of light, 306
AC electrical conductivity
  ac conductivity, 85
  frequency dependency, 85–86
  momentum of electrons, 84
  optical frequencies, 84–85
Adenosine triphosphate (ATP), 49
Adiabatic electron transfer, 572
Amperometric sensors, 593–594
Anodic bonding. *see* Field-assisted thermal bonding
Atomic force microscope (AFM), 21
Avalanche diode, 252

B

Bead arrays, DNA sequences
  decoding algorithm, 624
  gene-specific probes, 623
  3D liquid arrays, 622
  quantum dot, 622
  random self-assembly, bead pool, 623
  rehybridization, 624
  sequential hybridization, 623
Beam/ray direction, 325
Bell's theorem, 210–211
Bioluminescence
  aequorin, 617–618
  BIAs, 618
  coelenterazine, 617
  definition, 616
  fireflies, 617
  organisms, 616–617
  scaling, 619
Bioluminescence-based assays (BIAs), 618
Birefringence
  dichroism, 327
  fast/extraordinary wave, 327
  isotropic media, 326
  observation, 327
  optically anisotropic media, 326–327
  ordinary/slow wave, 327
  pleochroism, 327–328
  polaroid, 328
Birefringent, 312
Blackbody radiation
  Planck and Rayleigh–Jeans models, 96–97
  Stefan–Boltzmann law, 95
  temperatures, 95
  Wien displacement law, 95
Body-centered cubic (BCC) lattice, 39–40
Bohr's model
  angular momentum, 104–105
  attractive force, 103
  correspondence principle, 108–109
  ionization energy, 103–104
  limitations, 109
  negative energy, 104
  quantum number
    fourth quantum number, electron spin, 106
    magnetic number, 105–106
    principal quantum number, 104–105
    properties, 107
    second quantum number, 105
Bose–Einstein distribution, 162–163
Boundary layers
  continuity equation, 456
  diffusion layer, 458
  laminar flow, 456
  normal velocity scale, 457
  Péclet number, 457
  Prandtl number, 458
  thermal boundary layer thickness, 457
Brewster angle, 311

C

Casimir force, 520
  Casimir–Polder force, 543
  colloidal solutions, 542
  Heisenberg's uncertainty principle, 542
  NEMS and MEMS, 543
  virtual particles, radiation pressure, 542
  ZPE, 541–542
Charge-coupled devices (CCDs), 288
Chemiluminescence
  3-aminophthalate ion, 615
  Bolognian stone, 614
  cold light, 614
  Cyalume, 615
  luminol, 615
  luminometer, 616
  oxalate ester, 615–616
  oxetane intermediate reaction, 616
  scaling, 618–619
Circular polarization, 312
Colored rainbow, 309–310
Complementary metal oxide semiconductor (CMOS)
  advantage, 263
  mass manufacturability and economics, 268
  oxide leakage and device isolation, 271–272
  scaling issues, 264
Compositional superlattice, 429
Compton scattering, 99–100
Conduction band minimum (CBM), 379
Continuum theory, breakdown
  diffusion-based H-separators, 459
  macroscale physics, 458
  no-slip and slip condition
    definition, 459
    slip in gases, 459–462
    slip in liquids, 462–463
    slip length, 459
Coulomb interactions
  cations and anions hydration, 522
  dielectric constants, 522
  electrostatic interaction, 523
  gravitational interaction energy, 523
  magnitude of induced dipole, 524
  polarizability, 524
Critical angle, 311
Crystallography
  applied shear stress, 66
  Bravais lattice, unit cells, and basis
    BCC and FCC, 39–40
    conventional/crystallographic unit cells, 39
    7 crystal systems, 39
    cubic lattice classification, 39
    lattice constant, 39
    primitive cell, 38
    three-dimensional (3D) array, 38
    translational vector, 38–39
    Wigner–Seitz primitive cells, 39–40
  Brillouin zones, 60
  Burgers circuit, 63–64
  color centers, 62–63
  common defects, 61–62
  configurational entropy, 61
  definition, 37
  dislocation movement, 65–66
  doping, 62
  edge dislocation, 63–64

Frank-Read dislocation multiplication mechanism, 68
Gibbs free energy, 61
grain boundaries, 37, 68–69
impurities removal, 61
line defects, 63
line dislocation effect, 66
line tension, 67
materials classification, 37–38
microstates, 61
Miller indices
adjacent planes, 47
determination, 45
100 family planes, 45–46
plane orientation, 45–46
364 plane, SC cubic lattice, 46
SC cubic crystal, 45
symbols, 45–46
mixed dislocation, 63–64
nanoparticle, 38, 69
n_{eq}/N values, Al and Ni, 61
pinned dislocation, 67
plastic deformation, 69
point and space groups
anisotropy, 41
coordinate system, 43
crystal-like structure, 40
ice crystals, 40, 42
ICSD tables, 45
linear physical properties, solids, 41–42
mirror reflection, 40
point symmetry operations, 40–41
pyroelectricity, 43–44
scalar-vector and-tensor effect, 43
screw axis and glide plane, 44–45
tensor components, 43
tensor equation, 41
vector-vector effect, 42
point defects, 61–62
reciprocal lattice
definition, 57–58
diffraction pattern, 56
Ewald construction, 59–60
explicit algorithm, 58–59
graphical presentation, 57
Schmid's law, 68
screw dislocation, 63–64
semiconductor devices, 38
shear strain and modulus, 65
slip planes, 66–67
solution strengthening, 67
strain hardening effect, 67
theoretical shear stress, 64–65
types of defects, 61
X-ray analysis
diffraction (*see* X-ray diffraction)
Fourier transforms, 47–48
yield strengths, 65
zinc whisker, 70
Crystal momentum, 156
Czochralski (CZ) crystal pulling method, 225–226

D

DC electrical conductivity
Drudes model
charge carrier's drift mobility, 80
constant average velocity, 79
drift velocity, 79
electron collisions, 77–78
free electron gas, 77
macroscopic boundary conditions, 78
Newton's second law, 79
Ohm's law, microscopic interpretation, 78–80
resistivity, 79–80, 82–83
scatter time, 79
fermions, 84
resistivity, 76
Thompsons concept, 76–77
de Broglie matter waves
Bohr's model, 113–114
matter waves, 109–110
wave packets
Fourier integral and superposition, 111
particle velocity, 112
phase velocity, 112
wave function, 110–111
wave particle duality, 112–113
Debye–Hückel approximation, 535
Debye parameter, 535
Dense wavelength division multiplexing (DWDM), 411
Density of state (DOS) function
bulk materials, 169–172
dimension D, 191
quantum dots, 189–190
quantum wells
bound-state energies, 179
1D confinement, 177–178
2D electron gas, 178
frequency dependence, 179
optical transitions, 178–179
quantum wires, 182–183
DEP. *See* Dielectrophoresis
Dichroic crystals, 312
Dielectric constants. *See also* Relative permittivity
Coulomb interactions, 522
generalized dispersion relation, 376
plasma oscillation
bound and conduction electrons, 88
dielectric function, 87–88
displacement factor, free electron gas, 86
Drude–Lorentz model, 88
electric displacement field, 87
longitudinal oscillations, 86
macroscopic polarization density, 86–87
plasma frequency, 87–88
restoring force, 88
polar dielectric polaritons, 375
surface plasmons, 340, 342
Dielectric function, 315
Dielectrophoresis (DEP)
conductivity and permittivity, 488
dielectric force, 487
dielectric polarization effects, 486
3D electrodes, 489
insulator DEP, 490
Listeria, 488–489
maximal purification efficiency, 490
maximal retention efficiency, 490
motion of neutral matter, 487
negative and positive DEP, 486–487
particle separation, 489
Diodes
Avalanche diode, 252
current-voltage plot, 247–248
Esaki diode, 252–254
light-emitting diodes
advantages, 250
bandgaps, absorption coefficient, 249
crystal growth, 249
energy level difference, 248–249
forward based p-n junction, 248–249
laser diodes, 250
molecular beam epitaxy, 249
nanocrystals, 250
photon-generating transition, 249–250
quantum confinement, 250
photodiode, 250–251
p-n junction, 246–247
Shockley equation, 247–248
solar cells, 251–252
Zener diode, 252
Dirac's model, 208–209
Donnan potential, 559
Doping superlattice, 429
DOS function. *See* Density of state function
Double-gate transistor (DGT), 272
Double refraction. *See* Birefringence
Drude–Lorentz model
atomic polarizability
amplitude, 352
bounded electrons, 351
damping force/friction loss, 352
definition, 352
driven harmonic oscillator, 352
equation of motion, 352
FWHM, 353–354
group velocity, 354
Heisenberg uncertainty principle, 354
quality factor, 354
refractive index, 353
relative static permittivity, 354
single atom polarization, 352
displacement field, 350
free electron gas, 350
free electrons, 351
local field corrections, 358–359

multiple resonances, rarified media
atomic resonances, 354
complex refractive index, 355
dispersion curves, 356–357
Doppler broadening, 356
fast decay, 356
oscillator strength, 355
radiative relaxation, 355–356
rarified medium susceptibility, 354
resonant transitions, 355–356
spectral regions, 356–357
total classical polarization, 355
oscillators, 351
plasmon, 351
resonant frequency, 351
restoring force, 351
Drude model, 338
Dulong–Petit law, 89–90

E

EDL. *See* Electrical double layer
Effective mass, 155, 157–158
Elastic scattering
Mie scattering
back scattering maxima, 396
efficiency, 395–396
fogbows, glories, and coronas, 395–397
incident waves, 395
Mie phase function, 396
polar plots, 395
sunlight, 396
Rayleigh scattering
angular distribution, 389
Cartesian and spherical coordinates, 389
dipole, 388
harmonic motion of electrons, 388
intensity, 389–390
isolated dipole oscillator, 388
oscillating charges, 388
phase function, 389–390
planes, polarization, 389
total scattering power, 391
why the sky is blue, 391
Electrical double layer (EDL)
colloidal dispersions, 539–541
definition, 532
G–C model, 534–536
Grahame model, 537
Helmholtz model, 533–534
intermolecular forces, 518
liquid/solid interfaces, fractals, 532
Stern model, 536
zeta potential and electrokinetic phenomena, 537–539
Electric flux density. *See* Maxwell's displacement
Electric polarization, 310
Electrochemical and optical analytical techniques
attractive and repulsive forces, 517
electrochemistry (*see* Electrochemistry)
intermolecular forces
boiling point, 519
bonding, ionic, and nonbonding forces, 519–520
Casimir force (*see* Casimir force)
Coulomb interactions, 522–524
definition, 518
EDL, 518
electrostatic forces (*see* Electrical double layer)
HCl molecule, 518–519
hydrophilic *vs*. hydrophobic interactions, 519
interparticle interaction energy (U) and interaction forces (F), 521–522
Lennard–Jones potential, 528–529
nonbonding forces, 519
Van der Waals forces (*see* Van der Waals forces)
viscosity, 520
liquid/solid interface, 518
optical spectroscopy
absorption spectroscopy and cavity ring-down spectroscopy, 599–600
bead arrays, DNA sequences, 622–624
bioluminescence (*see* Bioluminescence)
chemiluminescence (*see* Chemiluminescence)
continuous immunosensors, optodes, 620–621
definition of luminescence, 601
Fabry–Pérot fringes, 620
fiberoptic sensors, 619–620
interferometric immunosensor, 621–622
intrinsic sensor, 619
IR and Raman spectroscopy, 598–599
PEBBLE sensors, 624–625
photoluminescence (*see* Photoluminescence)
quantum mechanics, 594–595
surface plasmon resonance, 620
UV/VIS absorption spectroscopy, 595–598
optical *vs*. electrochemical sensors, 625–626
Electrochemistry
amperometric immunosensor, 593–594
cyclic voltammetry
concentration and activation overpotential, 579
concentration gradient and polarization, 578
cyclic voltammograms, 579–581
diffusion-limited current, 579
Faradaic peak current, 581
limiting current density, 578
Nernst equation, 579
potentiostat-controlled three-electrode system, 577
definition, 543
dynamic equilibrium, 565
electrochemical potential, 566
electrodes of first kind, 555
electrodes of second kind
Ag/AgCl electrode, 555–556
glass pH electrodes, 556
ideal nonpolarizable electrode, 555
IPE, 556
ISE, 557
reference electrode, 555–556
SensIrOx, 557
solid-state pH sensor, 557
standard hydrogen electrode, 556
Faraday's law, 548
glucose sensor, 593
interface processes
counter electrode, 544
Daniell cell, 547
electrochemical cell, 544–545
electrode reactions, 544
electromotive force, 547
Faradaic and non-Faradaic process, 548
half-cell reaction, 545, 547
junction potentials, 545
nonaqueous electrochemistry, 548
redox reactions, 545
salt bridge, 544–545
SHE, 545–546
standard reduction potential, 546
stoichiometry coefficient, 547
ISEs (*see* Ion-selective electrodes)
mass and charge transport
conductance, 564
diffusion-limited current, 563
diffusion, migration, and convection, 563
molar ionic conductivity, 564–565
positive and negative ions, 565
terminal velocity, 564
transport number, 565
membrane-based gas-sensing probes, 562–563
membrane/electrolyte and mixed conductor/ electrolyte interfaces, 577
metal/electrolyte interface (*see* Metal/electrolyte interface)
mobile charge carriers, 543–544, 554
Nernst equation
Ag/AgCl reference electrode, 551–552
chemical potential, 550
dynamic equilibrium, 550
reference electrode, 552
potentiometric sensors (*see* Potentiometric sensors)
scaling, 584–585

semiconductor/electrolyte and insulator/electrolyte interface
electron equilibrium exchange current, 577
electron transitions, 575
energy diagram, 575–576
flat band potential, 576
forbidden bandgap, 575
free charge carriers, 574
ISFET, 577
isotropic and anisotropic etching, 575
reference electrode, 576
solid/electrolyte interface, 565
thermodynamic equilibrium, 566
thermodynamic interpretation, electromotive force
activity, 549–550
fugacity, 549
Gibbs function, 548
spontaneous process, 549
standard ambient temperature, 549
three-electrode cell configurations, 553–554
two-electrode cell configuration, 552–553
UME
advantages, 583
CNT arrays, 583
Cottrell equation for planar diffusion, 582
diffusion layer thickness, 581–582
diffusion-limited current, 582
Kelgraf electrode, 583
LIGA process, 584
microelectrodes, 582–583
nonlinear diffusion, 584
types, 581
Electrokinetic effects
alternating currents
applications, 493–494
DEP (*see* Dielectrophoresis)
particle levitation, 491–492
ROT, 486, 492
scaling considerations, 492–493
TWD, 490–491
direct current
charged particles, 476
electrokinetic injection, 481–482
electro-osmosis, 478–480
electrophoresis, 477–478
flow profiles, 480–481
fluidic channels, scaling, 483–484
IOF, 484–485
open electrophoresis, 485–486
planar fluidic networks, 482–483
surface electrophoresis, 486
Electromagnetic (EM) waves
boundaries
conditions, lossless media, 334
dimensionless reflectivity coefficient, 335
dissipative process, 336
Fresnel reflection, 334
integral equations, 334
Kramers–Krönig relation, 336–337
normal components, 334
optical coefficient properties, 337
planar interface, 335
reflectance and transmittance, EM radiation, 333
reflectance definition, 336
semiconductors, 337
tangential E field, 334–335
wave reflection and transmission, 334
complex refractive index and wave vector
absorption and propagation coefficient, 331–332
dispersion relation, 330
electric fields, 330
electric susceptibility tensor, 332
Kramers–Krönig relation, 331
optical intensity, 331
traveling damped wave, 331
energy (Poynting vector), 328–329
momentum, 329–330
Electron and hole mobilities, 159
Electron energy loss spectroscopy (EELS), 338
Electronic Numerical Integrator and Computer (ENIAC), 7
Electroosmosis, 537, 539
Electrophoresis, 537, 539
Electrorotation (ROT), 486, 492
Electrostatic bonding. *See* Field-assisted thermal bonding
Enzyme-linked immunosorbent assay (ELISA), 589–590
Esaki diode, 252–254
Euler equation, 439
Extended gate field-effect transistors (EGFET), 588
Extrinsic semiconductor
binding energy, 235
Bohr-like doping model, 236
Bohr radius, 235–236
conduction band, 236
conductivity, 234–235
donors and acceptors, 235
impurity levels, 236
vs. intrinsic, 236

F

Face-centered cubic (FCC), 39–40
Fanning friction factor, 455–456
Faradaic process, 544
Far-field diffraction, 302–303
Fermi quantities, 191–192
Fermi's golden rule (FGR)
absorption coefficient and transition matrix, 194
direct and indirect bandgaps, 193–194
energy surface, K-space, 193
incident photon flux, 192
joint density of states, 192
Si, absorption spectrum and band structure, 193
transition rate, 192
Fermi wavelengths, 365
Field-assisted thermal bonding, 14–15
Field effect transistor (FET), 10
FinFET approach, 273
Fluidics
centrifugal fluidic platform, CD fluidics
burst frequency, 497
capillary forces, 495, 497
coriolis force, 489–499
hydrophilic and hydrophobic valves, 496
LabCDT, 495
mechanical, osmotic, and acoustic pumping, 499–500
as microscope, 497–499
pressure, 496
prototyping methods, 495
pumping force, 497
two-point calibration disc, 497–498
continuum theory, breakdown (*see* Continuum theory, breakdown)
definition, 436
electrowetting, 494–495
fluid propulsion (*see* Electrokinetic effects)
forces at interfaces
capillary action, 466–467
definition of interface, 463
drops and bubbles, 464–465
surface tension, 463–464
wetting and contact angle-beading, 465–466
low Reynolds number fluids
micromixers (*see* Micromixers)
mixing, stirring, and chaotic advection, 468–469
Péclet number, 467–468
macroscale laws, fluid flow
absolute and kinematic viscosity, 437
Biot's number, 453
boundary layers, 456–458
continuum hypothesis, 439–440
density, 436
dynamic and average viscosity, 436
Euler equation, 439
friction factor, 454–456
heat conduction (*see* Heat conduction)
inviscid, 438
laminar, turbulent, and transitional flows, 439
Newtonian *vs.* non-Newtonian fluids, 437
Newton's law, 437
no-slip condition, 440–441

NSEs (*see* Navier–Stokes equations)
Reynolds number, 453–455
viscous regions, 438
micro-and nanomachined structures, 436
microchambers–microreactors
Damköhler numbers, 474–476
electrochemical reactions, 476–477
scaling, analytical separation equipment
amperometric technique, 499
band broadening, column, 502–505
calculated parameter sets, 505, 507
capillary electrophoresis, 508–509
detector miniaturization, 505
diffusion-controlled separations, 507–508
droplet microfluidics, 509–510
electrical Bodenstein number, 506
fluid handling, 500–501
Fourier numbers, 506
gas and liquid chromatography, 501
microdomain, 499–500
minimum migration time, 508
Péclet number, 506–507, 509
power per unit length constant, 508
pressure-driven chromatography, 509
proportionality factors, 507–508
reduced parameter set, 507
separation chemistry, 501–502
separation efficiency, 508
Fluorescence recovery after photobleaching (FRAP), 606
Fluorophores
chelation, 606
definition, 605
fluorescein molecule, 605
fluorescent dyes, 606
fluorescent probes, 606–607
FRET, 607–608
mauve and mauveine, 605
optical saturation, 608
photobleaching, 606
Focal length, 303
Forster resonance energy transfer (FRET), 607–608
Four-band model, 158–159
Fourier space. *See* Reciprocal space
Fourier transform process, 304
Fraunhofer diffraction pattern. *See* Far-field diffraction
Fresnel diffraction. *See* Near-field diffraction
Friction factor
force balance, 455
frictional resistance, 454
friction coefficient, 455
Moody friction factor, 455–456
Poiseuille number, 456
Reynolds numbers, 456
shear stress, 454–455
Full width at half maximum (FWHM), 353–354

G

Generalized dispersion relation
dielectric constant, 376
FEG *vs*. frequency, 376–377
forbidden gap, 377–378
group velocity, 376
light line, 376
light, polar solid, 377–378
photons and phonons coupling, 377–378
transverse electromagnetic waves, plasma, 376–377
Geometric optics, 308
Gouy–Chapman (G–C) model, 534–536

H

Hall coefficients
electron and hole charge carriers, 94
FEG predictions, 94–95
Hall resistance, 94–95
Hall voltage, 94
Lorentz force, 93–94
magnetic field, 93
Harmonic oscillators, 117–119
Heat conduction
Fourier's law, 450–451
Newton's cooling law, 451–452
radiation, 452–453
scaling laws, 450
Heat transport and capacity
classical and quantum oscillators, 197
Debye's heat capacity equation, 202–204
Einstein's heat capacity equation, 202
electrical *vs*. thermal conductance, 206–207
electron gas contribution, 195–196
Lattice contribution, 196–197
phonons
1D diatomic lattice, 200–201
1D monoatomic lattice, 197–200
radiation interaction, lattice modes, 201–202
quantized thermal conductance, 205–206
quantized thermal transport, 204–205
Heisenberg's uncertainty principle (HUP)
Casimir effect, 115
circular motion particle, 115
electron position, 114–115
graphic presentation, 115
noncommutability, 208
quantitative analysis, 115
wave function, 115–116
zero-point energy, 115
High electron mobility transistors (HEMTs), 128

I

Ideal lens
evanescent waves, 422–423
Fermat's principle, 422
near-field superlens, 423
Pendry lens, 421
resolution, 420–421
traditional *vs*. Veselago lens, 421–422
Ideal polarizable electrode (IPE), 556
Image formation, 303–304
Inorganic crystal structure database (ICSD), 45
International technology roadmap for semiconductors (ITRS), 31–32
IOF. *See* Isoelectric focusing
Ion implantation, 229, 233–234
Ion-selective electrodes (ISEs)
crystalline solid-state ISEs, 560
definition, 557
electrochemical cell potential, 558
glass electrodes, 558–560
LaF_3 crystal lattice, 557–558
polymer-based ISEs, 560–562
selectivity coefficient, 558
solubility and conductivity, 558
Ion-sensitive field-effect transistors (ISFET), 577, 585–587
Isoelectric focusing (IOF), 484–485
ITRS. *See* International technology roadmap for semiconductors

K

Kirchoff's law, 333
k-/momentum space. *See* Reciprocal space
Krönig–Penney model, 135
Bloch functions, 152
periodic lattice potential, 151
Schrödinger equation, 151
solution analysis, 152–155
k-vectors, 323

L

Laser tweezers, 330
LEDs. *See* Light-emitting diodes
Left-handed materials (LHMs)
development, 420
Doppler effect, 417
Drude model, 418
effective refractive index, 420
electric fields, 417
embodiments of structures, 418–419
magnetism, 418
Maxwell's equations, 325–326
microwave communications, 420
negative refraction, 420

Pendry's split rings, 418–419
periodic structure, 418
plasma frequency, 418
refraction, ε-μ plane, 417
SRR, 419
wave propagation, 420
Light-emitting diodes (LEDs)
advantages, 250
bandgaps, absorption coefficient, 249
crystal growth, 249
energy level difference, 248–249
forward based p-n junction, 248–249
laser diodes, 250
molecular beam epitaxy, 249
nanocrystals, 250
photon-generating transition, 249–250
photonic crystal optic fibers, 413–414
quantum confinement, 250
Light line, 322, 324
Light scattering
extinction coefficient, 387
light wavelength, 388
Mie/nonmolecular scattering, 387
photon flux, 386–388
Rayleigh scattering, 387
visible particles, different radii, 387
Lorentz condition for potentials, 321
Lorentz–Lorenz relationship, 358
Lorentz model for dielectrics
nonpolar dielectrics, 373
polar dielectric polaritons
broad reflection band, 373–374
dielectric constant, 375
KCl, high-energy transitions, 373–374
Lyddane–Sachs–Teller relationship, 375
optical phonons, 373
polariton, definition, 374
refractive index, 375–376
resonant frequency, 375
Restrahlen band, 376
total polarization, 374
Lorentz model for semiconductors
color and bandgap, 378
emission of photons, 381–382
excitons
Bohr model, 383
Bohr radius, 383
bulk materials, 383–384
definition, 382
energy levels, 383
Frenkel excitons, 382–383
Mott–Wannier excitons, 382–383
reduced dimensionality structures, 384
Rydberg energy, 383
stable excitons, room temperature, 384
thermal energy, 383
ultimate recombination process, 382
wave vector, 382
free carrier absorption, 378–379
light absorption, bandgap semiconductors
absorption α, Si and GaAs, 380–381
band-to-band absorption, 379–380
Cerenkov radiation, 381
Frölich interaction, 380
modulation spectroscopy, 380–381
photon wave vector, 380
VBM and CBM, 379
optical spectroscopy, 378
semiconductor absorption processes, 384–386
valence electrons, 378
Lorentz transformation
electromagnetic radiation, 345
Galilean addition of velocities, 346
Galilean transformation, 346
Lorentz contraction, 347
Lorentz–FitzGerald contraction/aether squeeze, 345
Lorentz's photography, 345
Michelson–Morley experiment, 346–347
optical phenomena, 346
reference frame, 346
relativity, 346
speed of light, 345
Low-pressure chemical vapor deposition (LPCVD), 243

M

Mach number, 438
Mallory bonding. *See* Field-assisted thermal bonding
Masers
amplification, stimulated emission of radiation, 423
energy levels, 424
incident photon with $E \geq E_g$, 423
laser diode, 425
lasing process, 424–425
metastable states, 424
multimode, 426
population inversion, 423
resonator, 425
VCSELs, 426
Maxwell–Boltzmann distribution, 81–82
Maxwell's displacement, 315
MEMs. *See* Micro-electromechanical systems
Metal/electrolyte interface
phenomenological description, exchange current
Butler–Volmer equation, 567–568
charge transfer resistance, 568
dynamic equilibrium, E^0, 566
Gibbs energy, 567
mixed potential, 566
Nernst expression, 567
overpotential, 566
Tafel regime, 567–568
semiquantitative description, exchange current
current under polarization, 570–571
density of state functions, 569–570
electron tunneling transfer rate, 571–574
Fermi–Dirac distribution functions, 570
Fermi level, 568
forward and backward currents, 570
Frank–Condon shift, 569
microreversibility, 570
redox molecules, 568
reorganization energy, 569
time-fluctuating energy levels, 568–569
Metalorganic chemical vapor deposition (MOCVD), 249, 428–429
Metal-oxide-semiconductor field-effect transistor (MOSFET)
carrier mobility limitations, 268–270
channel-length modulation effect, 263
channel pinch-off, 263
channel resistance, 261
constant field and constant voltage scaling, 266–267
drain current, 262
evolving materials and device structures, 273–274
mass manufacturability and economics, 268
metal oxide semiconductor capacitor
accumulation, 256–257
depletion, 257–258
flatband voltage, 258–259
pMOS and nMOS, 259
p-type semiconductor, 256–257
strong inversion, 258
threshold voltage, 259–260
nMOSFET, 259–261
oxide leakage and device isolation, 271–272
physical structure, 256
pinch-off process, 262–263
power budget, 270–271
scaling issues, 264–266
semiconductor, 256
threshold voltage control, 261–262
transistors current-voltage curve, 263
Metamaterials. *See* Left-handed materials; Right-handed materials
Microchambers–microreactors
Damköhler numbers
Biot number, 476
convective/advective transport, 475
definition, 475–476
diffusion coefficient, 474–475
Fick's second law, 474
Fourier number, 475
electrochemical reactions, 476–477

Micro-electromechanical systems (MEMs)
acronyms, 17
aspect ratio, 14
band theory, 2
Electronic Visions group, microscope, 16
Feynman, Richard, 13
field-assisted thermal bonding, 14–15
Gabriel, Kaigham, 16–17
gas chromatograph, Si wafer, 15–16
history line, 17–20
integrated photonic mirror array, 16
isotropic and anisotropic etching profiles, 14
marketing, 17
micromachining, 13, 15
microsystem technology, 13
miniaturize optical components, 3
NovaSensor founders, 16
optical properties, 2
piezoresistance, 13
piezoresistive pressure sensor featuring, 15
quantum mechanics, 2
sensitivity, optical sensing techniques, 3
Si accelerometer, 15
single-crystal Si band structure, 2
Si single-crystal gauge element, 13–14
Si substrate, 2
soft lithography process, 17
standing waves, 2D electron gas, 2
surface micromachining process, 16–17
Micromixers
active mixers
chaotic advection, 472
flapping bilayers, 473
flexible propelling structures, 473
hydrodynamic focusing, 472
multiple-hinged systems, 473
PPy/Au actuator, 473–474
Reynolds numbers, 473
diffusion coefficient, 468–469
diffusion constant, 468
molecular diagnostics, 469
passive micromixer
baker's transformation, 472
chaotic advection, 471–472
diffusion times and lengths, 469
diffusive extraction, 470
distribution networks, 469–470
Einstein–Smoluchowski equation, 469
H-filter, 470–471
multilaminated flow down-scales, 469
Reynolds number fluids, 471–472
staggered herringbone mixer, 471
transfer network, 470
T-shaped diffuser/mixer, 470
MOCVD. *See* Metalorganic chemical vapor deposition
Moody friction factor, 455–456
MOSFET. *See* Metal-oxide-semiconductor field-effect transistor

N

Nanocrystal semiconductors. *See* Quantum dots
Nano-electromechanical systems (NEMs). *See also* Micro-electromechanical systems
AFM, 21
Ashkin's report, 22
Buckminsterfullerene C60/buckyball, 22–23
carbon nanotubes, 23–24
Chu, Steven, 22
Davis discovery, 20
definition, 20
Dekkers demonstration, 23
dendrimers, 22
Feynman machines, 21
full-grown *Morpho rhetenor*, 25–26
history-line, 21
Iijima's discovery, 22–23
Kondo effect, 27
Lieber's report, 24
marketing, 28
metamaterials, 26
milestone chart, nanotechnology, 28–30
nanochemistry, 21
nanofabrication/nanomachining, 20–21
nanograph, Dr. M.Roco, 28
nanotech promoters, 28
optical scattering and gradient forces, 22
quantum dots, 23–24
refraction, 26–27
rotaxanes, 25
STM, 21–22
top-down gedanken experiment, 21
Veselago's prediction, 26
wavelength limitation, 22
yablonovite, 26
Navier–Stokes equations (NSEs)
Bernoulli's equation, 446–447
conservation of mass equation, 438
conservation of momentum, 438
continuity equation, 441
Couette shear flow, 448–449
3D vector equations, 437
Hagen–Poiseuille's law, steady duct pipe flow, 449–450
hydrostatics, 446
momentum equation
convective fluid transport, 441
falling sphere (I), viscid regime, 443
heat transfer problem, 441
scaling parameters, 442–443
total change of momentum, 442
x-momentum, 442
unidirectional flow, two infinite plates, 447–448
viscid, inviscid, and Froude flow
Boltzmann's constant, 444
diffusion constant, 444
drag coefficient, 444–445
falling sphere (II), inviscid flow, 445
Froude number, 446
gravitational forces, 446
Reynolds number, 444–445
Near-field diffraction, 302–303
Near-field scanning optical microscopy (NSOM). *See* Plasmonic applications 3
NEMs. *See* Nano-electromechanical systems

O

Optical properties of metals
conductors, EM radiation
absorption coefficient, 361
AC conductivity, plasma, 361
complex dielectric constant, 361
penetration/skin depth, 362
reflectance, 360–361
Drude–Lorentz theory
bound and conduction electrons, 363
free electrons and interband transitions, 364
ITO, 365
photon energy, 364
effective dielectric constant, 360, 363
energy *vs.* reflectance, 360
nanoparticles
absorption spectrum, 369
alchemists, 367
aspect ratio, 369–370
electromagnetic radiation, 367–368
electromagnetic wave interaction, 367
electron density, 368
energy absorption, 368
extinction coefficient, 369
gold and silver, 367
gold nanocages, 370
localized plasmons, 367
Lycurgus Cup, 366
metal shell, 370
nanoshell resonance frequency, 370
parameter α, 368
particle size, 367
quantum effects, 365–366
Renaissance pottery, 366
resonance, 369
resonance frequency and condition, 370–371
scattering, 368
surface plasmon bands, 367

plasma frequency, 359–360
reflectance curves, 360
surface plasmon applications II, 370, 372–373

P

Particle distributions functions
average interparticle spacing, 160
Bose–Einstein distribution, 162–163
de Broglie temperature, 160
Fermi–Dirac distribution
Boltzmann approximation, 160
electronic velocity vector component distribution, 161–162
energy and temperature, 161
exclusion principle, 160
Fermi energy, 161
Fermion and boson particle, 160–161
function of temperature, 161
heat capacity, 162
Pauli's principle, 160
thermal de Broglie wavelength, 159–160
Photodiode, 250–251
Photoelectric effect
classical wave model, 97
Einstein's light particles photons, 98–99
experimental setup, 97
frequency of incident light, 97–98
Planck's hypothesis, 98
Photoluminescence
fluorescence
analytical chemistry and scaling, 604–605
autofluorescence, 603
excited singlet and triplet states, 601–602
fluorescence quantum yield and lifetime, 603
fluorescent proteins, 613–614
fluorophores (*see* Fluorophores)
Franck–Condon principle, 602
Jablonski diagrams, 601
light-emitting phenomena, 602
quantum dots (*see* Quantum dots)
radiative processes, 601
reagent background, 603
singlet ground state, 601–602
spectrofluorometers and microplate readers, 604
Stokes law, 602
Stokes shift, 603
phosphorescence, 619
2D and 3D Photonic crystals manufacturing
holographic lithography, 409
inverse opal, 408
LPCVD, 409
macroporous silicon, 408
microwave range, 407–408
self-assembled silica opals, 409
synthetic opals, 408
two-photon lithography, 409
Yablonovite, 407
Photonics
beyond Maxwell
Einstein's general relativity, 348–349
Einstein's special relativity, 347–348
Lorentz transformation, 345–347
diffraction and image resolution
Abbe's limit, 304–306
far field and near field, 302–303
image definition, 303–304
Rayleigh's limit, 306–308
Galilean transformation, 299–300
lasers
masers, 423–426
quantum mechanics, 426–431
light interaction, small particles
geometric scattering, 397
light scattering, 386–388
Mie scattering-elastic scattering, 395–397
plasmonic applications 3, 397–402
Raman scattering-inelastic scattering, 392–395
Rayleigh scattering-elastic scattering, 388–391
light polarization, 311–312
Maxwell's equations
Ampère's law, 314–315
birefringence-double refraction, 326–328
characteristic vacuum and medium impedance, 323–324
curl equations, 318
Dirac's quantum field theory, 313
dispersion relation, 324–325
electromagnetic waves (*see* Electromagnetic (EM) waves)
electrostatics and magnetostatics, 318–319
Faraday's law, 316–317
Gauss and Stokes theorems, 313–314
Gauss' law for electric fields, 315–316
Gauss' law for magnetic fields, 316
mathematical formulation, 317
permittivity, 317
plane wave propagation, 318–319
plane wave solutions (*see* plane wave solutions)
plasmonics (*see* Plasmonics)
ray optics, 312
refractive index, 313
right-and left-handed materials, 325–326
time varying equations, 317–318
vector fields, 313
wave equations (*see* Wave equations, medium)
nature of light
cellular phones, 301
EM spectrum, 300–301
EM vibration, 300
EM wave formation, 301–302
far field and near field, 301
field lines, 301
gravitational lensing, 302
Maxwell's description, 300
plane waves, 301
visible light, 301
optical properties of materials
classification, 349
dielectrics (*see* Lorentz model for dielectrics)
Drude–Lorentz model (*see* Drude–Lorentz model)
generalized dispersion relation, 376–378
Lorentz model for semiconductors, 378–386
metals (*see* optical properties of metals)
microscopic origin, frequency dependency, 349–350
quantum dots *vs.* metal nanoparticles, 386
photons *vs.* electrons
applications, 413–414
Bloch invention, 405
Coulomb potential, 404
dispersion relationships, 404
lattice structure, 404–405
Maxwell's equations, 402–403
photonic bandgap, 405–407
photonic crystal optic fibers, 411–413
photonic crystals defect, 409–411
2D and 3D photonic crystals manufacturing, 407–409
prism effect, 411–412
wavelength, 402, 404
μ-ε quadrant and metamaterials
dispersion equation, 416
EM properties, 414–415
ideal lens, 420–423
left-handed materials, 416–420
plane representation, 416–417
right-hand rule, 416
reflectance and total internal reflectance, 310–311
refraction
refractive index, 309–310
Snellius law, 308–309
Planck constant, 109–110
Planck's hypothesis, 98
Plane wave equations, 70–71
Plane wave solutions
E⊥B⊥k, 322–323
E⊥k and B⊥k, 322
media with charges/current, 323
propagation, 321

Plasmon bands. *See* Surface plasmon bands
Plasmonic applications 3
  applications, 401–402
  diffraction limit break
    aperture-based proximal probes/tips, 399–400
    electron microscope, 397
    high-resolution imaging, 398
    lensless optics, 398
    light detection, 398
    local field enhancement technique, 400–401
    luminescent quantum dot imaging, 400
    near-field and far-field imaging, 397–398
    NSOM tip, 398–399
    perturbation, 400
    scattering based NSOM, 399
    SNOM, 398
    surface plasmon radiation, 400
  pros and cons, 401
Plasmonics
  applications
    AC current, 344
    information transmission, 343
    lithography, 345
    local electric fields, 345
    optical fibers, 343–344
    photonic device, 344–345
  bulk plasmons, 338–339
  EELS, 338
  localized plasmons, 343
  plasmon, 338
  polaritons types, 338
  subwavelength resolution, 337
  surface plasmons
    decay lengths, 343
    dielectric constant, 340, 342
    dispersion, 341–342
    eigenmodes, 339–340
    excitation methods, 342
    localized surface mode, 340
    metal/dielectric interface, 339
    momentum, 341
    optical frequencies, 341
    permittivity, 341
    polarization charges, 339–340
    p-polarization, 339
    quantum mechanics, 339
    transverse electric and magnetic field, 339–340
    wave vector, 340–341
Point spread function (PSF), 306–307
Poisson equation, 316
Polarizing/Brewster polarizing angle, 312
Polaroid, 312
Potentiometric sensors
  EGFET, 588
  ISFET
    CHEMFETS, 585
    Debye shielding, 586–587
    gate dielectrics, 586
    IC compatibility and drift, 587
    MOSFET, 585–586
    need for reference electrode, 586
    pH sensitivity, 585
  potentiometric immunosensors
    direct binding/label-less assays, 591
    DNA probes, 588
    ELISA, 589–590
    GOx enzyme, 589
    heterogeneous assay, 590
    immunoassay, 588–589
    micromachined potentiometric immunosensors, 591–593
    noncompetitive/sandwich assay, 590
    sensitivity, 590–591
    SWNTs, 590
Pseudoplastics, 437

Q

Quantum confinement
  bulk materials
    Brillouin zones, 165–169
    density of States, 169–172
    electronic conductivity, 172–175
    Fermi surface, 164–165
  DOS function, 164
  quantum dots (*see* Quantum dots)
  quantum wells (*see* Quantum wells)
  quantum wires (*see* Quantum wires)
  reduced-dimensional material geometries, 164
Quantum dots (QDs), 384
  BeadArray™, 612
  biomolecule conjugation schemes, 611
  CdS and ZnSe dots, 609
  CdSe QDs, 609–610
  colloid stabilizer, 610
  definition, 608
  density of states function, 189–190
  DNA microarray assays, 611
  electronic conductivity, 190–191
  Fermi surfaces and Brillouin zone, 188–189
  3-(mercaptopropyl) trimethoxysilane, 610
  vs. organic dyes, 610
  quantization, 188
  RSD, 609
  spectral barcodes, 612
Quantum electron dynamics, 209
Quantum mechanics and lasers
  photonic crystal, 431
  quantum cascade lasers
    degree of freedom, 431
    electron transitions, 430
    Esaki tunnel diode, 427
    GaAs, 428
    minibands formation, 430–431
    MOCVD, 428–429
    quantum well, 429
    resonant tunneling diode, 427–428
    superlattice, 428–429
  quantum confinement, 426–427
Quantum numbers
  fourth quantum number, electron spin, 106
  magnetic number, 105–106
  principal quantum number, 104–105
  properties, 107
  second quantum number, 105
Quantum wells
  density of states
    bound-state energies, 179
    1D confinement, 177–178
    2D electron gas, 178
    frequency dependence, 179
    optical transitions, 178–179
  electronic conductivity, 179–180
  Fermi surfaces and Brillouin zone, 177
  GaAs and $Al_xGa_{1-x}As$ layer, quantization
    blue shift, 177
    electron energy levels, 176–177
    electronic conductivity, 179–180
  Schrödinger's equation
    charge carriers, 128
    HEMT, 128
    one dimension, 127
    semiconductor heterostructures, 128
    Sommerfeld's mathematical model, 125
Quantum wires
  density of states function, 182–183
  electronic conductivity
    Bloch's theorem, 183
    conductance quantization, 186
    density of states, 183–184
    one-dimensional wire, 184–185
    quantized resistance, 185
    Van Hove singularities, 184
  Fermi surface, 182
  quantization
    2D quantum gas, 180–181
    energy dispersion function, 181
    $GaAs/Al_xGa_{1-x}As$ heterojunction, 181
    square cross-section, 182

R

Radiation pressure, 329
Radio detecting and ranging (RADAR), 7
Raman (inelastic) scattering
  advantages, Raman spectroscopy, 393
  chemical adsorption, 395
  decay curves, 394–395
  disadvantages, 393
  field and signal enhancement, 393
  gain, 394
  molecules polarizability, 392

physicist, Chandrasekhara Venkata Raman, 392
Raman shift, 392
Raman spectrum of CCl_4, 392
Raman *vs*. IR spectroscopy, 392–393
SERS, 393
Stokes lines, 392
total enhancement factor, 393–394
wavelength discovery, 392
Randles–Sevcik equation, 580
Random access memory (RAM), 10–11
Rayleigh's limit
Airy pattern, 306
circular aperture diffraction, 306
Fraunhofer diffraction patterns, 307
lens magnification, 307–308
lithography resolution, 308
numerical aperture, 308
PSF, 306–307
Rayleigh criterion, 307–308
small circular aperture, 307–308
smallest object radius, 307
Ray optics. *See* Geometric optics
Reciprocal space
definition, 57–58
diffraction pattern, 56
Ewald construction, 59–60
explicit algorithm, 58–59
graphical presentation, 57
Relative permittivity, 350
Relative standard deviation (RSD), 609
Reynolds number
hydraulic diameter, 454
hydrodynamic layer thickness, 454–455
intermolecular viscous forces, 454
laminar/turbulent flow, 453–454
mobile machines, 454
Right-handed materials (RHMs), 325–326
Rutherford experiment, 101–102

S

Scanning near-field optical microscopy (SNOM), 398
Scanning tunneling microscope (STM), 21
Schrödinger's equation
Born and von Karman's periodic boundary condition, 124–125
cat-like state of matter, 209–210
finite-depth potential wells, 134–136
finite height barriers, 136–137
free electrons, infinitely large 3D piece of metal, 124
free particles, wave function, 123–124
Hamiltonian, 116
harmonic potential wells, 140–141
infinitely deep, finite-sized 1D potential wells
boundary conditions, 125–126
dispersion function, 126–127
eigenfunctions, 126
energy levels, 126
probability distribution, 127
quantum wells (*see* Quantum wells)
Sommerfeld's mathematical model, 125
wave number, 126
infinitely deep, finite-sized 2D potential wells
vs. 1D potential wells, 129–130
quantum wires, 130–131
wave functions and energies, 129
infinitely deep, finite-sized 3D potential wells
energy levels, 131
k-states, 131–132
quantum dots, 132–133
wave vectors, 131
interfaces, 139–140
plausibility, 119–121
TISE, stationary states, 122
tunneling, 137–139
two-dimensional representation, 116
wave function interpretation, 121–122
zero point energy, 133–134
Sedimentation potential, 538
Severinghaus sensor, 563
SHE. *See* Standard hydrogen electrode
Silicon in integrated circuits
Apple I and II, 11
Babbage, analytical engine, 7
business, 13
Colossus coding and deciphering machine, 7
Crookes experiment, 5
crystal set radios, 6
De Forest, Lee, 6
Edison effect, 5–6
ENIAC, 7–8
FET, 10
Fleming valve/Fleming diode, 5–6
integration scale and circuit density, 12
junction transistors, 9
Kilby, Jack, 9–10
microcomputer kit, 11
Moore, Gordon (Intel cofounder), 12
Noyce, Robert, 9–10
personal computer (PC), 11
point-contact germanium bipolar transistor, 8
RADAR, 7
RAM, 10–11
Regency TR-1, 9
Shockley, Bardeen, and Brattain invention, 8
transferred resistance, 8
types of components, 9–10
vacuum tubes, 6, 12
von Neumann, John, 8
Silicon single crystal (SSC)
crystal growth
Czochralski (CZ) crystal pulling method, 225–226
float zone crystal growth, 226
monocrystalline silicon wafer, 227
silicon boules, 226–227
wafer flats, 226–227
crystallography
diamond-hexagonal silicon, 216
diamond-type lattice, 216
energy bandgap difference, 216
FCC lattices, 216
MEMS structure, 217–218
silicon substrates, 217
silicon wafers, 217
diffusion
activation energy, 231–232
boundary condition, 230
diffusion coefficient, 232
dopant diffusivities, 232
dopant surface concentration, 230–231
drive-in diffusion, 231
Fick's first law, 229
Fick's second law, 229
gaseous doping, 229–230
gradients, 230
impurity profile, 232
vs. ion implantation, 233
lateral diffusion, 232
mechanisms, 229
oxide film, 228
two-step junction formation, 231
wafer coating, 228–229
extrinsic *vs*. intrinsic semiconductors, 227–228
ion implantation, 229, 233–234
mechanical sensors
elasticity constants, 277–278
piezoresistivity, 281–285
residual stress, 278–279
stress-strain behaviors, 275–277
MEMS, biocompatibility, 290–293
optical properties, 288–289
[100]-oriented silicon, 218–222
[101]-oriented silicon, 222–223
oxidation
Bruce Deal–Andy Grove model, 238
field oxidation, 240
fluxes, 238
gas-phase oxidation method, 237–238
gate oxide thickness, 245
high-pressure oxidation, 240
kinetics, 241–242
LOCOS process, 240–241
nonthermal methods, 242–243

oxide thickness color, 239
parabolic law, 239
semiconductors, 237
shallow trench isolation, 240
silicon and oxide thickness, 238
silicon dioxide growth, 237
silicon dioxide thickness *vs.* growth time, 238–239
thermal oxides, 239
thermal SiO_2, properties, 243–245
ultrathin oxides, 240
structure and conductivity, 222–225
thermal properties, 288
transistors (*see* Transistors)
Single-wall carbon nanotubes (SWNTs), 590
Solar cells, 251–252
Solid-state p-n junction diode, 246
Sound navigation and ranging (SONAR). *See* Radio detecting and ranging
Specific heat capacity, 89–90
SSC. *See* Silicon single crystal
Standard hydrogen electrode (SHE), 545–546
Stern potential, 536
Streaming potential, 537
Surface-enhanced Raman scattering (SERS), 393
Surface plasmon bands, 367

T

Thermal conductivity, 90–93
Thermalization process, 381–382
Time-independent Schrödinger equation (TISE), 122
Tin-doped indium oxide (ITO), 365
Transistors
bipolar junction transistor
vs. BiCMOS, 255
vs. CMOS, 255
elements, 254
high *vs.* low-resistance circuit, 255–256
planar process, 254–255
p-n diodes, 254
bipolar transistors, 254–256
FinFET approach, 273
MOSFETS (*see* Metal-oxide-semiconductor field-effect transistor)
Traveling wave dielectrophoresis (TWD), 490–491

U

Ultramicroelectrodes (UME)
advantages, 583
CNT arrays, 583
Cottrell equation for planar diffusion, 582
diffusion layer thickness, 581–582
diffusion-limited current, 582
Kelgraf electrode, 583
LIGA process, 584
microelectrodes, 582–583
nonlinear diffusion, 584
types, 581
UV/VIS absorption spectroscopy, 595–598

V

Valence band maximum (VBM), 379
Van der Waals forces
definition, 524
dipole-dipole interactions
hydrogen bonding, 526–528
Keesom equation, 524–525
dipole-induced dipole interactions, Debye, 525
dispersion forces, 524
induced dipole-induced dipole, dispersive/London forces, 525–526
solid surfaces
energy of adhesion, 530
Hamaker constants, 530
interaction potentials, 531
for molecules, 529
sphere-plane interaction, 531–532
surface energy, 530–531
A/V ratio, 532
types, 524
Velocity of charge carriers, 156–157
Vertical cavity surface-emitting lasers (VCSELs), 426
Voltammetry, 553–554
Voltammogram, 553

W

Wave equations, medium
with charges and currents, 320–321
without charge/current, 319–320
Wave velocity, 309
Working electrode, 552
Worldwide IC and electronic equipment, 32–33

X

X-ray diffraction
Bragg's law, 49–50
discovery, 48
electron density map, ATP, 49
electron density/molecular structure image, 49–50
Fourier transform, 49
Laue equations
connection with Bragg's law, 52–53
diffraction angles, 50
Henry and Bragg, 52
incident x-ray beam scattering, 51
magnitude of vector, 53
reflection, 52
sodium deoxyribose nucleate, 53
two cones, 51
two scattering atoms, 50–51
von Laue's photography, 52
pattern, 48–49
phase problem, 49
x-ray intensity and structure factor
Fraunhofer diffraction patterns, 54
Fresnel diffraction patterns, 53–54
interference, 55
lost phase regeneration, 54
phase problem, 54, 56
scattering, x-ray, 54
spatial frequency spectrum/harmonic analysis, 54
wave amplitude, 54–55
wave vector, 56
$y = \sin^2 Mx/\sin^2 x$ plot, 55–56

Z

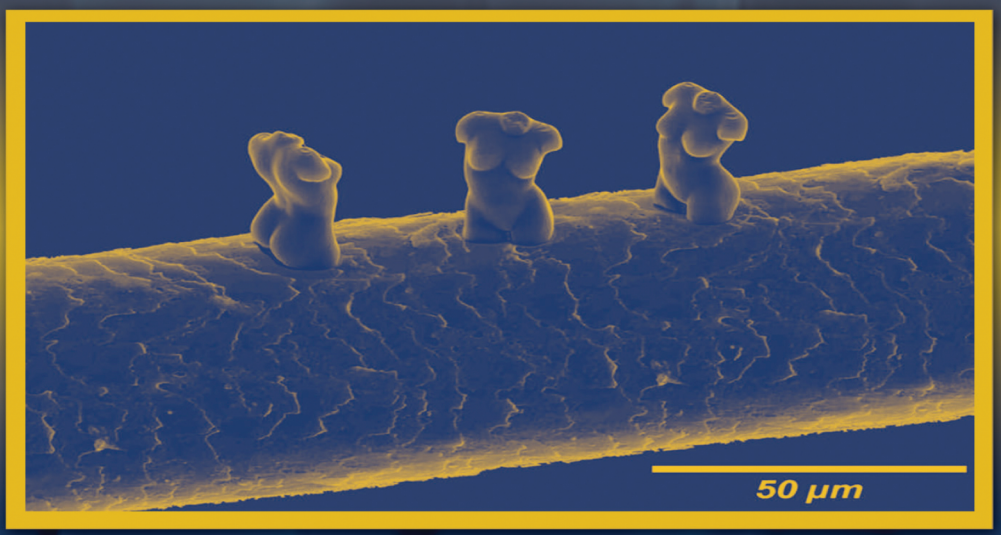
Zener diode, 252
Zero point energy (ZPE), 541–542

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME II

Manufacturing Techniques for Microfabrication and Nanotechnology



Marc J. Madou

CRC Press
Taylor & Francis Group

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME II

Manufacturing Techniques for Microfabrication and Nanotechnology

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME II

Manufacturing Techniques for Microfabrication and Nanotechnology

Marc J. Madou

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2012 by Taylor and Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-5519-1 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

I dedicate this third edition of Fundamentals of Microfabrication to my family in the US and in Belgium and to all MEMS and NEMS colleagues in labs in the US, Canada, India, Korea, Mexico, Malaysia, Switzerland, Sweden and Denmark that I have the pleasure to work with. The opportunity to carry out international research in MEMS and NEMS and writing a textbook about it has been rewarding in terms of research productivity but perhaps even more in cultural enrichment. Scientists have always been at the frontier of globalization because science is the biggest gift one country can give to another and perhaps the best road to a more peaceful world.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Roadmap ix
Author xi
Acknowledgments xiii

PART I
Lithography

Introduction to Part I 2

1 Photolithography 3

2 Next-Generation Lithographies and Lithography Research 89

PART II
Pattern Transfer with Subtractive Techniques

Introduction to Part II 148

3 Dry Etching 155

4 Wet Chemical Etching and Wet Bulk Micromachining—Pools as Tools 215

5 Thermal Energy-Based Removing 319

6 MechanicalEnergy-Based Removing 351

PART III
Pattern Transfer with Additive Techniques

Introduction to Part III 386

7 Physical and Chemical Vapor Deposition—Thin Film Properties and Surface Micromachining 391

8 Chemical, Photochemical, and Electrochemical Forming Techniques 509

9 Thermal Energy-Based Forming Techniques—Thermoforming 567

10 Micromolding Techniques—LIGA 591

Index **643**

Roadmap

Manufacturing Techniques for Microfabrication and Nanotechnology consists of three parts and ten chapters.

In Part I, we review different forms of lithography, detailing those that differ most from the miniaturization processes used to fashion integrated circuits (ICs). Chapter 1 starts with a short historical note about the origins of lithography, followed by a description of photolithography, including developments that have allowed the printing of the ever-shrinking features of modern ICs. After reviewing the limits of photolithography, we detail alternative next-generation lithographies (NGEs) in Chapter 2, including extreme ultraviolet (EUV), x-ray, and charged-particle (electron and ion) lithographies, followed by lithography techniques in the early research and development (R&D) stage. Lithography approaches in the R&D stage include very thin resist layers and block copolymers, zone plate array lithography (ZPAL), quantum lithography (two-photon lithography), and proximal probe-based techniques such as atomic force microscopy (AFM), scanning tunneling microscopy (STM), dip-pen lithography (DPL), near-field scanning optical microscopy (NSOM), and apertureless near-field scanning optical microscopy (ANSOM). Also considered are holographic lithography, plasmonic lithography, and lithography with superlenses ("Pendry's dream"). As an example of a lithography method that is capable of patterning resist on nonplanar substrates (e.g., patterning a resist on the surface of a cylinder) we detail soft lithography.

In Part II we deal with material removal processes, and additive technologies are the focus of Part III. We review gas and vapor phase dry etching processes, including chemical, physical (purely mechanical), and chemical-physical (mostly chemical with some mechanical assist) in Chapter 3. Chapter 4 consists of two parts: one covers wet *chemical etching* and the other *wet bulk micromachining*, the latter being chemical etching as well but optimized for the manufacture of ICs, nano-, and micromachines. Chapter 5 discusses thermal removing techniques where thermal energy, provided by a heat source, melts and/or vaporizes material from the work piece. Examples of thermal techniques covered in this chapter are electric discharge machining (EDM) (including μ-EDM), laser- and electron-beam machining (LBM and EBM), and plasma arc cutting. In Chapter 6, we cover traditional and nontraditional mechanical removing techniques. Traditional mechanical material removing technologies covered are mechanical precision machining and abrasive wheel machining. In nontraditional mechanical machining, we deal with ultrasonic drilling, electrolytic in-process dressing (ELID) grinding, water jet, abrasive water jet, abrasive jet machining, and focused ion-beam milling (FIB). Also in this chapter, we highlight recent developments toward miniaturized mechanical manufacturing equipment, the so-called desktop factories.

In Part III we cover additive (forming) processes, where materials are added, usually in a selective manner, to a work piece or a device under construction. Physical vapor deposition processes (PVD) and chemical vapor deposition (CVD) are topics of Chapter 7. In the same chapter we survey thin film properties and detail surface micromachining, an important application of thin film deposition techniques. In surface micromachining, features are built up, layer by layer, on the surface of a substrate (e.g., a single-crystal silicon wafer). Dry etching defines the surface features in the x,y-plane, and wet etching releases them from the plane by undercutting. In surface micromachining, shapes in the x,y-plane are unrestricted by the crystallography of the substrate. In Chapter 8 we discuss additive chemical, photochemical, and electrochemical processes. The chemical forming methods covered in Chapter 8 are geared toward BioMEMS applications. In photochemical forming, photoenergy solidifies a material into a 3D shape, as illustrated in rapid prototyping and microphotoforming (stereolithography). In the case of electrochemical forming, the energy at the work piece is electrochemical. Chapter 9 covers thermal forming techniques where thermal energy provided by a

heat source "transforms" a material as in crystallization, sintering, alloying, annealing, decomposition, and pyrolizing. Replication methods based on plastic molding and LIGA—a sophisticated type of replication technique combining x-ray lithography with electroplating and plastic molding—are discussed in Chapter 10. Plastic molding techniques include RIM, thermoplastic injection molding (IM), and compression molding or hot embossing.

Note to the Reader: *Manufacturing Techniques for Microfabrication and Nanotechnology* was originally composed as part of a larger book that has since been broken up into three separate volumes. *Manufacturing Techniques for Microfabrication and Nanotechnology* represents the second volume in this set. The other two volumes include *Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology* and *From MEMS to Bio-NEMS: Manufacturing Techniques and Applications*. Cross-references to these books appear throughout the text and will be referred to as Volume I and Volume III, respectively. The interested reader is encouraged to consult these volumes as necessary.

Author

Dr. Madou is the Chancellor's Professor in Mechanical and Aerospace Engineering (MEA) at the University of California, Irvine. He is also associated with UC Irvine's Department of Biomedical Engineering and the Department of Chemical Engineering and Materials Science. He is a Distinguished Honorary Professor at the Indian Institute of Technology Kanpur, India, and a World Class University Scholar (WCU) at UNIST in South Korea.

Dr. Madou was Vice President of Advanced Technology at Nanogen in San Diego, California. He specializes in the application of miniaturization technology to chemical and biological problems (bio-MEMS). He is the author of several books in this burgeoning field he helped pioneer both in academia and in industry. He founded several micromachining companies.

Many of his students became well known in their own right in academia and through successful MEMS start-ups. Dr. Madou was the founder of the SRI International's Microsensor Department, founder and president of Teknekron Sensor Development Corporation (TSDC), Visiting Miller Professor at UC Berkeley, and Endowed Chair at the Ohio State University (Professor in Chemistry and Materials Science and Engineering).

Some of Dr. Madou's recent research work involves artificial muscle for responsive drug delivery, carbon-MEMS (C-MEMS), a CD-based fluidic platform, solid-state pH electrodes, and integrating fluidics with DNA arrays, as well as label-free assays for the molecular diagnostics platform of the future.

To find out more about those recent research projects, visit http://www.biomems.net.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Acknowledgments

I thank all of the readers of the first and second editions of Fundamentals of Microfabrication as they made it worthwhile for me to finish this completely revised and very much expanded third edition. As in previous editions I had plenty of eager reviewers in my students and colleagues from all around the world. Students were especially helpful with the question and answer books that come with the three volumes that make up this third edition. I have acknowledged reviewers at the end of each chapter and students that worked on questions and answers are listed in the questions sections. The idea of treating MEMS and NEMS processes as some of a myriad of advanced manufacturing approaches came about while working on a WTEC report on International Assessment Of Research And Development In Micromanufacturing (http://www.wtec.org/micromfg/report/Micro-report.pdf). For that report we travelled around the US and abroad to visit the leading manufacturers of advanced technology products and quickly learned that innovation and advanced manufacturing are very much interlinked because new product demands stimulate the invention of new materials and processes. The loss of manufacturing in a country goes well beyond the loss of only one class of products. If a technical community is dissociated from manufacturing experience , such as making larger flat-panel displays or the latest mobile phones, such communities cannot invent and eventually can no longer teach engineering effectively. An equally sobering realization is that a country might still invent new technologies paid for by government grants, say in nanofabrication, but not be able to manufacture the products that incorporate them. It is naïve to believe that one can still design new products when disconnected from advanced manufacturing: for a good design one needs to know the latest manufacturing processes and newest materials. It is my sincerest hope that this third edition motivates some of the brightest students to start designing and making things again rather than joining financial institutions that produce nothing for society at large but rather break things.

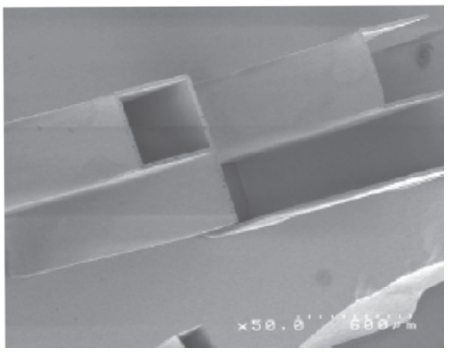
Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Part I

Lithography

(a)

(b)



(a) Engraving of Cardinal d'Amboise. The earliest example of photolithography followed by wet etching. (Photograph from the Science Museum. Courtesy of the Royal Photographic Society.) (b) Ultrahigh-aspect-ratio SU-8 photoresist structures. Microfeatures with designed thickness of 40 μm and height of 2000 μm. (Courtesy of Dr. Wanjun Wan, Louisiana State University.)

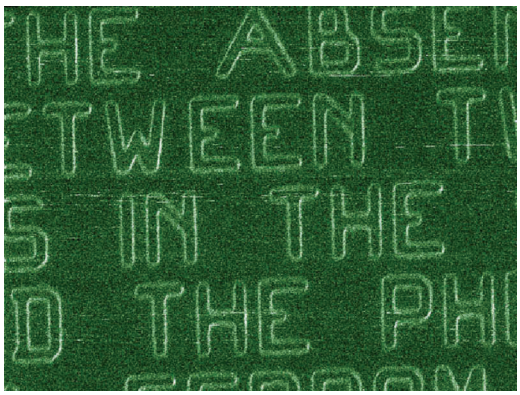
> There are those who still believe that America can move from the strongest industrial core economy in the world to essentially a service economy without losing its greatness, its dynamism, and industrial health. I am absolutely certain we cannot.
>
> **David Halberstam**
> *The Next Century* (1991)

Introduction to Part I

Chapter 1 Photolithography
Chapter 2 Next-Generation Lithographies (NGLs) and Lithography Research

Introduction to Part I

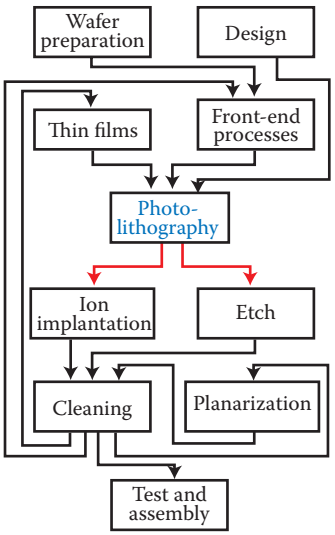
For many miniaturization tasks, lithography, the technique used to transfer copies of a master pattern onto the surface of a solid material such as a silicon wafer, is of central importance. In Part I of this book, we review different forms of lithography, detailing those that differ most from the miniaturization processes used to fashion integrated circuits (ICs). Chapter 1 starts with a short historical note about the origins of lithography, followed by a description of photolithography, including developments that have allowed the printing of the ever-shrinking features of modern ICs. After reviewing the limits of photolithography, we detail alternative next-generation lithographies (NGLs) in Chapter 2, including extreme ultraviolet (EUV), x-ray, and charged-particle (electron and ion) lithographies, followed by techniques in the early research and development stage.



Text by William Gilbert (from De Magnete), written using a 100-keV electron microscope in a common resist (PMMA). The height of each letter is 250 nm. The pattern was then imaged using the same electron microscope. (From http://www.lancs.ac.uk/staff/koltsov/nanoimage.html.)

1

Photolithography



The image of an IC, micro-, or nanomachine pattern is transferred to a substrate in a process known as photolithography. Substrates are first coated with a thin emulsion of photoactive organic material called a *photoresist*. A pattern is then aligned and projected onto the substrate using a light source (e.g., deep UV). Subsequently, the resulting image is developed and inspected.

Outline

Introduction

Historical Note: Lithography's Origins

Photolithography Overview

Critical Dimension, Overall Resolution, Line Width Metrology

Lithographic Sensitivity and Intrinsic Resist Sensitivity (Photochemical Quantum Efficiency)

Resist Profiles

Contrast and Experimental Determination of Lithographic Sensitivity

Resolution in Photolithography

Photolithography Resolution Enhancement Technology

Appendix 1A: Spin Coating Process Troubleshooting

Questions

References

Introduction

We start this chapter with a short historical note about the origins of photolithography and follow with details on all the steps involved, including resolution enhancement techniques (RET), that have allowed the printing of the ever-shrinking features of modern integrated circuits (ICs). In this photolithography treatise we expand especially on those aspects that are of most importance to microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS) and that differ the most from typical lithography processes for IC fabrication.

Historical Note: Lithography's Origins

Within the space of 40 years two developments in photography laid the foundations for the photoresists we use today. In 1782, the Swiss pastor Jean Senebier (1742–1808) of Geneva investigated the property of

certain resins that become insoluble in turpentine after exposure to sunlight. Inspired by Senebier, Nicéphore Niépce, after experimenting with various resins in sunlight, managed to copy an etched print on oiled paper by placing it over a glass plate coated with bitumen (asphalt) dissolved in lavender oil (France, 1822). In so doing, he resurrected an ancient Egyptian embalming technique that involved the use of what is now known as Syrian asphalt. After two or three hours of sunlight, the unshaded areas in the bitumen became hard compared with the shaded areas, which remained more soluble and could be washed away with a mixture of turpentine and lavender oil. Niépce's concoction, we will learn below, corresponds to a negative-type photoresist. Five years later, in 1827 (talk about fast turnaround time!), using strong acid, the Parisian engraver Lemaître made an etched copy of an engraving of Cardinal d'Amboise from a plate developed by Niépce (see (a) on title page of Part I). The latter copy represents the earliest example of pattern transfer by photolithography and chemical milling [i.e., photochemical milling (PCM); see Chapter 4]. The accuracy of the technique was 0.5–1 mm.[1]

The word *lithography* [Greek for the words *stone* (*lithos*) and to *write* (*gráphein*)] refers to the process invented in 1796 by Aloys Senefelder. Senefelder found that stone (he used Bavarian limestone), when properly inked and treated with chemicals, could transfer a carved image onto paper. As a result of the chemical treatment of the stone, image and nonimage areas became oil receptive (water repellent) and oil repellent (water receptive), respectively, attracting ink onto the image area and attracting water on nonimage areas.

Photomasking, followed by chemical processing, led to the photolithography now used in fabricating integrated circuits (ICs) and in miniaturization science. Not until World War II, more than 100 years after Niépce and Lemaître, did the first applications of the printed circuit board (PCB) [also printed wiring board (PWB)], invented in 1943 by the Austrian Paul Eisler, come about. Interconnections were made by soldering separate electronic components to a pattern of "wires" produced by photoetching a layer of copper foil that was laminated to a plastic board. Jack Kilby of Texas Instruments and Robert Noyce of Fairchild Semiconductor developed the first IC in the late 1950s. By 1961, Fairchild Semiconductor had introduced the first commercial IC in which photoetching processes produced large numbers of transistors on a thin slice of Si. Now, ICs are ubiquitous, and we find them in microprocessors, audio and video equipment, dishwashers, garage openers, security systems, automobiles, and so on. With the first ICs, patterns had a resolution not better than 5 μm. Today photolithography, x-ray lithography, and charged particle lithography all achieve submicrometer-printing accuracy. Intel introduced the Prescott Pentium IV chip, an IC chip incorporating 90-nm-sized features using UV photolithography, in the fourth quarter of 2003. Chips based on a 65-nm technology became available in 2006, and the 32-nm node was achieved in 2009.

For details on the projected feature size of future ICs, see the International Technology Roadmap for Semiconductors (ITRS) in Appendix 1A of Chapter 1 in Volume I or on the web at http://public.itrs.net. In the ITRS, technology modes have been defined, i.e., the feature sizes that have to be in volume manufacturing at a fixed date (year of production).

Photolithography Overview

Introduction

The most widely used form of lithography is photolithography. In the IC industry, pattern transfer from masks onto thin films is accomplished almost exclusively via photolithography. The combination of accurate registration and exposing a series of successive patterns leads to complex multilayered ICs. This essentially two-dimensional (2D) process has a limited tolerance for nonplanar topography, creating a major constraint for building non-IC miniaturized systems, which often exhibit extreme topographies. Photolithography has matured rapidly, and its ability to resolve ever-smaller features is constantly improving. For the IC industry, this continued improvement in photolithography resolution has impeded the adaptation of alternative, higher-resolution lithography techniques, such as x-ray lithography. Research during the past 10 years in high-aspect-ratio resist features to satisfy needs of both IC and

non-IC miniaturization is also finally improving dramatically photolithography's capacity to cover wide ranges of topography. Performance of a photolithographic process is determined by its resolution, the minimum feature size that can be transferred with high fidelity; the registration, how accurately patterns on successive masks can be aligned; and throughput, the number of wafers that can be transferred per hour (a measure of the efficiency of the lithographic process).

Photolithography and pattern transfer involve a set of process steps summarized in Figure 1.1. As an example, we use an oxidized Si wafer and a negative photoresist system. For simplicity, not all the steps are detailed in this figure, as they will be covered in the subsequent text. A short preview follows.

An oxidized wafer (Figure 1.1a) is coated with a 1-µm-thick negative photoresist layer (Figure 1.1b). After exposure (Figure 1.1c), the wafer is rinsed in a developing solution or sprayed with a spray developer, which removes the unexposed areas of photoresist and leaves a pattern of bare and photoresist-coated oxide on the wafer surface (Figure 1.1d). The photoresist pattern is the negative image of the pattern on the photomask. In a typical next step after development, the wafer is placed in a solution of HF or HF + NH_4F, meant to attack the oxide but not the photoresist or the underlying silicon (Figure 1.1e). The photoresist protects the oxide areas it covers. Once the exposed oxide has been etched away, the remaining photoresist can be stripped off with a strong acid such as H_2SO_4 or an acid-oxidant combination such as H_2SO_4-Cr_2O_3, attacking the photoresist but not the oxide or the silicon (Figure 1.1f). Other liquid strippers include organic solvent strippers and alkaline strippers (with or without oxidants). The oxidized Si wafer with the etched windows in the oxide (Figure 1.1f) is ready now for further processing, which might entail a wet anisotropic etch of the Si in the etched windows with SiO_2 as the etch mask.

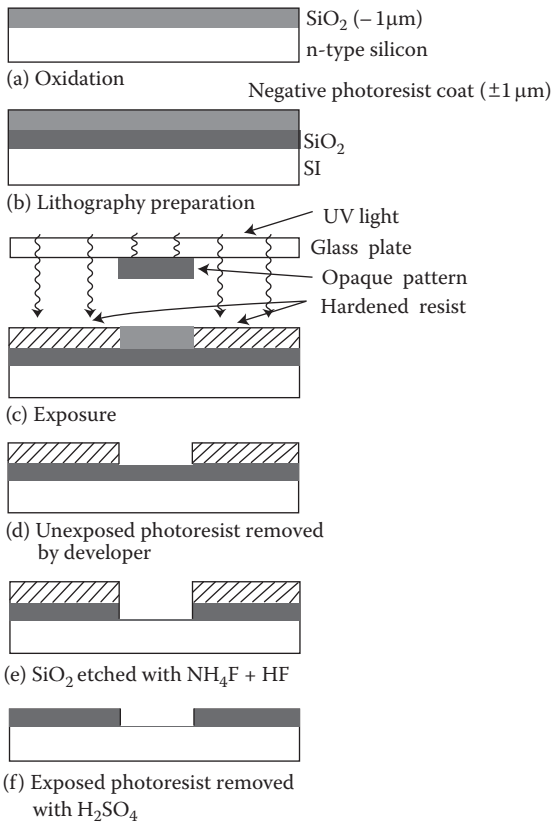


FIGURE 1.1 Basic photolithography and pattern transfer. Example uses an oxidized silicon wafer and a negative photoresist system. Process steps include exposure, development, oxide etching, and resist stripping. Steps (a) through (f) are explained in the text.

Wafer Cleaning and Contaminants: The Clean Room

An important step, even before lithography proper, is wafer cleaning. Contaminants include solvent stains (e.g., methyl alcohol, acetone, trichloroethylene, isopropyl alcohol, and xylene), dust from operators and equipment, and smoke particles. Stains or films may lead to adverse effects during oxidation and evaporation processes. Particulates, chunks of granular matter, may cause undesirable masking effects and scratches on the photomask during contact printing (see Figure 1.6). Solvent stains and other contaminants on a Si wafer can be easily observed in dark field microscopy (see Volume III, Chapter 6 on metrology and MEMS, NEMS modeling). All lithography processes take place inside a semiconductor clean room, which is a specially constructed enclosed area environmentally controlled with respect to airborne particulates, temperature (±0.1°F), air pressure, humidity (from 0.5–5% relative humidity), vibration, and lighting. In Table 1.1, some common sources of clean room contaminants are listed, and in Figure 1.2, the clean room classification system is elucidated. In a Class 1 clean room, the particle count does not exceed 1 particle

TABLE 1.1 Some Common Clean Room Contaminant Sources

• Location: a clean room near a refinery, smoke stack, sewage plant, cement plant spells big trouble
• Construction: the floor is an important source of contamination. Also, items such as light fixtures must be sealed, and room construction tolerances must be held very tight
• Wafer handling: transfer box
• Process equipment: never use fiberglass duct liner; always use 100% polyester filters, eliminate all nonessential equipment
• Chemicals: residual photoresist or organic coatings, metal corrosion
• Attire: only proper attire and dressing in the anteroom
• Electrostatic charge: clean room must have a conductive floor
• Furniture: only clean room furniture
• Stationary: use ballpoint pen instead of lead pencil, only approved clean room paper
• Operator: no eating, drinking, smoking, chewing gum, or makeup of any kind

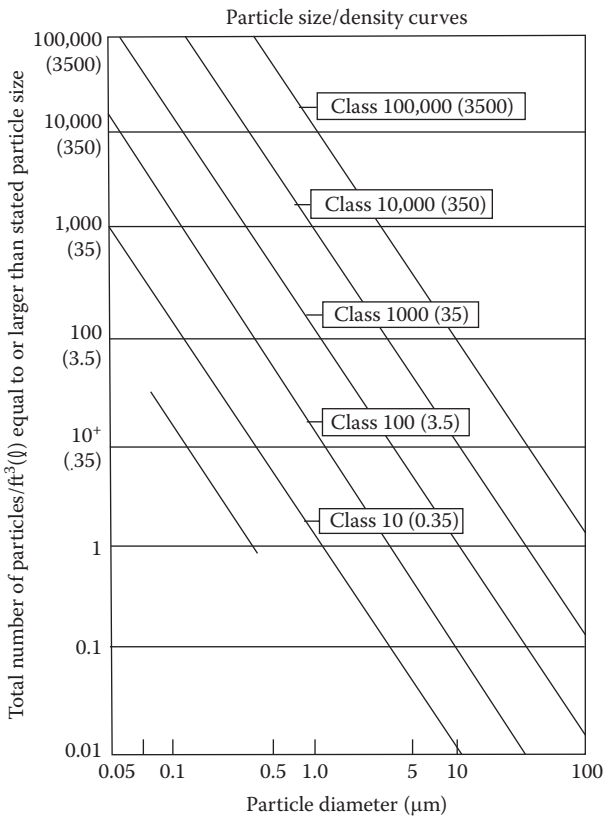


FIGURE 1.2 U.S. Federal Standard 209b for clean room classification. The bottom solid line shows the definition of Class 1. (From Cunningham, J.A. 1992. The remarkable trend in defect densities and chip yield. *Semicond Int* 15:86–90. With permission.[2])

per cubic foot with particles 0.5 μm and larger, and in a Class 100 clean room, the particle count does not exceed 100 particles per cubic foot with particles 0.5 μm and larger. The allowable contamination particle size in IC manufacture has been decreasing hand in hand with the ever-decreasing minimum feature size. With a 64-kB dynamic memory chip [dynamic random access memory (DRAM)], for example, one can tolerate 0.25-μm particles, but for a 4-MB DRAM, one can only tolerate 0.05-μm particles. The smallest feature sizes in these two cases are 2.5 and 0.5 μm, respectively. As a reference point, a human hair has a diameter of 75–100 μm (depending on age and race); tobacco smoke contains particles ranging from 0.01–1 μm; and a red blood cell ranges from 4–9 μm.

Example 1.1

Problem: If a 125-mm diameter wafer is exposed for 1 minute to an air stream under a laminar-flow condition at 30 m/min, how many dust particles will land on the wafer in a Class 10 clean room?

Solution: In a Class 10 clean room, there are 350 particles (0.5 μm or larger) per cubic meter. The air volume that passes over the wafer in 1 min is (30 m/min) × π(0.125 m / 2)^2 × 1 min = 0.368 m^3. The number of dust particles (0.5 μm or larger) contained in that air volume is 350 × 0.368 = 128 particles. Therefore, if there are 200 IC chips on the wafer, the particle count amounts to one particle on each of 64% of the chips. Fortunately, only a fraction of the particles that land on the surface of the wafer adhere to the wafer surface, and of those only a fraction are at a circuit location critical enough to cause a failure. However, the calculation demonstrates the importance of the clean room.

Many different dry and wet methods for wafer cleaning currently in use are listed below. RCA1 and RCA2, developed by W. Kern in 1965 while working for the Radio Corporation of America (RCA), which are well known, use mixtures of hydrogen peroxide and various acids or base followed by deionized (DI) water rinses. Others include vapor cleaning; thermal treatment, for example, baking at 1000°C in a

Table 1.2 Wet versus Dry Cleaning Attributes

Attribute	Wet	Dry
Particle removal	+	–
Metal removal	+	–
Heavy organics, i.e., photoresist	+	–
Light organics, i.e., outgassed hydrocarbon residues	+	+
Throughput	+	–
Process repeatability	+	+
Water usage	–	+
Process chemical cleanliness	–	+
Environmental impact, purchase, and disposal cost	–	+
Single wafer use applicability	–	+

Dry cleaning usually requires wet followup. UV ozone can effectively remove light organic contamination.[3]
Source: Iscoff, R. 1991. Wafer cleaning: can dry systems compete? *Semicond Int* 14:48–54. With permission.[3]

vacuum or in oxygen; and plasma or glow discharge techniques, for example, in Freons with or without oxygen. Mechanical methods include ultrasonic agitation, polishing with abrasive compounds, and supercritical cleaning. Ultrasonic cleaning, which is excellent for removing particulate matter from the substrate, is unfortunately prone to contamination and mechanical failure of deposited films. Attributes of wet versus dry cleaning techniques are compared in Table 1.2. Except for environmental concerns, wet etching still outranks other cleaning procedures.

The prevalent RCA1 and RCA2 wet cleaning procedures are as follows:

- RCA1: Add 1 part of NH_3 (25% aqueous solution) to 5 parts of DI water; heat to boiling, and add 1 part of H_2O_2. Immerse the wafer for 10 minutes. This procedure removes organic dirt (resist).
- RCA2: Add 1 part of HCl to 6 parts of DI water; heat to boiling, and add 1 part of H_2O_2. Immerse the wafer for 10 minutes. This procedure removes metal ions.

The second RCA cleaning process is required to keep oxidation and diffusion furnaces free of metal contamination. Both cleaning processes leave a thin oxide on the wafers. Before a further etch of the underlying silicon is attempted, oxide must be stripped off by dipping the wafer in a 1% aqueous HF solution for a very short time. Water spreads on an oxide surface (hydrophilic) and beads up on a bare Si surface (hydrophobic). This behavior can be used to establish whether any oxide remains.

In most IC labs, processing a wafer previously exposed to KOH is not allowed as it is feared that the potassium will spoil the IC fabrication process. In more lenient environments, carefully cleaned wafers using RCA1 and RCA2 are allowed.

Supercritical cleaning (Figure 1.3) with CO_2 is especially suited for microstructure cleaning.[4] These fluids possess liquid-like solvative properties and gas-like diffusion and viscosity that enable rapid penetration into crevices with complete removal of organic and inorganic contaminants contained therein. During wet cleaning of surface micromachined structures, thin, mechanical microstructures tend to stick to one another or to the substrate through surface tension (stiction) (Chapter 7). Consequently, dry vapor phase and supercritical cleaning with low or no surface tension are preferred.

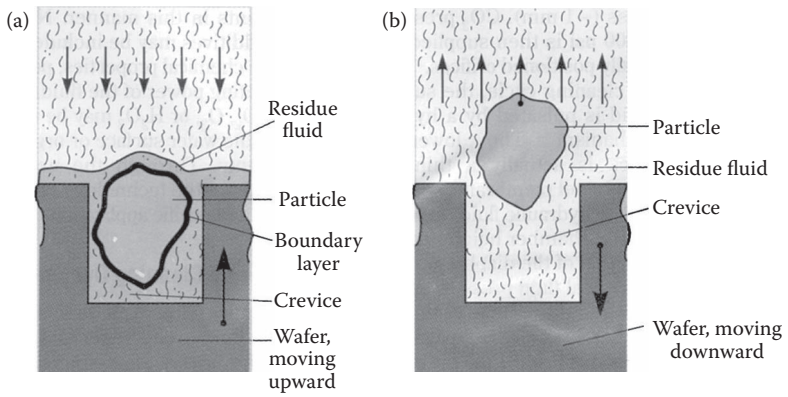


FIGURE 1.3 (a) Wet cleaning. (b) Supercritical cleaning.

Vapor phase cleaning also uses significantly fewer chemicals than wet immersion cleaning.[5]

We will learn more about the emerging importance of supercritical fluids as developers for completely dry resist processes further below.

Wafer cleaning has become a scientific discipline in its own right with journals such as *Microcontamination* and *The Magazine for Ultraclean Manufacturing Technology*, books such as *Handbook of Contamination Control in Microelectronics: Principles, Applications and Technology*[6] and the *Handbook of Semiconductor Wafer Cleaning Technology*,[7] and dedicated conferences (e.g., the Microcontamination Conference). Visit the Semiconductor Subway (http://www-mtl.mit.edu/semisubway.html) for frequent updates in this important area.

Silicon-Based Device Manufacturing Costs

The manufacturing of a Si-based device involves many processing steps, and each step adds to the cost of the finished wafer/device. In the most general terms, the finished wafer costs depend on the number of masks used, the complexity of the devices, and the clean room requirements. The costs increase with the number of layers in a device in a nonlinear fashion as defects are introduced in each new layer, but with each new layer more defects accumulate in the underlying layers as well. Cost per device/chip further increases the smaller the feature size because of more stringent requirements for the lithography steps and overall process control. In first world countries, an added concern is that semiconductor fabrication processes are highly toxic; therefore, environmental and worker protection laws need to be considered. In the case of an IC fab, the current cost is in the $2–3 billion range. A typical fab line occupies one city block and employs a few hundred workers. The most profitable period is typically in the first 18 months to 2 years. It is important to note that for large-volume ICs, packaging and testing constitute the largest costs, and that for low-volume ICs, design costs may swamp out the manufacturing costs. In the case of non-IC micro- and nanomachines, testing and packaging become yet more significant contributors to costs. Moreover, standardization is much more difficult, and multiphysics design software packages required for these devices are not yet sophisticated enough.

In Table 1.3 we list the relative costs of IC production processes (does not include design, test, and packaging); lithography is the most expensive step (at 35%), and cleaning adds another 20%.

Variability in the IC manufacturing processes listed in Table 1.3 can lead to deformations or nonconformities in finished products. The product yield is defined as the percentage of devices or circuits that meet a nominal performance specification. One distinguishes between a functional (hard) yield (e.g., opens, shorts, particles) and a parametric (soft) yield (e.g., speed, noise, immunity, power consumption). Functional yield is affected by the presence of defects. Defects result from a number of random sources such as contamination from equipment or processes, handling, mask imperfections, and airborne particles. Functional yield Y is a function of A_c, a critical area in which, if a defect occurs, it has a high probability of inducing a fault, and D_0, the density of defects per unit area, or:

$$Y = f(A_c, D_0) \tag{1.1}$$

This relation depends on circuit geometry, density, feature sizes, number of processing steps, and so on, and is sometimes represented as:

$$Y = \frac{1 - e^{-D_0 A_c}}{D_0 A_c} \tag{1.2}$$

Manufacturing issues that influence hard yield Y include photoresist shrinking/tearing, variations in material deposition, variations in temperature, variations in oxide thickness, impurities, variations between lots, variations across the wafer, and so on.

Table 1.3 Relative Costs of IC Production Processes

Manufacturing Process Step	Percentage of Wafer Processing Cost per Square Centimeter*
Lithography	35%
Multilevel materials and etching	25%
Furnaces and implants	15%
Cleaning/stripping	20%
Metrology	5%

*Excludes packaging, test, and design costs.

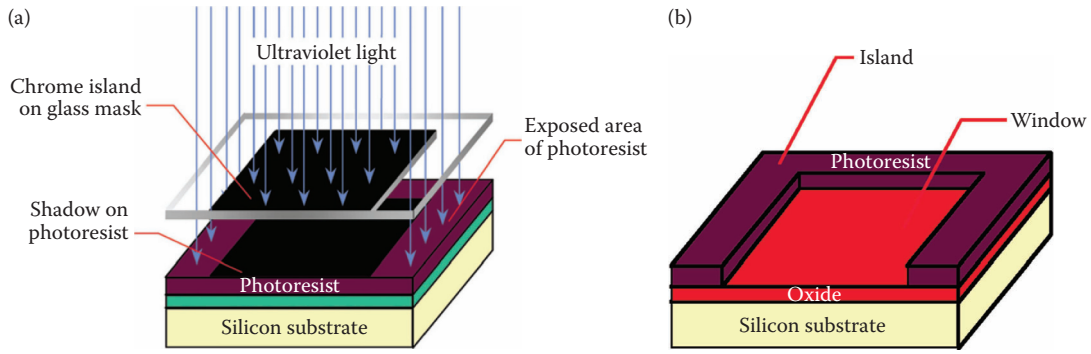


FIGURE 1.4 (a) Photomask in an exposure setup. (b) Resulting pattern in case a negative photoresist is used.

The parametric yield measures the "quality" of functioning systems. Even in a defect-free manufacturing environment, random process variations can lead to varying levels of system performance. Parametric yield may include variations in threshold voltage (because of oxide thickness, ion implantation, poly variations), variations in R, C (because of changes in doping or poly-Si, metal variations in height and width), and shorts and opens (because vias are not cut all the way through, or undersized vias that have too much resistance or oversized vias that short).

Masks

Standard Photolithography Masks

The stencil used to repeatedly generate a desired pattern on resist-coated wafers is called a *mask*. In typical use, a photomask—a nearly optically flat glass [transparent to near ultraviolet (UV)] or quartz plate (transparent to deep UV) with an absorber pattern metal (e.g., an 800-Å-thick chromium layer)—is placed above the photoresist-coated surface, and the mask/wafer system is exposed to UV radiation (Figure 1.4). The absorber pattern on the photomask is opaque to UV light, whereas the glass or quartz is transparent. The absorber pattern on the mask is generated by e-beam lithography, a technique that yields higher resolution than photolithography. In e-beam lithography, a pattern drawn on a computer-aided design (CAD) system is exposed onto the mask. Like resists, masks can be positive or negative. A *positive* or *dark field* mask is a mask on which the pattern is clear with the background dark. A *negative* or *clear field* mask is a mask on which the pattern is dark with the background clear. A light field or dark field image, known as *mask polarity* (Figure 1.5), is then transferred to the semiconductor surface. This procedure results in a 1:1 image of the entire mask onto the silicon wafer.

Masks making direct physical contact (also referred to as *hard contact*) with the substrate are called *contact masks*. Unfortunately, these masks degrade faster because of wear than noncontact, proximity masks (also referred to as *soft contact masks*), which are slightly raised, e.g., 10–20 μm, above the wafer. The defects resulting from hard contact masks on both the wafer and the mask make this method of optical pattern transfer unsuitable for **v***ery* **l***arge* **s***cale* **i***ntegration* (VLSI) manufacturing. In VLSI, ICs have between 100,000 and 1 million components, and in **u***ltra***l***arge* **s***cale* **i***ntegration* (ULSI), there are more than 1 million circuit elements on a single chip. In 2005, Intel began manufacturing computer chips with an average feature size of 65 nm (the so-called *65-nm technology node*, see Volume I, Appendix 1A). This new generation of chips has more than half a billion transistors and constitutes a ULSI chip.

We review hard contact masks because they are still used in research and development, in mask making itself, and for prototyping. Contact mask

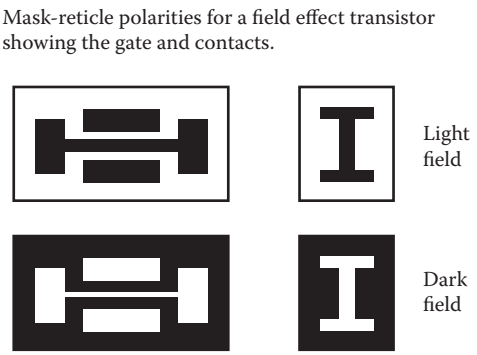


FIGURE 1.5 Light field and dark field.

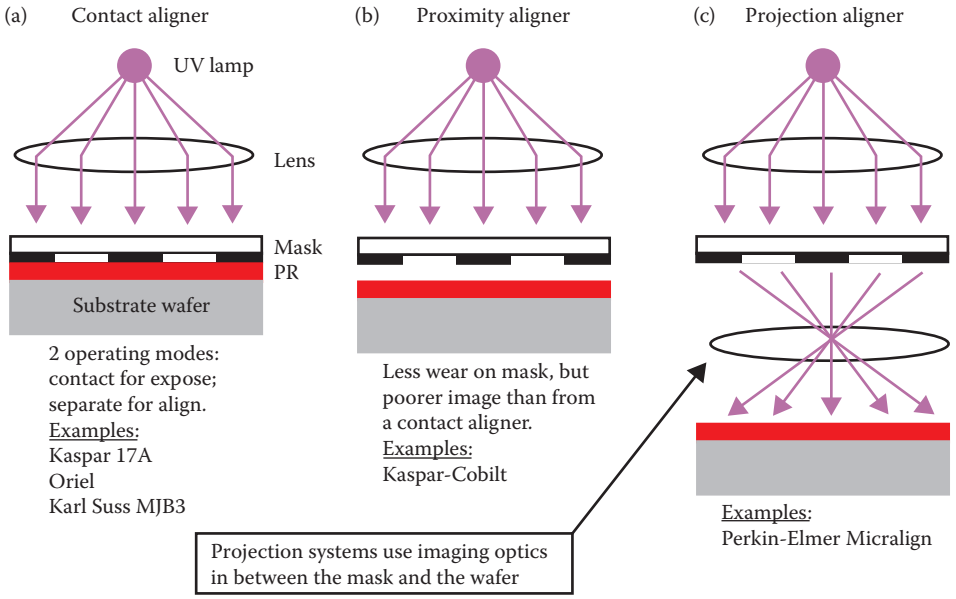


FIGURE 1.6 (a) Contact printing, (b) proximity printing, and (c) projection printing.

and proximity mask printing are collectively known as *shadow printing*. A more reliable method of masking is *projection printing*, where, rather than placing a mask in direct contact with (or in proximity of) a wafer, the photomask is imaged by a high-resolution lens system onto the resist-coated wafer. In the latter case, the only limit to the mask lifetime results from operator handling. The imaging lens can reduce the mask pattern by 1:5 or 1:10, making mask fabrication less challenging. In Figure 1.6, we compare contact, proximity, and projection lithography, and example commercial systems are listed also.

The design of electron beam-generated masks for ICs and miniaturized machines is generally fairly straightforward and requires some suitable CAD software and a platform on which to run it (Figure 1.7). Electron-beam lithography is maskless and is discussed later in this chapter; mask design and suitable CAD software will be addressed in Volume III, Chapter 6.

Gray-Tone Lithography Masks

As described above, photolithography constitutes a binary image transfer process: the developed pattern consists of regions with resist (1) and regions without resist (0). In contrast, in grayscale lithography, the partial exposure of a photoresist renders it soluble to a developer in proportion to the local exposure dose, and as a consequence, after development, the resist exhibits a surface relief.

Gray-tone masks (GTMs), letting varying amounts of light pass through, are of great potential use in

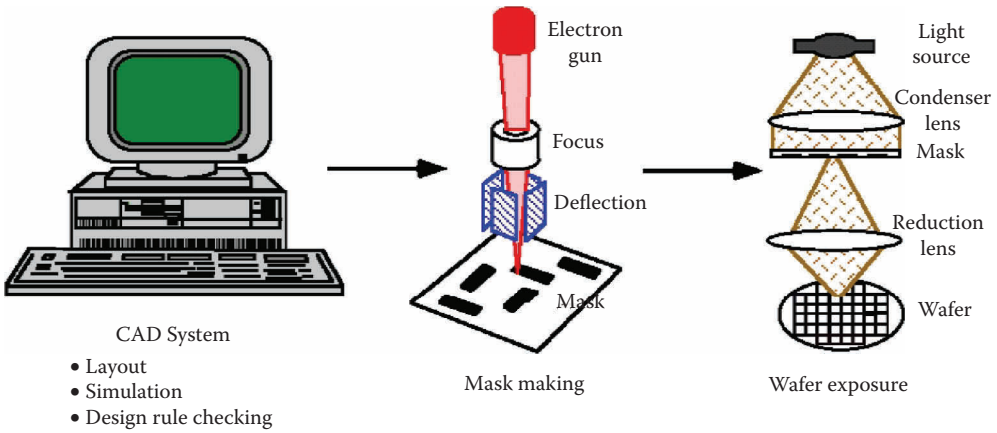


FIGURE 1.7 E-beam writing of a photomask and subsequent projection printing.

miniaturization science as they allow for the mass production of micromachines with varying topography. The possibility of creating profiled micro-3D structures offers tremendous additional flexibility in the design of microelectronic, optoelectronic, and micromechanical components. In optical integrated circuits (OICs), for example, GTMs can be used for the fabrication of tapered waveguides, gratings with saw tooth profiles (blazed gratings), microlenses, and so on.[8] As an example of GTM use, in Figure 1.8, we show experiments with a Rohm and Haas dual-tone resist.[9,10] The right-hand path in Figure 1.8a

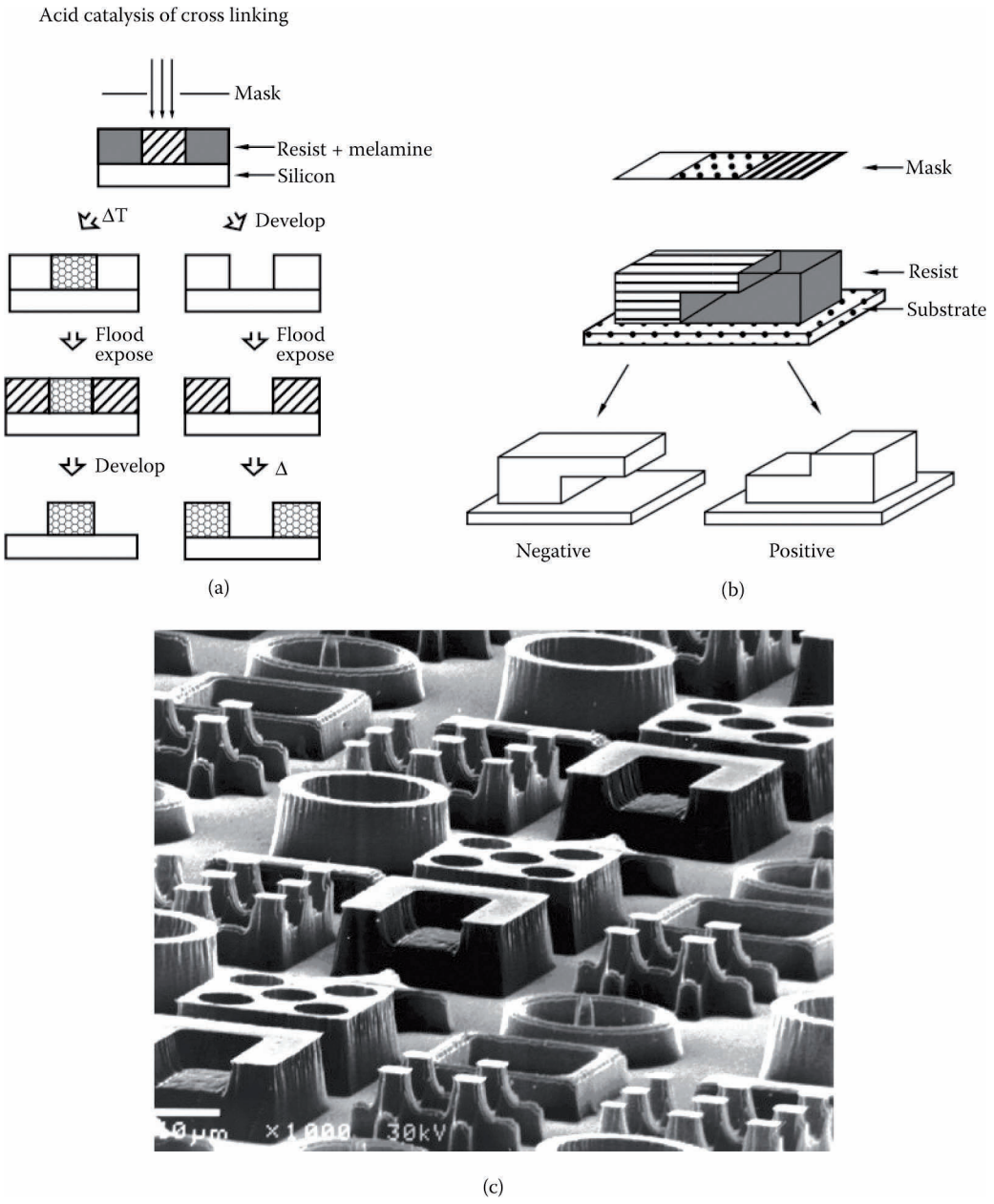


FIGURE 1.8 Rohm and Haas dual-tone resist. (a) Process options for Rohm and Haas dual-tone resist. The right-hand path produces positive-tone images in which the remaining resist is cross-linked and does not flow when exposed to temperatures above the T_g of the novolak-base resin. The left-hand path provides thermally stable negative-tone images. (b) Exposure of the Rohm and Haas resist with a mask exhibiting nominally 0% transmission, 50% transmission, and 100% transmission allows generations of patterns with controlled variations in thickness and/or overhangs and cantilevered structures. (c) Example of the types of features that can be printed using the Rohm and Haas resist. Note the structures with full thickness and 50% thickness. (Courtesy of Drs. Feely, Rohm, and Haas. Feely, W. E. 1988. *Technical digest: solid state sensor and actuator workshop.* Hilton Head Island, SC: IEEE.[9])

produces positive-tone images in which the remaining resist is cross-linked and does not flow when exposed to temperatures above the T_g of the novolak-base resin. The left-hand path provides thermally stable negative-tone images. The mask used here has only 0% or 100% transmission fields. Exposure of the same Rohm and Haas resist with a GTM, exhibiting nominally 0% transmission, 50% transmission, and 100% transmission, allows generations of patterns with controlled variations in thickness and/or overhangs and cantilevered structures. An example of the types of features that can be printed using the Rohm and Haas resist with a GTM is shown in Figure 1.8c. Note the structures with full thickness and 50% thickness.

The fabrication of 3D components by gray-tone technology is performed in three principal steps. The first and most critical step deals with the realization of the GTM itself. The challenge in this step is the definition of zones of variable optical transmission that represent the various gray levels. In a second step, the wafers are exposed, and the light intensity modulation by the gray-tone areas on the mask generates depth variations in the photoresist (Figure 1.9). The profile in the photoresist can also be transferred proportionally into the substrate during a third step, i.e., a dry etching step.[11]

For each substrate/resist combination, the dry etching process has to be established to obtain resist/substrate selectivity close to 1:1.

One approach to make multilevel photoresist patterns directly, without a physical mask, is by variable-dose e-beam writing, in which the electron dosage (current × dwell time) is varied across the resist surface.[13] Variable-dose e-beam writing is slow and costly, and GTMs are a desirable alternative, especially if high throughput production is required. A laser writer can produce the same result but at a lower resolution.

Possible methods for making variable transmission masks include magnetron sputtering of amorphous carbon onto a quartz substrate. Essentially any transmittance (T) desired in the $0\% < T < 100\%$ can be achieved by controlling the film thickness (t) in the $200\text{ nm} > t > 0\text{ nm}$ range with subnanometer precision.[14] Perhaps more elegantly, gray levels may be created by the density of dots that will appear as transparent holes in a chromium mask. These dots are small enough not to be transferred onto the wafer because they are below the resolution limit of the exposure tool. In a CAD program, the dot matrices associated with the gray levels are then combined with the 2D geometries to define the final 3D structures. Both electron-beam and laser-beam pattern generators have been used for manufacturing GTMs. The electron-beam pattern generator is better adapted to this application because of its flexibility and higher resolution.[11] Optimization of the exposure and the development parameters improves the resist depth and profile. The resist profile may then be

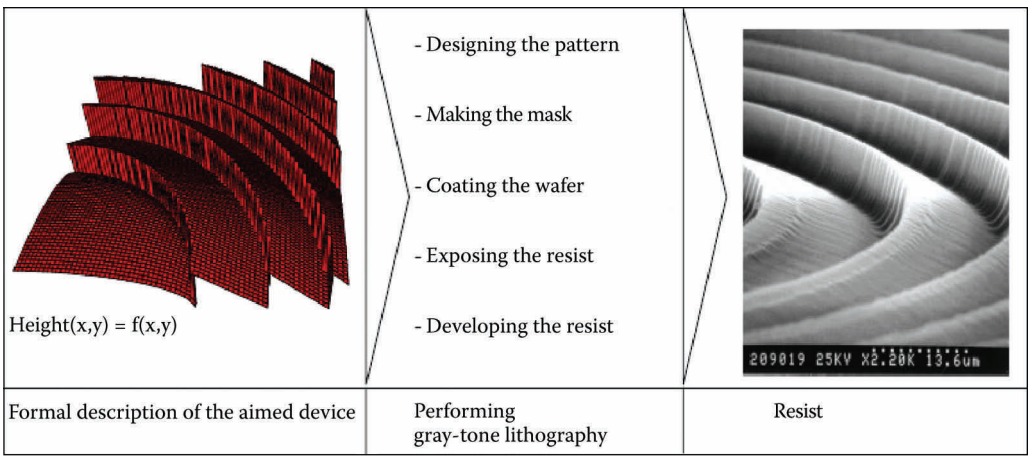


FIGURE 1.9 Process flow chart for transferring the image on a gray-tone mask into a substrate. Gray-tone lithography is an inexpensive one-step lithographic method to fabricate 3D microstructures using a raster-screened (gray-tone) photomask. (Reimer, K., R. Engelke, U. Hofmann, P. Merz, K.T.K.v. Platen, and B. Wagner. 1999. *MF-Conf., micromachine technology for diffractive and holographic optics*. Santa Clara, CA: SPIE.[12])

used directly; it may function as a mold; or it is conformally transferred into the substrate using plasma etching with controlled selectivity. Figure 1.10 shows a blazed grating proportionally transferred from a patterned resist to the underlying fused silica. Blazed gratings with a period of 16 μm and 8 μm covering a surface of 3×3 mm^2 were fabricated with an eight-level GTM. A reflection grating in which the grooves are asymmetric with respect to the surface normal is said to be blazed. A blazed grating is used to direct the light only into the diffracted order of interest with high efficiency. From the atomic force measurements of the grating profile shown here, the distribution of light in the various diffraction orders was computed and found to be in agreement with the experimental optical measurements. Compared with multiple mask techniques, the GTM approach allows the realization of different depths from the same GTM.

Another attractive way to implement grayscale lithography is with high-energy beam-sensitive (HEBS) glass. HEBS glass turns dark on exposure to an electron beam; the higher the electron dosage, the darker the glass turns. In HEBS glass, a top layer, a couple of micrometers thick, contains silver ions in the form of silver-alkali-halide $(AgX)_m$ $(MX)_n$ complex nanocrystallites that are about 10 nm or smaller in size and are dispersed within cavities of the glass

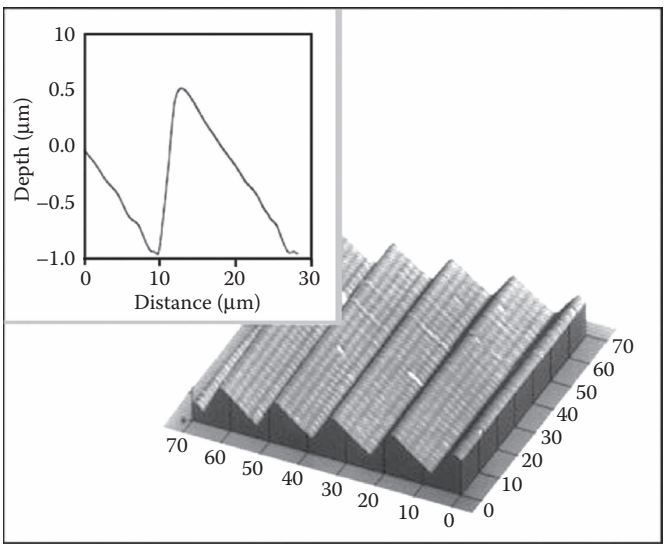
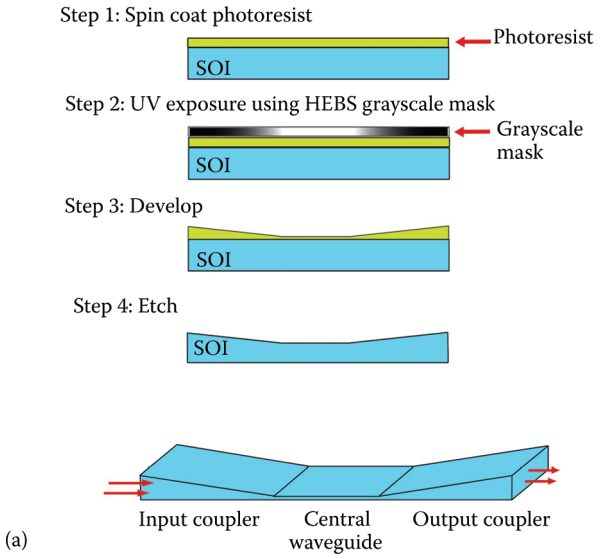


FIGURE 1.10 Atomic force microscope profile of a 16-μm pitch blazed grating in fused silica: the grating was proportionally transferred from resist to fused silica. A sawtooth (blazed) profile notably increases the efficiency of an optical grating.

SiO_4 tetrahedron network. Chemical reduction of the silver ions produces opaque specks of silver atoms on exposure to a high-energy electron beam (>10 kV). A key part in the development of a grayscale process is the characterization of the resist thickness as a function of the optical density in the mask for a given lithographic process. It is desirable to use photoresists that exhibit a low contrast to achieve a wide process window (see "Contrast and Experimental Determination of Lithographic Sensitivity," this chapter). Ideally, the resist response can be linearized to the optical density within the mask. Using an HEBS glass mask, Sure et al.[8] fabricated 3D silicon tapers to couple light from an optical fiber into an optical waveguide for an OIC. Fabrication involves writing a single grayscale mask in HEBS glass with a high-energy electron beam, UV grayscale lithography, and inductively coupled plasma etching (see Figure 1.11a).[8] In Figure 1.11b we show an example HEBS mask for the fabrication of an array of lenses (http://www.canyonmaterials.com).

Inexpensive Masks and Maskless Optical Projection Lithography for Research and Development

In miniaturization science, one often is looking for low-cost and fast-turnaround methods to fabricate masks. This may involve manually drawn patterns on cut-and-peel masking films and photoreduction, affording fast turnaround without relying on outside photomask services. Alternatively, it may involve direct writing on a photoresist-coated plate with a laser plotter (~2-μm resolution).[15] Simpler yet, using a drawing program such as Canvas® (Deneba Systems, Inc.), Freehand® (Macromedia, Inc.), Illustrator® (Adobe Systems, Inc.), or L-Edit® (Tanner Research, Inc.), a mask design can be created on a computer and saved as a Postscript® file to be printed with a high-resolution printer on a transparency.[16] A common laser printer has a resolution of 72 lines per inch (lpi) and between 300 and 1200 dots per inch (dpi) (see Figure 1.12). The average dot size for a 1200-dpi printer is 21 μm, and if we print a vertical line approximately 100 μm wide with this printer, approximately 5 dots make up the 100-μm width, but the space between successive lines of 5 dots is 332 μm. Thus, this vertical



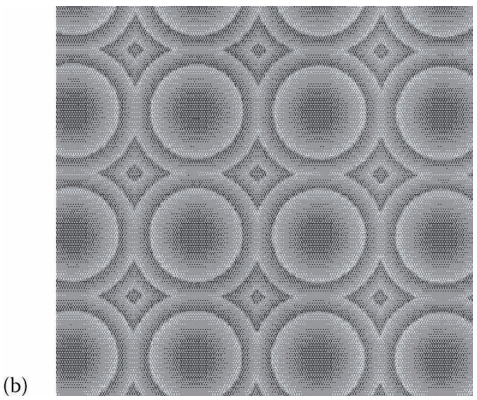


FIGURE 1.11 (a) Fabrication steps for a silicon taper. (Sure, A., T. Dillon, J. Murakowski, C. Lin, D. Pustai, and D.W. Prather. 2003. Fabrication and characterization of three-dimensional Si tapers. *Optics Express* 11:555–61. Drawing by Mary Amasia.[8]) (b) HBS mask for the fabrication of a diffractive microlens array (http://www.canyonmaterials.com).

line will look like a dotted line. Using a typesetting type of printing can avoid these resolution problems (16,000 dpi).

The transparency with the printed image may then be clamped between a presensitized chrome-covered mask plate (i.e., a vendor such as Nanofilm has preapplied the resist) and a blank plate to make a traditional mask from it. After exposure and development, the exposed plate is put in a chrome etch for a few minutes to generate the desired metal pattern, and the remaining resist is stripped off. Simpler yet, the printed transparency may be attached to a quartz plate to be used as a mask directly.

Lithography is still mostly carried out using masks, but because of problems caused by masks, such as expense and time in fabricating them, contamination introduced by them, their disposal, and the difficulties in their alignment, research into maskless optical projection lithography (MOPL) is growing rapidly and broadly. One MOPL approach already on the market (http://www.intelligentmp.com) is based on the digital micromirror device (DMD) chip from Texas Instruments Inc. (TI) and relies on the same spatial light modulation technology used in digital light processing projectors and high-definition television (see Chapter 7, Example 7.10). Enormous simplification of lithography hardware is feasible by using the movable mirror arrays in a DMD chip to project images on the photoresist. This technique is capable of fabricating micromachined elements with any surface topography and can, just like e-beam lithography or laser writing, be used for

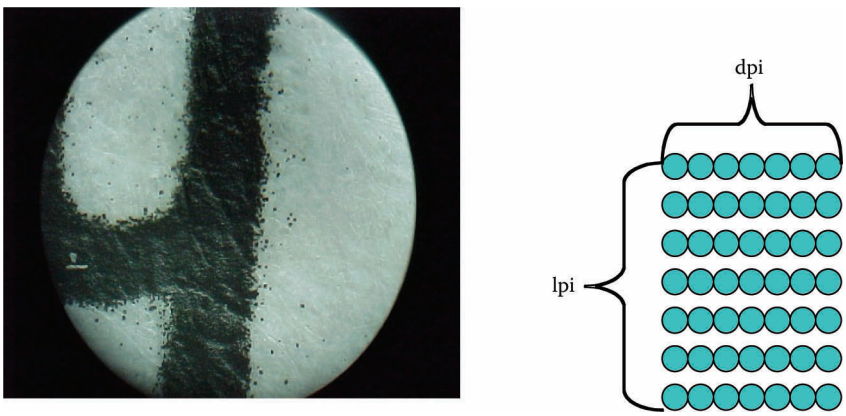


FIGURE 1.12 Laser printer resolution. Image of a laser-printed letter. Note the scatter of small dots. Technology is not so precise as to place dots in a straight line, and the effective dpi might be partially the result of overspray. (From Dr. James Lee, Ohio State University.)

implementing maskless binary and grayscale lithography. The resolution of DMD-based maskless photolithography (about 5 μm) is significantly less than with e-beam lithography (0.25 μm) or laser writers (<1.0 μm), but it is a parallel technique, and for many applications, i.e., in microfluidics, the lower resolution might not be an obstacle. Using DMD technology, on-chip electronics, in conjunction with a control computer, can generate a pattern by turning individual mirrors "on" or "off." DMD chip's micromirrors are mounted on tiny hinges that enable them to tilt either toward the light source in the projection lens (ON) or away from it (OFF) (see Example 7.10), creating a light or dark pixel on the projection surface. The bit-streamed image code entering the semiconductor directs each mirror to switch on and off at a rate of up to 10,000 times/s. This pattern is then demagnified and focused on the wafer to expose the photoresist while light from the remaining off mirrors is deflected away from the optics. The use of refractive optics prevents the extension of MOPL to wavelengths much shorter than 157 nm. The unique capability of representing a grayscale is probably the most essential merit of this type of maskless lithography. When a mirror is switched on more frequently than off, it reflects a light gray pixel; a mirror that is switched off more frequently reflects a darker gray pixel. In this way, the mirrors in a DMD system can reflect pixels in up to 1024 shades of gray to convert the video or graphic signal entering the DMD chip into a highly detailed grayscale image. The method to fabricate the micromirrors used in DMDs is based on MEMS and is detailed in Chapter 7 on surface micromachining. In Figure 1.13 we show a photo of Intelligent Micro Patterning's lithography system, the SF-100, based on TI's DMD with a line-width resolution of 5 μm (http://www.intelligentmp.com/SF100Photos.htm).

Maskless zone-plate-array lithography (ZPAL)[17] is another optical maskless technology. This technology operates on the principle of diffraction rather than refraction. Instead of a single large numerical aperture (NA) lens, an array of thousands of nanofabricated Fresnel zone plate lenses is used, each focusing a beam of light onto the substrate. A computer-controlled array of micromechanical spatial-light modulators turns the light to each lens on or off as the substrate is scanned under the array, thereby printing the desired pattern in a "dot-matrix" fashion.

FIGURE 1.13 Intelligent MicroPatterning's SF-100 lithography system based on DMDs (http://www.intelligentmp.com/SF100Photos.htm).

ZPAL is reviewed in Chapter 2 as a nanolithography technique that can be adapted to work with an extreme UV light source.

Photoresist Deposition

Introduction

Lithography is the most expensive step in microelectronics technology, representing up to 35% of the wafer manufacturing cost (see Table 1.3). Within lithography, photoresist coating is one of the more expensive steps. Photoresists may cost $100/L (and it is about $40/gallon for a typical developer), and several of the resist deposition processes used waste significant amounts of material. For silicon ICs, the resist thickness after a prebake (see below for what a prebake does) typically ranges between 0.5 and 2 μm. For miniaturized 3D structures, much greater resist thicknesses are often required, and complex topography might also call for a conformal resist coat over very high-aspect-ratio features. For thick resist coats, techniques such as casting and the use of thick sheets of dry photoresists replace the ineffective resist spinners. For conformal coating, resist spraying or, better yet, electrodeposition (ED) of photoresist might be preferable. In this section we emphasize methods used to deposit liquid photosensitive materials in ICs, MEMS, and NEMS, including spin coating, spray coating, dip coating,

meniscus coating, electrodeposited (electrophoretic) photoresist, and casting methods such as roller, curtain, and extrusion coating. Photoresist coating is an additive technology, and in Chapter 8 we present a broader review of available processes for the deposition of organic films in general, including photoresists and nonphotosensitive organics for applications such as biosensors, membranes, packaging, protein arrays, DNA arrays, and so on. Very thin photoresist coatings (<<1 μm) are reviewed further below under the section on Thin Photoresists, and in Chapter 10 we explore very thick photoresist coatings (>>20 μm) for molding applications. Lamination of dry photoresist films to a substrate and a comparison of wet versus dry photoresists are covered below in the section on Permanent and Dry Resists.

Spin Coating

A common step before spinning on a resist with Si as the substrate is the growth of a thin layer of oxide on the wafer surface by heating it to between 900 and 1150°C in steam or in a humidified oxygen stream (see Figure 1.1a). Dry oxygen also works, but wet oxygen and steam produce faster results. The oxide can serve as a mask for a subsequent wet etch or boron implantation. As the first step in the lithography process itself, a thin layer of an organic polymer, a photoresist sensitive to UV radiation, is deposited on the oxide surface (see Figure 1.1b). The photoresist is dispensed onto the wafer lying on a wafer platen in a resist spinner (see Figure 1.14).[18] A vacuum chuck holds the wafer in place. A speed of about 500 rpm is commonly used during the dispensing step, enabling the spread of the fluid over the substrate. After the dispense step it is common to accelerate to a relatively high speed to thin the fluid to near its final desired thickness. Typical spin speeds for this step range from 1500–6000 rpm, depending on the properties of the fluid (mostly its viscosity), as well as the substrate. This step can take from 10 seconds to several minutes. The combination of spin speed and time selected for this step will generally define the final film thickness. At these speeds, centrifugal force causes the solution to flow to the edges, where it builds up until expelled when surface tension is exceeded. The resulting polymer thickness, *T*, is a function of spin speed, solution

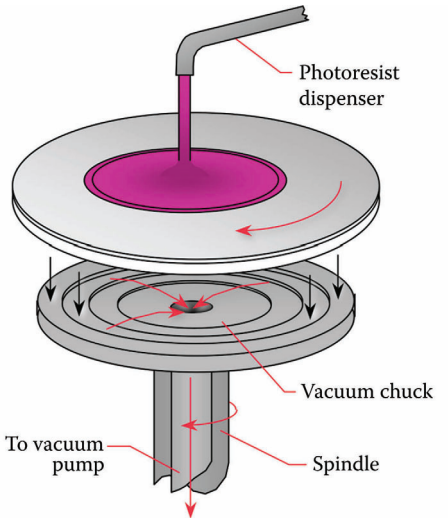


FIGURE 1.14 Resist spinner.

concentration, and molecular weight (measured by intrinsic viscosity). The empirical expression for *T* is given by:

$$T = \frac{KC^{\beta}\eta^{\gamma}}{\omega^{\alpha}} \tag{1.3}$$

where K = the overall calibration constant
C = the polymer concentration in g/100-mL solution
η = the intrinsic viscosity
ω = the number of rotations per minute

Once the various exponential factors (α, β, and γ) have been determined, Equation 1.3 can be used to predict the thickness of the film that can be spun for various molecular weights and solution concentrations of a given polymer and solvent system.[19]

The resist spinning process is of primary importance to the effectiveness of pattern transfer. The quality of the resist coating determines the density of defects transferred to the device under construction. The photoresist film, after application to the substrate, must have a uniform thickness and must be chemically isotropic so that its response to exposure and development is uniform. The coating thickness of the thin, glassy resist film depends on the chemical resistance required for image transfer and the fineness of the lines and spaces to be resolved. The coating process should take place in Class 100 laminar flow air conditions, and a minimum spin time of 30 seconds is necessary for coating uniformity. The

photoresist should also be filtered just before application. The application of too much resist results in edge covering or run-out, hillocks, and ridges, reducing manufacturing yield. Application of too little resist may leave uncovered areas. Optimization of the "regular" photoresist coating process in terms of resist dispense rate, dispense volume, spin speed, ambient temperature, venting of the resist spin station, and humidity presents a growing challenge.

To ensure reproducible line widths and development times in subsequent steps, a resist thickness of 1 μm, with a repeatability of ±10 Å (±0.1%) is becoming the norm. In such cases, even small changes in air draft over the resist-coated substrate could influence the results. In Table 1.4 we present a handy spin coating-troubleshooting chart based on data presented at http://www.brewerscience.com/products/cee/technical/spintheory. In Appendix 1A a more detailed spin coating process troubleshooting list is presented.

Since 2005, computer chips with an average feature size of 65 nm have been on the market, and for these, as well as smaller feature sizes, yet better control over thinner and thinner resist layers is sought. An on-line film thickness monitor, possibly a technique based on reflection spectroscopy, will become essential for statistical process control of such demanding photoresist coatings.[18] The need for an alternative photoresist deposition technique might arise as the amount of waste material generated by spin coating is high, with most of the resist solution (>95%) thrown off the substrate during the spin casting process (the wasted resist must be disposed of as a toxic material). Also inherent to this process is the formation of edge beads, which require an additional removal process (see Figure 1.15a) before subsequent process steps. The edge of a wafer might exhibit resist ridges that are about 10 times the mean thickness on the rest of the substrate. A typical bead removal solution is AZ edge bead remover solvent (http://www.glue.umd.edu/~nima/doc/AZEBRSolventMSDS.pdf). With the current trend toward larger substrate formats, these limitations will be exacerbated. The main obstacle to using spin coating in MEMS is caused by varying topography: deeply etched features cause a physical obstruction to the solution flow, preventing complete coverage and often causing striation or resist thickness variation. For example, resist thickness variations can occur on the near and far sides of a cavity or in cavities at different locations on the substrate. Sizes and shapes of the cavities also have influence on the resist uniformity and coating defects. For substrates with moderate topography, alternative coating techniques, like spray coating (see the next section below), offer better prospects. Wafers with extreme topography, e.g., as encountered in radio frequency-MEMS, require a process with yet higher shape conformality. For the latter application we will see below that ED photoresists are a possible solution.

A good review on spin coating can be found at http://www.brewerscience.com/cee/technical/spintheory.html.

Table 1.4 Spin Coating Troubleshooting Chart

Film too thin		
Spin speed too high	→	Select lower speed
Spin time too long	→	Decrease time during high-speed step
Film too thick		
Spin speed too low	→	Select higher speed
Spin time too short	→	Increase time during high speed step
Exhaust volume too high	→	Adjust exhaust lid or house exhaust damper

Source: Based on http://www.brewerscience.com/cee/technical/spintheory.html.

Alternative Photoresist Deposition Methods

In this section, we compare alternatives to spin coating (Figure 1.15a). We follow closely James Webster's treatise of the subject.[20]

Spray Coating In spray coating, the substrates to be coated pass under a spray of photoresist solution. A diagram depicting this process is shown in Figure 1.15b. The spray system includes an ultrasonic spray nozzle that generates a distribution of droplets in the micrometer range. The resist is pushed out of a pressurized tank via a supply pipe to the spray head. The spray head has a defined aperture where the spray mist is formed. The shape of the spray pattern can be adjusted by a secondary air cushion. Compared with electrostatic spraying (see

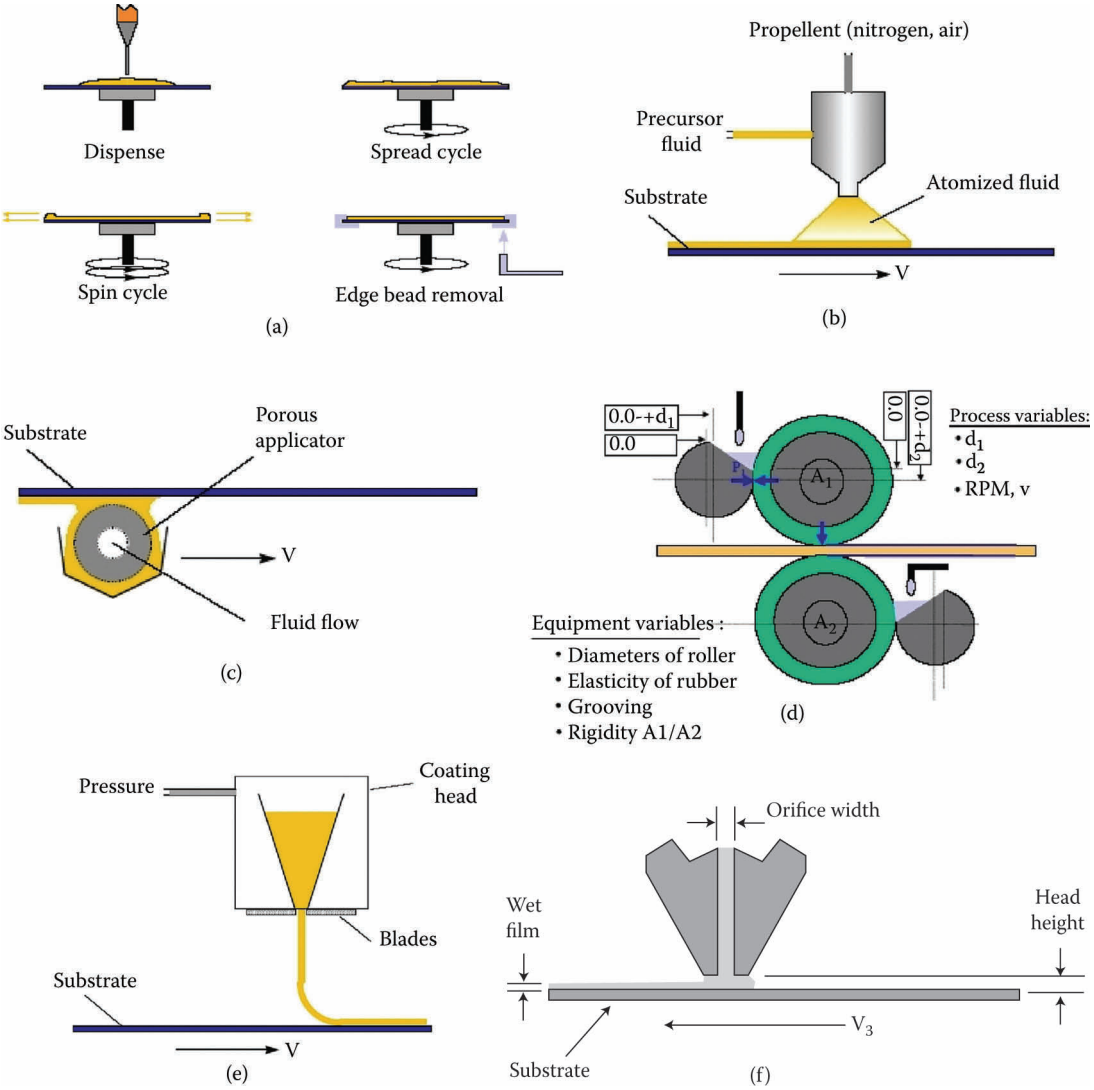


FIGURE 1.15 Photoresist deposition methods: spin coating (a), spray coating (b), meniscus coating (c), roller coating (d), curtain coating (e), and extrusion coating (f) V represents a movement in the direction of the arrow. (Webster, R.J. 1998. *Thin film polymer dielectrics for high-voltage applications under severe environments.* Master's thesis. Virginia Polytechnic Institute and State University.[20])

the coating method described immediately below), the droplets formed in the spray mist are not as small. This can lead to problems in the coating. To avoid the undesired overspray the spray gun must be kept close to the substrate (e.g., 5 cm compared with 30 cm in electrostatic spray). Compared with spin coating, spray coating does not suffer from the resist thickness variation caused by the centrifugal force because the resist droplets are supposed to stay where they are deposited. Spray coating is often used where thicker photoresist films are required, and the uniformity of spray coatings can be held within 5% tolerance when mechanized spray arms are used. Another major advantage is its ability to coat uniformly over nonuniform surfaces, making the technique appropriate for MEMS. During spray coating, the wafer may be rotated slowly while the swivel arm of the spray-coating unit is moved across the wafer. A low spin speed (30–60 rpm) is used to minimize the centrifugal force effects. The rotation allows for resist coverage in all the angles of MEMS cavities. Importantly, sprayed coatings do not have the internal stress forces that are common to spin-coated films. However, control of the deposited film thickness is not as precise as with spin- and extrusion-coated substrates (see last coating method reviewed in this section), and some waste of photoresist solution must be accepted as part of this

processing technique. It is difficult to deposit layers of resist thicker than 20 µm with spray coating. The process can be automated and may coat substrates double-sided.

Electrostatic Spraying Electrostatic spraying or electrodeposition is a variant of spray coating. During the atomization of the resist by air or nitrogen pressure, the droplets formed are given a static charge by applying a large voltage (e.g., 20 kV). The charge causes the droplets to repel each other, maintaining the integrity of the mist of resist formed. Because the droplets follow the electric flux lines, the amount of overspray can be minimized by a set of charged wires keeping the cone of spray to within the desired coating area, thus minimizing overspray. The substrate to be coated can be passed horizontally or vertically through the spray. The latter allows for simultaneous coating on both sides of substrates.

Meniscus Coating Meniscus coating is another viable coating method, especially attractive for large area substrates such as flat panel displays. The method is illustrated in Figure 1.15c. Photoresist solution is pumped through a porous tube (10-µm pores), and a gravity-assisted laminar flow of the liquid is established around the perimeter of the tube. The inverted substrate is made to touch and pass the solution flowing around the tube, and the meniscus of the resist solution adheres to the surface of the moving substrate. Excess fluid that does not adhere to the substrate is collected and recirculated to the coating head. Important process parameters include the separation of the coating head to the substrate, coating velocity, solids content in the resist solution, and evaporation rate of the carrier solvent. These variables control the coating uniformity and thickness. Two important advantages of meniscus coating are the reduction in coating materials loss (only 5–10% loss) and the avoidance of edge buildup (unlike spin coating). Coating techniques involving intimate contact between the substrate and the photoresist supply source, such as meniscus coating or silkscreen printing, lead to good coverage if the adhesion of the resist to the substrate surface is adequate.

Silkscreen Printing Silkscreening can also be used to apply thin, even films of resist for small production runs. Silkscreening of organics is covered in Chapter 8 on chemical, photochemical, and electrochemical-based forming (see Figure 8.10).

Plasma-Deposited Resist Another interesting method is the deposition of resist from the gas phase, i.e., plasma-deposited photoresist. In this method, a monomeric coating precursor is required, which may be evaporated at ambient temperatures and forms a polymer after deposition on the substrate. The latter technique is covered in Chapters Seven and Eight (see for example Figure 8.9).

Electrodeposition (Electrophoresis)[21] Electrophoretic photoresist deposition (ED) is an appropriate technique for the coating of substrates with extreme topography. The ED process requires electrically conductive substrates and electrical biasing during the resist-coating process. The coating solution contains charged micelles comprising resist, solvent, and dye and photoinitiator molecules. The sizes of the micelles are in the 50–200 nm ranges, and coulombic repulsion keeps the particles sufficiently separated to avoid flocculation and settlement. The electrophoretic process was explained in detail in Volume I, Chapter 6; it suffices to mention here that the solution must be rather resistive for a strong enough electrophoretic field to be established. The ED of the photoresist occurs onto a cathodic (negative) or anodic (positive) polarized conductive substrate.[22] In the case of a cataphoretic resist emulsion, the ionized polymer forms positively charged micelles. When the electric field is applied, the positively charged micelles migrate by electrophoresis toward the cathode. When the micelles reach the cathode, their positive surface charges are neutralized by hydroxide ions produced by the electrolysis of water at that electrode. The micelles then become destabilized and coalesce on the surface of the cathode to form a self-limiting, insulating photoresist film. The growth of the film is self-limiting because the resist layer is nonconductive, and once sufficient resist has been deposited to insulate the substrate surface, no further "plating" can occur. Both positive and negative photoresist chemistries for ED are

available (e.g., 2400 ED and Eagle 2100 ED from Shipley Ltd.). Typical coating thicknesses, highly dependent on the voltage and the temperature, are in the range of 5–10 µm, but specific resist systems can be deposited up to about 35 µm on both sides of the substrate simultaneously. The main advantage is that it yields a pore-free deposit of the resist, even at very low thicknesses. Although the most extreme surface topographies can be coated with this technique, this process needs metal plating of substrates and has some associated process complexities, such as the requirement for wafer electrical biasing during the resist-coating process.

Dip Coating In dip coating, one dips the substrate into a solution of liquid resist to apply a coating. The method is simple in that the substrate after cleaning and drying is dipped into a tank containing the photoresist and withdrawn slowly. A thin coating of the resist adheres to the substrate and is fixed by mild baking. The rate of withdrawal and the viscosity and solid contents of the resist are adjusted to obtain the desired thickness. However, it is not possible to get thicknesses greater than about 8–10 µm through this method. The technique does not enable much thickness or uniformity control either, although this method is commonly used in small-volume shops or for odd-sized substrates. The benefits of dip coating are obvious. As a double-sided coating method, all types of formats can be coated. It allows coating thick plates, as well as thin sheets and even wires. Simultaneous coating of multiple sheets increases the productivity significantly. A high and consistent coating quality can be ensured because the few and basic coating parameters can easily be maintained. Dip coating, a process often used to make needle-type biosensors, is covered in more detail in Chapter 8. In the dip-coating process, multiple dip coatings and drying cycles follow each other.

Roller Coating Roller, curtain, and extrusion coating are all variations of directly casting the coating solution on the substrate. Roller coating allows for single- or double-sided coating, and the equipment is less expensive than that used in curtain coating (see next section). Double-sided roller coating is illustrated in Figure 1.15d. Grooved rubber rolls are used to transfer the liquid resist, which is constantly fed into a nip between a doctor bar or roller and the coating roll, to the substrate surface. The pressure between the doctor and coating roller affects the final resist film thickness, as does the pitch of the grooves on the roller, the nonvolatile content of the resist, and the coating roller-to-substrate pressure (see Figure 1.15d). Roller coating can produce very even coatings of low thickness. Excess resist flows back to a sump for recycling via an automatic viscosity controller and filter unit, thus limiting wastage. Roller coaters are often used in-line with a photoresist dryer. The roller-coating technique can also be useful for the application of thick photoresist films to thin substrates. In general, however, it is incapable of producing uniform coatings less than 5 µm in thickness. Other advantages of roller coating include low photoresist waste and adaptability to automation.

Curtain Coating Curtain coating is illustrated in Figure 1.15e; the substrate is moved on a conveyor through a sprayed "curtain" of the resist. The liquid resist is pumped into a head from which the only exit is a thin nip on the head's underside. The resist forced through this nip forms a curtain of resist through which the substrates to be coated are passed. This method can only coat single-sided. The distance from the resist delivery point (the slit) to the substrate is large compared with other coating methods. This method gives high transfer efficiency; however, the thin layer of resist falling through the air results in a significant loss of solvent. Typically, thickness in the 25–60 µm range can be obtained with curtain coating. Undeposited material is recirculated back to the coating head. By careful control of material viscosity, belt speed, and pump speed, reproducible thicknesses can be achieved and maintained over the substrate surface with less than 10% variation in the overall thickness.

Extrusion Coating A method closely related to curtain coating is extrusion coating, as illustrated in Figure 1.15f. In extrusion coating, the extrusion head is positioned at a short, predetermined height above the substrate. Film thicknesses from less than 1 µm to greater than 150 µm in a single coating

pass have been demonstrated with uniformities better than ±3% on substrates up to 350 mm × 400 mm.[20] An important commercial advantage of extrusion coating is that it can be easily integrated into high throughput microelectronic fabrication lines. However, there are some disadvantages associated with extrusion coating, such as the variation in the substrate surface uniformity, which is transferred to the extrusion head during the deposition process. Furthermore, edge beads form along the leading edge of extrusion-coated substrates, although not to the extent encountered in spin-coated films. A thin curtain of the resist is made to fall on the substrate, which is moving horizontally at a controlled rate. As it falls on the substrate, the liquid curtain gently covers it just like a cloth and gets fixed as a film during conveyor movement and drying. This method obviously requires a lot of controls to achieve good results and, if performed well, can yield very satisfactory coatings. It should be noted, however, that coating the two sides of the laminate requires two operations. Control schemes exist in which the fluid flow into the extrusion head and the internal head pressure are measured and passed to a computer that determines the proper motion profile for the extrusion head, thereby improving the uniformity of the coating along the leading edge. The resist extrusion process has been proposed as a lower cost alternative to spin coating for producing highly uniform thin films on large area substrates.

There is no forced drying during roller, curtain, and extrusion coating other than evaporation. Therefore, the coating material has time to flow and planarize over surface features. The degree of coverage into deep features is highly dependent on the surface wettability and the solution viscosity.

A number of the photoresist deposition processes reviewed above are depicted in Figure 1.15.

In Table 1.5, three coating methods of photoresist—spin, spray, and electrodisposition—on large topography MEMS surfaces are compared.[23] Characteristics of each method, as well as advantages and disadvantages, are outlined.

Soft Baking or Prebaking

After resist coating, the resist still contains up to 15% solvent and may contain built-in stresses. Therefore,

Table 1.5 Comparison of Spin Coating, Spray Coating, and Electrodeposition (ED) of Photoresists

	Spin Coating	Spray Coating	ED
Process	Simple Difficult to automate process	Simple Possible for batch production	More complicated (equipment, solution) Batch production
Surface materials	Insulating or conductive	Insulating or conductive	Only on conductive layer Electrical connection to wafer
Parameters	Viscosity Spin speed	Solid content of solution Resist dispensed volume Scanning speed Spray pressure	Voltage Temperature
Photoresist	Several commercially available types	Resist solution with viscosity <20cSt Use much less resist	Special ED resist Resist bath needs frequent refreshing
Resist uniformity	Difficult to control Poor reproducibility Dependent on position of cavities in wafer	Controllable Reproducible Independent on position of cavities	Reproducible and good uniformity (best among three methods)
Suitable application	Transfer patterns to the bottom of etched cavities One level etched and large cavities preferable	Transfer patterns to bottom of etched cavities Cavities with comparable size preferable	Transfer patterns that run in and across cavities Metal patterning preferable

Source: Boellaard, E., P.N. Pham, L.D.M.v.d. Brekel, and G.J. Bertens. 2003. Electrodepositable photoresist processing for RF-devices in MEMS technology. *MEPTEC Rep* Quarter Two: 29–31; and Pham, N.P., E. Boellaard, P.M. Sarro, and J.N. Burghartz. 2002. *SAFE 2002.* Veldhoven, the Netherlands: STW.[22,23]

the wafers are soft baked (also pre-exposure baked or prebaked) at 90–100°C for about 20 minutes in a convection oven or at 75–85°C for 1–3 minutes with a vacuum hot plate to remove solvents and stress and to promote adhesion of the resist layer to the substrate. This is a critical step in that failure to sufficiently remove the solvent will affect the resist profile. Excessive baking destroys the photoactive compound and reduces sensitivity. Thick resists may benefit from a longer bake time. The resist thickness, for both negative and positive resists, is typically reduced by 10–25% during soft baking. Hot plating the resist is faster, more controllable, and does not trap solvent like convection oven baking. In convection ovens, the solvent at the surface of the resist is evaporated first, and this can cause an impermeable resist skin, trapping the remaining solvent inside. Therefore, heating in a convection oven must proceed slowly to avoid solvent bursts. On a hot plate, a smooth substrate surface is needed for good thermal contact and heating uniformity, and the hot plate should be perfectly horizontal. In this case, the temperature increase starts at the bottom of the wafer and works upward, more thoroughly evaporating the coating solvent and leading to a generally much faster and more suitable approach for automation. Commercially, microwave heating or IR lamps are also used in production lines. The optimization of the prebaking step may substantially increase device yield.

Exposure and Postexposure Treatment

After soft baking, the resist-coated wafers are transferred to an illumination or exposure system where they are aligned with the features on the mask (see Figure 1.16). For any lithographic technique to be of value, it must provide an alignment technique capable of a superposition precision of mask and wafer that is a small fraction of the minimum feature size of the devices under construction. In the simplest case, an exposure system consists of a UV lamp illuminating the resist-coated wafer through a mask without any lenses between the two. The purpose of the illumination is to deliver light with the proper intensity, directionality, spectral characteristics, and uniformity across the wafer, allowing a nearly perfect transfer or printing of the mask image onto the resist in the form of a latent image.

In photolithography, wavelengths of the light source used for exposure of the resist-coated wafer range from the very short wavelengths of extreme UV (10–14 nm) to deep UV (150–300 nm) to near UV (350–500 nm). In near UV, one typically uses the g-line (435 nm) or i-line (365 nm) of a mercury lamp. The brightness of shorter-wavelength sources is severely reduced compared with that of longer-wavelength sources, and the addition of lenses further reduces the efficiency of the exposure system. For example, the total collected deep UV power for a 1-kW mercury-xenon lamp in the 200–250 nm range is only 30–40 mW; the additional optics absorb more

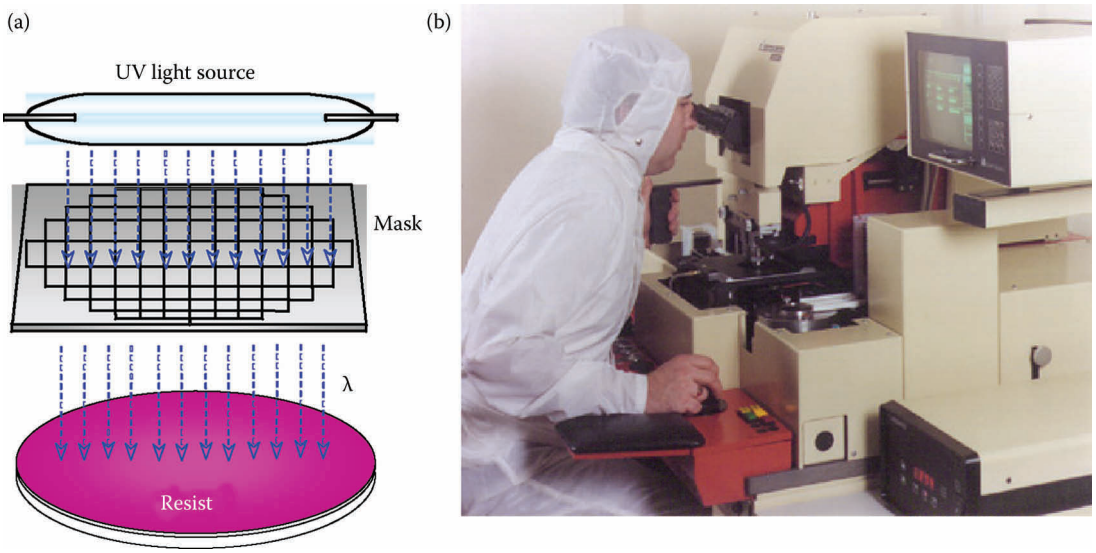


FIGURE 1.16 (a) Exposure of resist-coated wafer to UV light source. (b) An exposure station.

energy of the short wavelengths passing through them. As a consequence, with shorter wavelengths, higher resist sensitivity is required, and newer deep UV sources that produce a higher flux of deep UV radiation must be used. For example, a KrF excimer laser with a short wavelength of 248 nm and a power of 10–20 W at that wavelength is used. In general, the smallest feature that can be printed using projection lithography is roughly equal to the wavelength of the exposure source; in this example, 248 nm would be expected. The same laser, in combination with sophisticated resolution-enhancing techniques (RETs), may be used to produce more advanced circuits with transistor gate features of 160 nm and less. RET methods (see below) enable one to go a bit beyond the conventional Rayleigh diffraction limit. Other exposure systems now available include two deep UV excimer lasers, the ArF at 193 nm and the F_2 at 157 nm, and extreme UV lithography at 13.4 nm. The current generation of lithography uses 193-nm light[24] from ArF lasers. In the case of extreme UV, a plasma or synchrotron source and all-reflective reduction optics (4×) (at this wavelength all materials absorb!) are used. Refractive optical elements are too absorbing at those wavelengths.[25] Extreme UV as a commercial tool is in the exploratory phase only and is discussed in detail in Chapter 2 on next generation lithographies (NGLs). Step and scan printing extreme UV systems are expected to come on-line by the end of this decade. In Figure 1.17 we show the various technologies under investigation for the development of pilot and production lines of large scale integrated (LSI) circuitry.

The incident light intensity (in W/cm^2) multiplied by the exposure time (in seconds) gives the incident energy (J/cm^2) or dose, D, across the surface of the resist film. Radiation induces a chemical reaction in the exposed areas of the photoresist, altering the solubility of the resist in a solvent either directly or indirectly via a sensitizer. During the latent image-forming reaction, the sensitizer in the resist usually *bleaches*; in other words, exposed resist is rendered transparent to the incoming wavelength. This bleaching allows the use of thick films with high absorbency because light will reach the substrate through the bleached resist. The absorbency of the unexposed resist should not reach 40% to avoid degradation of the image profile through the resist depth because

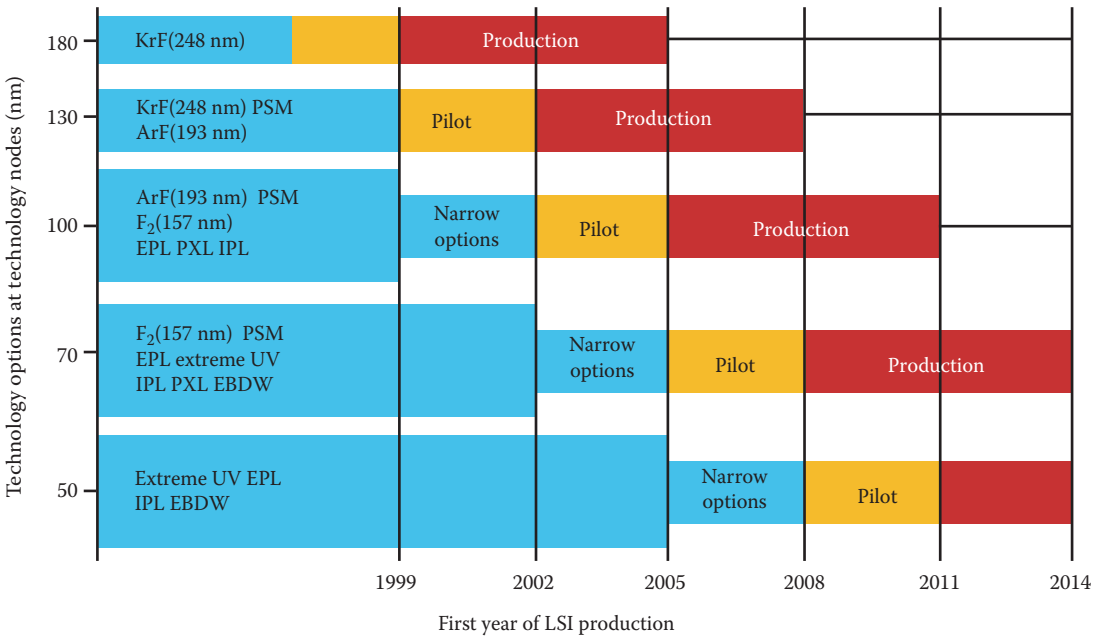


FIGURE 1.17 Various technologies under investigation for the development of pilot and production lines of LSI circuitry. These include KrF (248 nm), KrF excimer laser lithography with wavelength 248 nm; ArF (193 nm), ArF excimer laser lithography with wavelength 193 nm; F_2 (157 nm), F_2 excimer laser lithography with a wavelength of 157 nm; PSM, phase-shifting mask applied to KrF, ArF, and F_2; EPL, electron-projection lithography; PXL, proximity X-ray lithography; IPL, ion-projection lithography; EBDW, electron-beam direct writing. (Ito, T., and S. Okazaki. 2000. Pushing the limits of lithography. *Nature* 406:1027–31.[26])

too large a percentage of the light is absorbed in the top layer. On the other hand, with the absorbency far less than 40%, exposure times required to form the image become too long. The smaller the dose needed to "write" or "print" the mask features onto the resist layer with good resolution, the better the lithographic sensitivity of the resist.

A postexposure treatment of the exposed photoresist is often desired because the reactions initiated during exposure might not have run to completion. To halt the reactions or to induce new ones, several postexposure treatments are in use: postexposure baking (PEB), flood exposure with other types of radiation, treatment with a reactive gas, and vacuum treatment. PEB (sometimes in a vacuum) and treatment with reactive gas are used in image reversal and dry resist development. In the case of a chemically amplified resist, the PEB is most critical. Although reactions induced by the catalyst that forms during exposure take place at room temperature, their rate is greatly increased by baking at 100°C. The precise control of this type of PEB critically determines the subsequent development itself. A PEB improves line-width control by improving adhesion, reduction of scumming (resist left behind after development), increasing contrast and resist profile (higher edge-wall angle), and reducing the effects of standing waves (see Figures 1.69c and 1.70) in regular positive resist. A typical PEB used for the positive resist OCG 895i involves a temperature of 115°C on a hot plate held for 1 minute. Image reversal, dry resist development, and chemically amplified resists will be discussed below.

Development

Development is the dissolution of unpolymerized resist that transforms the latent resist image, formed during exposure, into a relief image that will serve as a mask for further subtractive and additive steps. During the development of an exposed resist, selective dissolving takes place (see Figure 1.1d). Two main technologies are available for development: wet development, illustrated in Figure 1.18, is widely used in circuit and miniaturization manufacture in general, and dry development, which is starting to replace wet development for some of the ultimate line-width resolution applications.

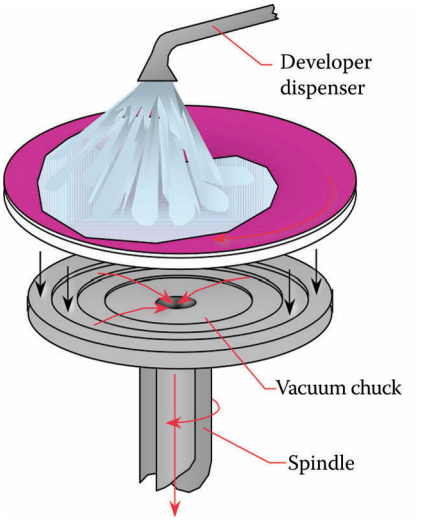


FIGURE 1.18 Wet development.

Wet development by solvents can be based on at least three different types of exposure-induced changes: variation in molecular weight of the polymers (by cross-linking or by chain scission), reactivity change, and polarity change.[27] Two main types of wet development setups are used: immersion (Figure 1.18) and spray developers (Figure 1.19). During batch immersion developing, cassette-loaded wafers are batch-immersed for a timed period in a developer bath and agitated at a specific temperature.

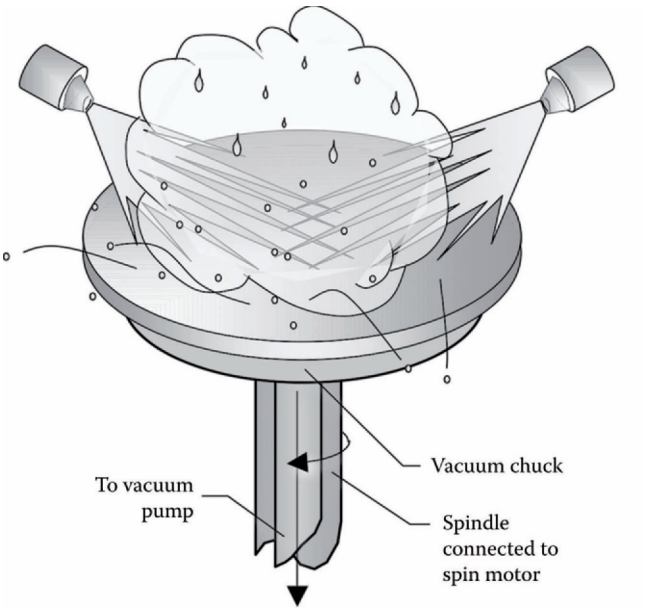


FIGURE 1.19 Spray developer: fresh developing solution is directed across wafer surfaces by a fan-type spraying nozzle. The renewal of developer allows uniform developer strength to be maintained.

During batch spray development, fan-type sprayers direct fresh developing solution across wafer surfaces. Positive resists are typically developed in aqueous alkaline solutions, and negative resists are developed in organic ones. For alkaline developers, simple solutions of NaOH (Shipley 351) or KOH (AZ/Clariant 400K) may be used, but because of the possibility of mobile ion contamination in metal oxide semiconductor (MOS) devices, developers free of metal ions are preferred. The latter are usually tetramethyl ammonium hydroxide (TMAH) based (Shipley CD-26, MF-321, or AZ/Clariant 300-MIF). Each developer has a different dilution, and some require longer development times than others. Developers are generally matched to a type of photoresist. Although they may be interchangeable to some extent, changing the type of developer used in a process will usually change the exposure time necessary to resolve the pattern. It should be noted that all of the above-mentioned alkaline developers etch aluminum. If the micromachine or microchip under construction contains aluminum features, Shipley Microposit Developer Concentrate, a mixture of proprietary alkaline salts (mostly phosphates) with the slowest aluminum etch rate, is a better developer choice.

Aqueous development is highly favored for health reasons. The aqueous development rate depends on the pH of the developer and the temperature, which needs to be controlled to within ±0.5°C.[19] Surfactants and other wetting agents added to the developer ensure uniform wetting, and buffers provide a more stable operating window and a longer lifetime. The more newly developed negative resists also may be developed in aqueous solutions.

The use of organic solvents leads to some swelling of the resist (especially for negative resists) and loss of adhesion of the resist to the substrate. Dry development overcomes these problems, as it is based either on a vapor phase process or a plasma.[28] In the latter, oxygen-reactive ion etching is used to develop the latent image. The image formed during exposure exhibits a differential etch rate to oxygen-reactive ion etching rather than differential solubility to a solvent.[19] Dry developed resists should not be confused with dry film resists, which are resists that come in film form and are laminated onto a substrate rather than spin coated.

With continued pressure by the U.S. Environmental Protection Agency for a cleaner environment, dry development and dry etching (see Chapter 3) are becoming the predominant technologies to use. Dry developed resists, such as the DESIRE process, where the surface of the exposed resist is treated with a silicon-containing reagent, will be discussed below.

Descumming and Postbaking

A mild oxygen plasma treatment, so-called *descumming*, removes unwanted resist left behind after development. Negative and, to a lesser degree, positive resists leave a thin polymer film at the resist/substrate interface. The problem is most severe in small (<1 μm) high-aspect-ratio structures where the mass transfer of a wet developer is poor. Patterned resist areas are also thinned in the descumming process, but this is usually of little consequence.

Before etching the substrate or adding a material, the wafer must be postbaked. Postbaking or hard baking removes residual coating solvent and developer and anneals the film to promote interfacial adhesion of the resist that has been weakened either by developer penetration along the resist/substrate interface or by swelling of the resist (mainly for negative resists). Hard baking also improves the hardness of the film and avoids solvent bursts during vacuum processing. Improved hardness increases the resistance of the resist to subsequent etching steps. Postbaking frequently occurs at higher temperatures (120°C) and for longer times (e.g., 20 min) than soft or prebaking. The major limitation for heat application is excessive flow or melt, which degrades wall profile angles and makes it more difficult to remove the resist. Postbake induces some stress and resist shrinkage. Special care needs to be taken when the baking temperature is above the glass transition temperature, T_g, when impurities are easily incorporated into the resist because of the plastic flow of the resist. Positive resists withstand higher heating temperatures than negative resists, but their stripping proves more difficult. Descumming and postbaking both follow step d in Figure 1.1.

Resist does not withstand long exposure to etchants well. As a consequence, with 1:7 buffered

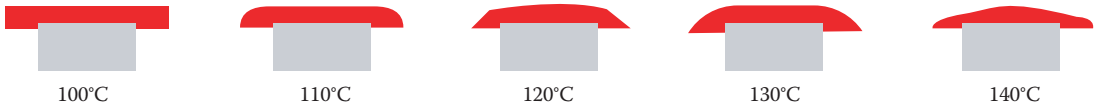


FIGURE 1.20 Tailoring resist sidewalls through resist reflow.

HF (a mixture of 1 part 49% aqueous HF solution and 7 parts NH_4F that is used to strip SiO_2), the postbake sometimes is repeated after 5 minutes of etching to prolong the lifetime of the resist layer (see Figure 1.1e). Also, postbaking should be prolonged before electroplating but is not needed for processes in which a soft resist is desired, e.g., in metal liftoff patterning (see Figure 1.37). As mentioned before, photoresist will undergo plastic flow with sufficient time and/or temperature, and this resist reflow may be used for tailoring resist sidewalls (Figure 1.20).

Resists

The principal components of photoresists are a polymer (base resin), a sensitizer, and a casting solvent. The polymer changes structure when exposed to radiation; the solvent allows spin application and formation of thin layers on the wafer surface; sensitizers control the chemical reactions in the polymeric phase. Resists without sensitizers are single-component or one-component systems, whereas sensitizer-based resists are two-component systems. Solvent and other potential additives do not directly relate to the photoactivity of the resist.

Resist Tone

If the photoresist is of the type called *positive* (also *positive tone*), the photochemical reaction during exposure of a resist weakens the polymer by rupture or scission of the main and side polymer chains, and the exposed resist becomes more soluble in developing solutions (e.g., 10 times more soluble). In other words, the development rate, R, for the exposed resist is about 10 times faster than the development rate, R_0, for the unexposed resist. If the photoresist is of the type called *negative* (also *negative tone*), the reaction strengthens the polymer by random cross-linkage of main chains or pendant side chains, becoming less soluble (slower dissolving). Exposure, development, and pattern-transfer sequences for negative and positive resists are shown in Figure 1.21. Figure 1.22 illustrates chain scission (positive resists) and cross-linking (negative resists).

Positive Resists Two well-known families of positive photoresists are the single component poly (methylmethacrylate) (PMMA) (Figure 1.23) resists and the two-component diazonaphtoquinone (DNQ)

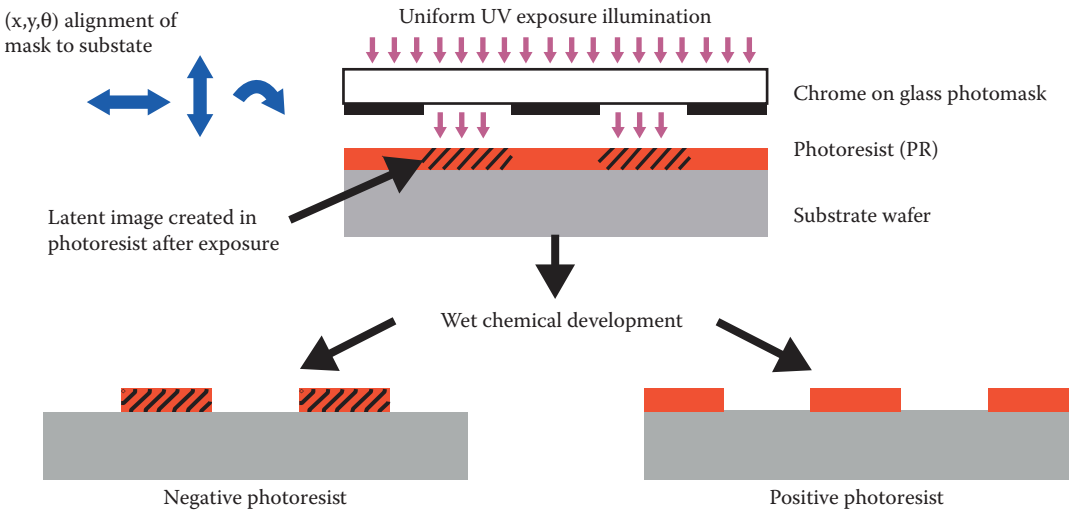


FIGURE 1.21 Positive and negative resist: exposure, development, and pattern transfer. (a) Negative resists remain in the exposed region. (b) Positive resists develop in the exposed region.

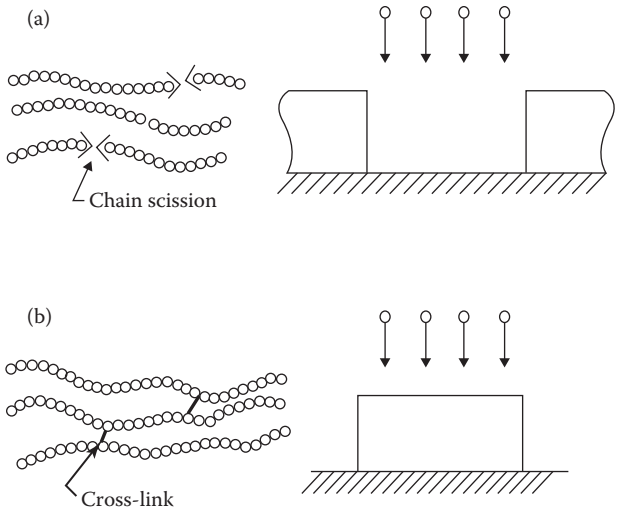


FIGURE 1.22 (a) Polymer chain scission (positive resists) and (b) cross-linking (negative resists).

(Figure 1.24) resists composed of a photoactive component such as DNQ ester (20–50 wt%) and a phenolic novolak resin.

PMMA becomes soluble through chain scission under deep UV illumination. The maximum sensitivity of PMMA lies at 220 nm; at greater than 240 nm the resist becomes insensitive. PMMA resin by itself constitutes a rather insensitive or slow deep UV photoresist requiring doses >250 mJ/cm^2. Exposure times of tens of minutes were required with the earliest PMMAs available.[28] By adding a photosensitizer such as t-butyl benzoic acid, the UV spectral absorbency of PMMA is increased, and a 150-mJ/cm^2 lithographic sensitivity can be obtained. PMMA is also used in electron beam, ion beam, and x-ray lithography. Although the resolution of PMMA is very good, its plasma etch tolerance is very low.

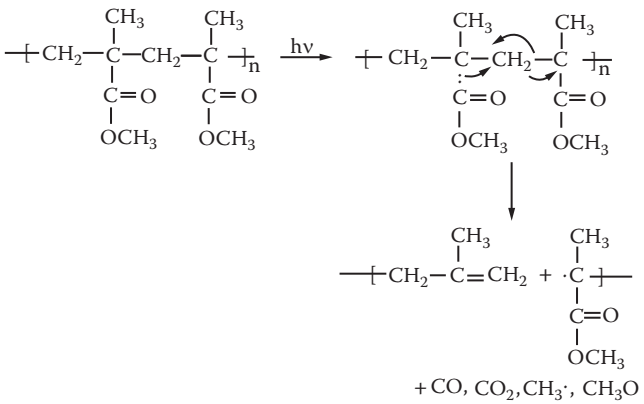


FIGURE 1.23 Poly(methylmethacrylate) or PMMA. Photoinduced chain scission of PMMA resist.

Also, dissociation of PMMA changes the chemistry of a plasma etch and often leads to polymeric deposits on the surface of the substrate.

The DNQ resist system is a "workhorse," near-UV, two-component positive resists, which photochemically transform into a polar, base-soluble product.[24] The hydrophilic novolak resin (N) is in itself alkali soluble because of the OH groups. Diazonaphthaquinone (DQ) is a hydrophobic and nonionizable compound, and when phenolic resins are impregnated with DQ, they become hydrophobic and are rendered insoluble. The addition of 20–50 wt% DQ forms a complex with the phenol groups of the novolak resin and reduces the solubility rate of the unexposed resist to less than 1–2 nm·s^{-1}. During exposure, DNQ undergoes photolysis that destroys the inhibitory effect of DQ on film dissolution (Figure 1.24). The photolysis causes the DNQ to undergo a reaction forming a base-soluble carboxylic acid that can be rapidly developed in aqueous solution of hydroxide ions (e.g., 1% NaOH). In contrast with cross-linked resists, the film solubility is controlled by chemical and polarity difference rather than molecular size. The matrix novolak resin itself is a condensation product of a cresol

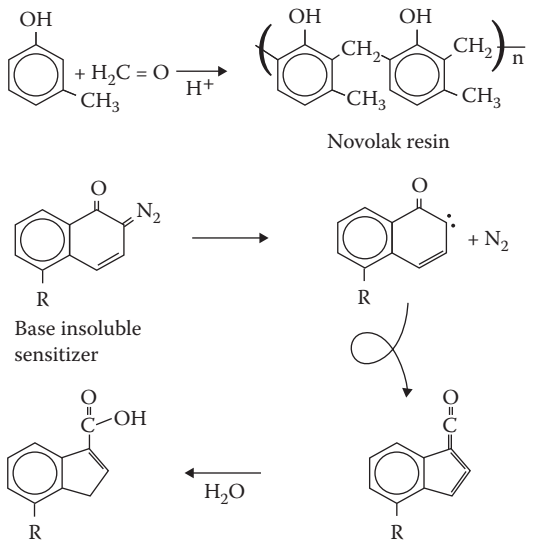


FIGURE 1.24 Phenolic novolak resin and the influence of diazonaphthoquinone ester (DNQ) on it. The novolak matrix resin is prepared by acid copolymerization of cresol and formaldehyde. The base-insoluble sensitizer, a DNQ, undergoes photolysis to produce a carbine, which then undergoes a rearrangement to form a ketene. The ketene reacts with water, present in the film, to form a base-soluble, indenecarboxylic acid photoproduct.

isomer (*para-cresol*) and formaldehyde consisting of hydrocarbon rings with two methyl groups and 1 OH group attached. Phenolic resins are readily cross-linked by thermal activation into rigid forms (bakelite was the first thermosetting plastic). A novolak resin absorbs light less than 300 nm, and the DNQ addition adds an absorption region around 400 nm. The 365-, 405-, and 435-nm mercury lines can all be used for exposure of DNQ. The intense absorption of aromatic molecules prevents the use of this resist at exposing wavelengths less than about 300 nm; at those shorter wavelengths, linear acrylate and methacrylate copolymers have the advantage. Positive attributes of novolak-based resist are that the unexposed areas are essentially unchanged by the presence of the developer. Thus, line width and shape of a pattern are precisely retained. Moreover, novolak is a long-chain aromatic ring polymer that is fairly resistant to chemical attack. Therefore, the photoresist is a good mask for the subsequent plasma etching. Moreover, unlike radical-based systems, it shows no oxygen inhibition.

Most positive resists are soluble in strongly alkaline solutions (a fact which is taken advantage of for stripping of the resist; see Figure 1.1f and text below) and develop in mildly alkaline ones (as shown in Figure 1.1d). Some typical industrially used developers for positive resists are KOH (e.g., a 0.05–0.5 N aqueous solution and a surfactant), TMAH, ketones, and acetates. As mentioned above, besides changing the molecular weight of the resist, radiation-induced reactions may also change the resist's solubility by altering its hydrophilicity (polarity) or its reactivity. Typical casting solvents for positive resists are Cellosolve* acetate, methyl Cellosolve, and aromatic hydrocarbons.

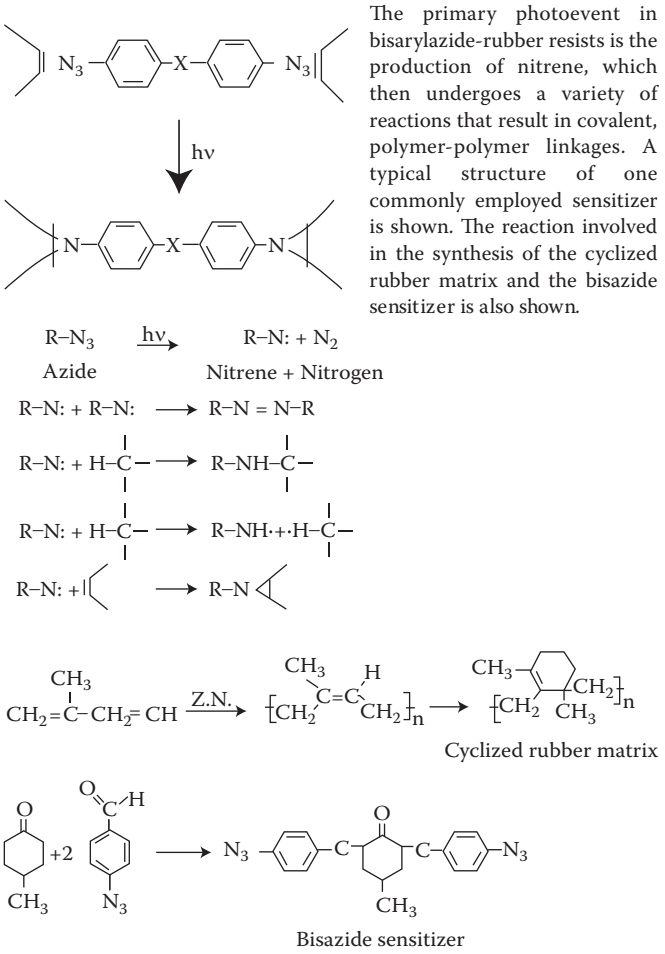
Negative Resists The first negative photoresists were based on free radical-initiated photo cross-linking processes of main or pendant polymer side chains, rendering the exposed parts insoluble. They were the very first types of resists used to pattern semiconductor devices and still comprise the largest segment of the overall photoresist industry, being widely used to define circuitry in printed wiring boards (PWBs).[29] A negative photoresist becomes insoluble in organic (more traditional negative resists) or water-based developers (newer negative resist systems) on exposure to UV radiation. The insoluble layer forms a "negative" pattern that is used as a stencil (usually temporarily) to delineate many levels of circuitry in semiconductors, MEMS, and PWBs. The insolubilization of radiated negative resists can be achieved in one of two ways: the negative resist material increases in molecular weight through UV exposure (traditional negative resists), or it is photochemically transformed to form new insoluble products (newer negative resist products). The increase in molecular weight is generally accomplished through photoinitiators that generate free radicals or strong acids, facilitating polymeric cross-linking or the photopolymerization of monomeric or oligomeric species. Photochemical transformation of negative photoresists may also lead to hydrophobic or hydrophilic groups, which provide another means of inducing preferential solubility between the exposed and unexposed resist film.

Commonly used negative-acting, two-component resists are bis(aryl)azide rubber resists (Figure 1.25), whose matrix resin is cyclized poly(*cis*-isoprene), a synthetic rubber. Cyclized rubber is obtained by acid treatment of poly(*cis*-isoprene), a process that leads to some ring formation in the polymer and stiffens it, thereby raising its glass transition temperature (T_g). For use in a photoresist, a resin should have a molecular weight of around 100,000–200,000 to ensure proper viscosity, melting point, softening point, and stiffness. To get a molecular weight of 100,000 using isoprene (M = 68.12 g/mol), one needs chains of an average length of n = 100,000/68.12 or 1,468 monomer units. This represents a molecule that is too long for proper photolithographic resolution, so the chains need to coil up to make them shorter and to increase the mechanical stiffness of the resist. Therefore, *cis*-isoprene is used to curl up the chains into rings; *trans*-isoprene would not work because of steric hindrance of the CH_3 groups.

The earliest negative photoresists, introduced in the 1960s, were based on the photo cross-linking of

* Cellosolve is a trade name for solvents based on esters of ethylene glycol; these solvents have been identified as possible carcinogens.

Bis(aryl)azide-sensitized rubber resists with cyclized poly(cis-isoprene) as matrix resin



The primary photoevent in bisarylazide-rubber resists is the production of nitrene, which then undergoes a variety of reactions that result in covalent, polymer-polymer linkages. A typical structure of one commonly employed sensitizer is shown. The reaction involved in the synthesis of the cyclized rubber matrix and the bisazide sensitizer is also shown.

FIGURE 1.25 Bis(aryl)azide-sensitized rubber resists with cyclized poly(*cis*-isoprene) as matrix resin. The primary photoevent in bis(aryl)azide rubber resists is the production of nitrene, which then undergoes a variety of reactions that result in covalent, polymer-polymer linkages. A typical structure of one commonly used sensitizer is shown.

"cyclized rubber" using an additive that contained two azide groups such as bis(aryl)azide. When this bis-azide is photolyzed, it loses nitrogen and produces two highly reactive nitrene moieties.

The nitrene intermediates undergo a series of reactions that results in the cross-linking of the resin. The nitrenes can form bonds with the cyclized rubber in a variety of ways, the most common of which is aziridine ring formation. Oxidation of the nitrene intermediate, with oxygen from the ambient or dissolved in the polymer to form azoxy and nitrose products, often competes with the addition and insertion reactions, reducing the extent of the polymer-polymer cross-linking. This phenomenon is called oxygen inhibition. In other words, polymerization can be inhibited by the quenching of the cross-linking reactions through scavenging of the nitrene photoproduct by oxygen. This competing reaction represents a disadvantage, as exposure must be carried out under a nitrogen blanket or in a vacuum. This type of resist was successfully used in the semiconductor industry when printed in "hard contact," minimizing the effect of the oxygen inhibition. With the introduction of soft contact (proximity mode) and off-contact (projection mode), the resist showed severe film thinning (scalping) as a direct consequence of the reduced cross-linking density near the air/polymer interface. Today, cyclic poly(*cis*-isoprene)-based resists play only a small roll in the semiconductor industry.

Another disadvantage of negative resists is that the resolution is limited by film thickness. The cross-linking process starts topside, where the light hits the resist first. Consequently, overexposure is needed to render the resist insoluble at the substrate interface. The greater the desired resist thickness, the greater the overdose needed for complete polymerization and the larger the scattered radiation. Scattered radiation at the resist/substrate interface in turn reduces the obtainable resolution. Moreover, the organic solvent developer swells the cross-linked negative image, which further degrades the resolution. In a practical situation, this leads to a 2–3-μm maximum resolution in a 1-μm-thick resist layer. To improve the resolution of a negative resist, thinner resist layers can be used; however, when using thin layers of negative resist, pinholes become problematic. Xylene is the most commonly used aromatic solvent for negative resists, although almost any organic solvent will do. Aromatic solvent developers may pose environmental, health, and safety concerns. Newer negative resists are water developable.

An example of a commercial, two-component negative photoresist is the Kodak KTFR [an azide-sensitized poly(isoprene) rubber] with a lithographic sensitivity (also photospeed) of 75–125 mJ/cm^2. Negative photoresists, in general, adhere very well to the substrate, and a vast amount of compositions are available (stemming from research and development work in paints, UV curing inks, and adhesives

all based on polymerization hardening). Negative resists are highly resistant to acid and alkaline aqueous solutions, as well as to oxidizing agents. As a consequence, a given thickness of negative resist is more chemically resistant than a corresponding thickness of positive resist. This chemical resistance ensures better retention of resist features even during a long, aggressive wet or dry etch. Negative resists also are more sensitive than positive resists but exhibit a lower contrast (γ smaller; see "Contrast and Experimental Determination of Lithographic Sensitivity," this chapter).

A comparison of negative and positive photoresist features is presented in Table 1.6. This table is not exhaustive and is meant only as a practical guide for selection of a resist tone. The choice of whether to use a negative or a positive resist system depends on the needs of the specific application, such as resolution, ease of processing, speed, and cost. The choice of resist tone will even depend on the specific intended pattern geometry, which is known as the *optical proximity effect*. For example, an isolated single line most easily resolves in a negative resist (higher resolution line), whereas an isolated hole or trench is most easily defined in a positive resist. Because traditional negative resists used to have a line-width limit of only about 2–3 µm and because the industry moved away from organic solvent-based systems in favor of less toxic, water-based developers, positive resists gained in popularity. However, traditional negative resists continue to be used in the production of PWBs and low-cost, high-volume chips

Table 1.6 Comparison of Traditional* Negative and Positive Photoresists

Characteristic	Resist Type	
	Positive Resist	Negative Resist
Adhesion to silicon	Fair (priming required)	Excellent (priming not required)
Available compositions	Many	Vast
Baking	In air (+)	In nitrogen (–)
Contrast γ	Higher, e.g., 2.2	Lower, e.g., 1.5
Cost	More expensive	Less expensive
Developer	Temperature sensitive (–) and aqueous based (Ecologically sound)	Temperature insensitive (+) and organic solvent (–)
Developer process window	Small	Very wide, insensitive to overdeveloping
Influence of oxygen	No (+)	Yes (–)
Liftoff	Yes [usually with multiple layer resist (MLR)]	Yes [even in single layer resist (SLR)]
Mask type	Dark-field: lower defects	Clear-field: higher defects
Opaque dirt on clear portion of mask	Not very sensitive to it	Causes printing of pinholes
Photospeed	Slower	Faster
Pinhole count	Higher	Lower
Pinholes in mask	Prints mask pinholes	Not so sensitive to mask pinholes
Plasma etch resistance	Not very good	Very good
Proximity effect	Prints isolated holes or trenches better	Prints isolated lines better
Residue after development	Mostly at <1 µm and high aspect ratio	Often a problem
Resolution	High	Low (>1 µm)
Sensitizer quantum yield Φ	0.2–0.3	0.5–1
Step coverage	Better	Lower
Strippers of resist over Oxide steps Metal steps	 Acid Simple solvents	 Acid Chlorinated solvent compounds
Swelling in developer	No	Yes
Thermal stability	Good	Fair
Wet chemical resistance	Fair	Excellent

*Newer resist systems are discussed under "Photolithography Resolution Enhancement Technology," this chapter.

because they require only small amounts of sensitizers and therefore are substantially less expensive than positive resists. Moreover, great progress has been made in improving the resolution of new types of water-soluble negative resists. These are used in new generations of ICs and in high-aspect-ratio miniaturized systems.[29,30] In working with different resists, it is also important to be aware of such properties as shelf life, flash point, and threshold limit value (TLV) rating. The flash point is the temperature at which the resist vapors ignite in the presence of an open flame. The TLV is the toxicity rating that specifies the maximum ambient concentration (in parts per million) to which a worker can be safely exposed during a normal workday.

Table 1.7 lists some common positive and negative resists used in various lithography strategies along with their lithographic sensitivities. For charged particles (e-beam lithography and ion-beam lithography), sensitivity is expressed in coulombs per centimeter square (C/cm^2); for photons (optical and x-ray), joules per centimeter square (J/cm^2) is used. Ideally, in charged-particle lithography, one should select a resist with sensitivity in the range of 10^{-5}–10^{-7} C/cm^2, and in photon lithography, 10–100 mJ/cm^2, to minimize the exposure duration.

Resist research and development are producing better resolution in both positive and negative resists. Results have been especially impressive in the latter. Negative resists used to be relegated to low-resolution chips and PWBs, but newer systems offer wider processing latitude and higher resolution. Progress has been so swift that new negative resists are now used in advanced complementary metal oxide silicon (CMOS) logic device manufacture. Several of these new negative resists also enable very high-aspect-ratio microfabrication.

Finally, in Table 1.8, we list the development of photoresists as they keep up with the scale of integration in ICs. In the inset of Table 1.8, we show the photoresist evolution with the type of lithography. In this inset, SCALPEL, ion-projection lithography, and x-ray lithographies are next generation lithographies (NGLs) that will be reviewed in Chapter 2; PSM stands for phase-shifting masks and OAI for off-axis illumination, two resolution-enhancing techniques (RETs) that are reviewed further below.

Permanent Resists and Dry Resists

Permanent Resists Resists typically are removed (stripped) once they have served their function as temporary stencils. Some negative resists, hardened through UV exposure, are used as permanent components of miniature devices. Two prominent examples in this category are polyimide and SU-8-based resists.

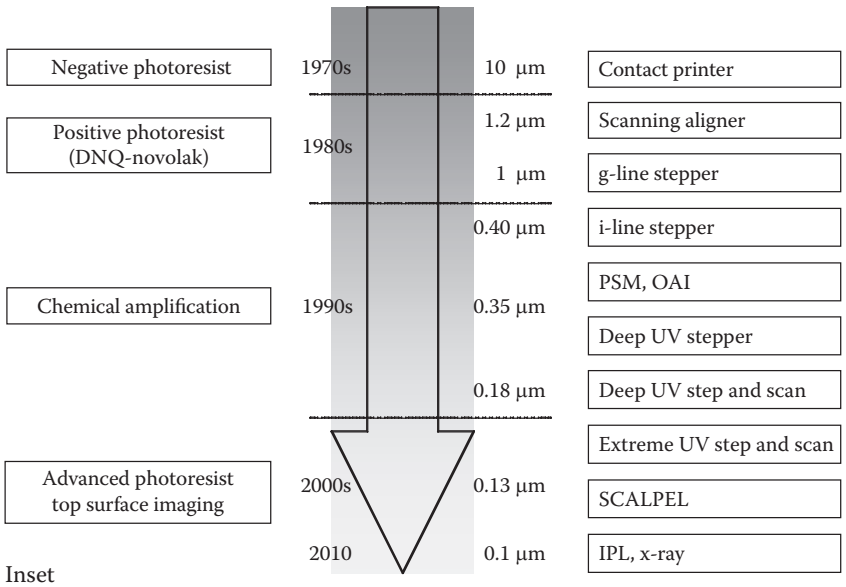
Polyimides are a class of polymers synthesized from two monomers: a dianhydride and a diamine.

Table 1.7 Typical Negative and Positive Photoresists and Their Lithographic Sensitivity

Class of Resist	Resist Name	Tone (Polarity)	Lithographic Sensitivity
Optical	CAMP-6 (OCG)	Positive	100 mJ/cm^2
	APEX-E (IBM and Shipley)	Positive	75 mJ/cm^2
	Kodak 747	Negative	9 mJ/cm^2
	AZ-1350J	Positive	90 mJ/cm^2
	PR102	Positive	140 mJ/cm^2
	XP-2198 (Shipley)	Positive	30 mJ/cm^2
	KRF (from UCB-JSR)	Negative	20–30 mJ/cm^2
e-Beam	COP [Copolymer-(α-cyano ethyl acrylate-α-amido ethyl acrylate)]	Negative	0.5 $\mu C/cm^2$
	GeSe (germanium selenide)	Negative	80 $\mu C/cm^2$
	PBS [poly-(butene-1-sulfone)]	Positive	1 $\mu C/cm^2$
	PMMA	Positive	100 $\mu C/cm^2$
X-ray	COP	Negative	100 mJ/cm^2
	DCOPA	Negative	14 mJ/cm^2
	PBS	Positive	170 mJ/cm^2
	PMMA	Positive	6500 mJ/cm^2

Table 1.8 Development of Photoresists during the Past 38 Years

Year	Scale of Integration	Minimum Line Width	Photoresist
1970	1 Kb	10 μm	Rubber-based negative
1973	4 Kb	8 μm	Rubber-based negative
1976	16 Kb	5 μm	Rubber-based negative
1979	64 Kb	2 μm	Rubber-based negative
1982	256 Kb	2 μm	Novolak-type positive
1985	1 Mb	1.2 μm	High-resolution positive photoresist of DQ-novolak type
1988	4 Mb	0.8 μm	High-resolution positive photoresist and multilayer process
1992	16 Mb	0.5 μm	Deep UV resist, multilayer process
1995	256 Mb	0.25 μm	I-line (356 nm) resist
1999	1 Gb	0.18 μm	KrF (248 nm) resist
2002	4 Gb	0.13 μm	ArF (248/193 nm) resist
2005	16 Gb	0.1 μm	F2 (193/157 nm) resist
2008	64 Gb	0.07 μm	Extreme UV (13 nm) resist



Commercial products are supplied as soluble polyamic acid (PAA) intermediates, which undergo a thermal imidization with the evolution of water to form an insoluble polyimide (Figure 1.26). Polyimide is used as an interlayer dielectric because of its desirable characteristics such as a high thermal stability with a glass transition temperature, T_g > 300°C, a low dielectric constant ε, good chemical stability, and ease of processing. However, its tendency to absorb moisture causes the polyimide layers to swell and their dielectric constants to increase significantly. Various polyimide analogs have been developed to avoid the latter undesirable property.[31] A common approach to reduce the dielectric constant of polyimides is the inclusion of organofluorine components, in the form of pendant perfluoroalkyl groups. The polyamic acid precursor for fluorinated polyimides is commonly based on hexafluorodianhydrideoxydianiline (HFDA-ODA). The inclusion of fluorinated monomers in the polyimide backbone reduces moisture absorption and the dielectric constant. Unfortunately, these modified polyimides generally show an increased susceptibility to chemical attack, making their use in multilayer fabrication questionable. However, newer formulations have improved the chemical resistance of polyimides and have demonstrated a unique wet etch capability for making vias with aspect ratios approaching 1.2:1. Polyimide can be "wet etched" using a strong alkaline solution

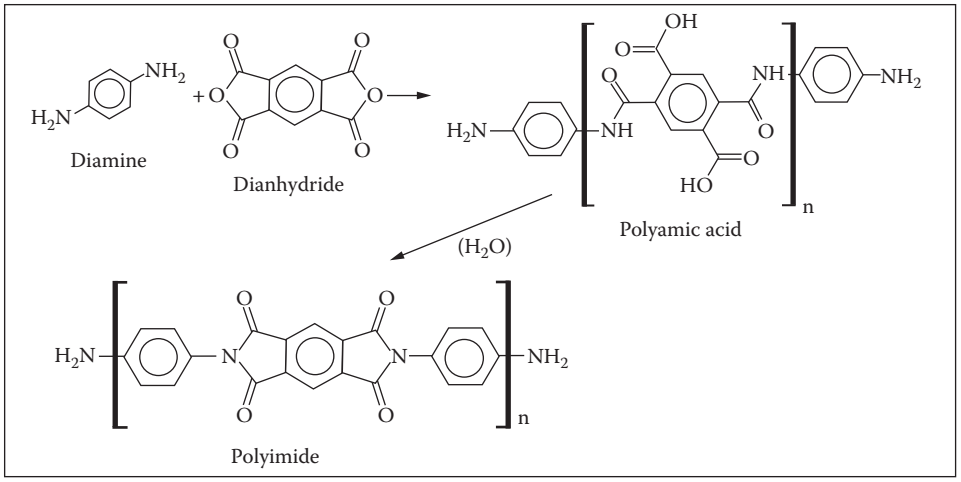


FIGURE 1.26 Typical imidization process used to form polyimides from polyamic acid.

such as TMAH. For dry etching of polyimide, one can use plasma-enhanced chemical vapor deposition (PECVD) silicon nitride as a mask with an O_2 plasma etch (possibly adding some CF_4). Tougher polyimide formulations are commercially available in polyamic acid form from Amoco under the trade name Ultradel. They have a slightly modified chemical structure based on hexafluorodianhydride-aminophenoxy-biphenyl (HFDA-APBP). The newest polyimides are extremely versatile materials and are used, for example, as stress-release layers in multichip modules, low dielectric insulation layers,[32] and even as optical waveguides based on fluorinated polyimides[33] (Figure 1.27; see also Chapter 3 on dry etching, Example 3.4). Ultradel 9000D is a soluble, preimidized, fluorinated polymer with properties optimized for integrated optical applications.

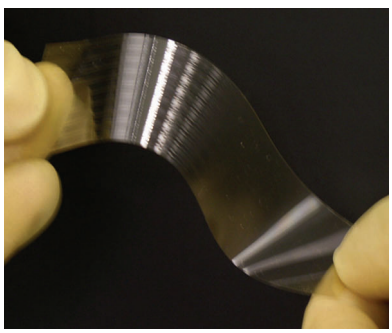


FIGURE 1.27 Flexible polyimide waveguide from Mitsui Chemicals. (Shioda, T. 2007. Next generation packaging technology which supports the ubiquitous devices. Application of the flexible optical waveguide to the consumer products. *Electr Parts Mater* 46:38–42.[35])

Typical absorption losses in the near-IR range from 0.2–0.4 dB/cm. In MEMS, polyimides have been used, for example, as flexible hinges in mechanical miniaturized structures (see Figure 7.86).[34]

Both nonphotosensitive polyimides and photosensitive polyimides (negative and positive) are available. Photosensitive polyimides have been processed in mass production lines using g-line (436 nm) and i-line (365 nm) exposure tools. Using photosensitive polyimides in manufacturing reduces the number of process steps dramatically compared with the use of nonphotosensitive polyimides. This is illustrated in Figure 1.28 for making polyimide vias in packaging. Positive tone polyimides, as used in Figure 1.28, are relatively new and were developed by HD Microsystems, a joint venture between Hitachi Chemical and Dupont (http://www.hdmicrosystems.com). The positive tone polyimides are more environmentally friendly than negative tone polyimides because of reduced organic solvents and associated volatile organic compounds. These positive tone resists also enable the process to use industry-standard TMAH developers. Polyimide films very strongly absorb UV light less than 350 nm. This absorbency is a result of the polymer's high aromaticity that is also responsible for the exceptional thermal properties of polyimides, which allows for processing at greater than 400°C. Polyimides in packaging and MEMS are further reviewed in Volume III, Chapter 4 on packaging.

Whereas in IC fabrication, sacrificial resist layers of 1 μm or less are common, in miniaturization

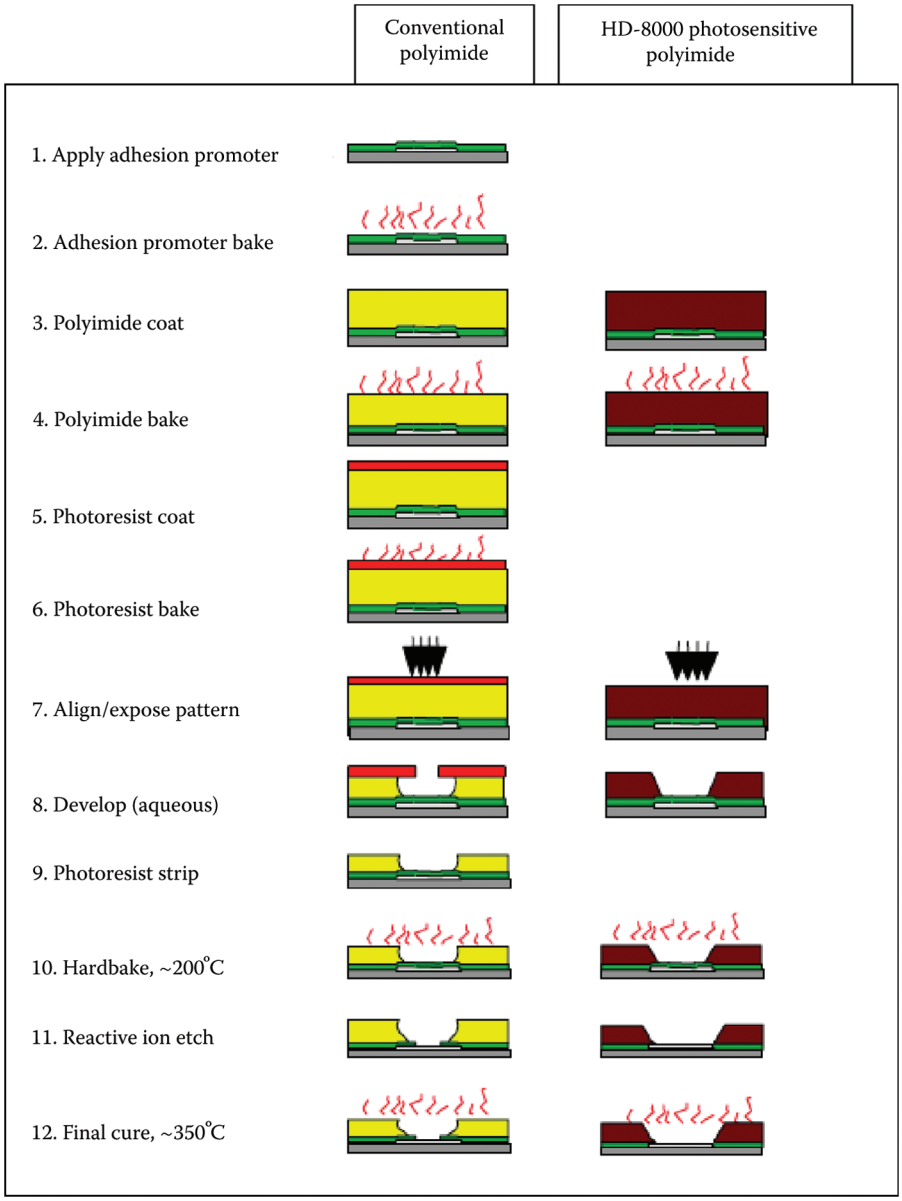


FIGURE 1.28 Comparison of conventional polyimide process and positive acting photosensitive polyimide process in packaging. The photosensitive polyimide process eliminates multiple steps and decreases cycle time. (Flack, W., S. Kulas, and C. Franklin. 2000. Process characterization of an aqueous developable photosensitive polyimide as a broadband stepper. SPIE. 3999–45.[36])

science, 3D structures frequently require thick permanent resist layers, capable of high resolution and high aspect ratios. Relatively new, chemically amplified deep UV resist systems, such as SU-8, can be applied in very thick layers (>100 μm) and have excellent sensitivity, high resolution, low optical absorption, high aspect ratio, and good thermal and chemical stability. These properties make SU-8 a good candidate for molds and other permanent MEMS structures. SU-8 is an acid-catalyzed negative photoresist, made by dissolving EPON®-SU-8 resin (registered trademark of Shell Chemical Company, http://www.shell.com) in an organic solvent such as cyclopentanone solvent or GBL (γ-butyrolactone) and adding a photoinitiator. The viscosity, and hence the range of thicknesses accessible, is determined by the ratio of solvent to resin. The EPON resist is a multifunctional, highly branched epoxy derivative that consists of bisphenol-A novolak glycidyl ether (Figure 1.29). On average, a single molecule contains eight epoxy groups, which explains the "8" in the name SU-8. The material has become a major

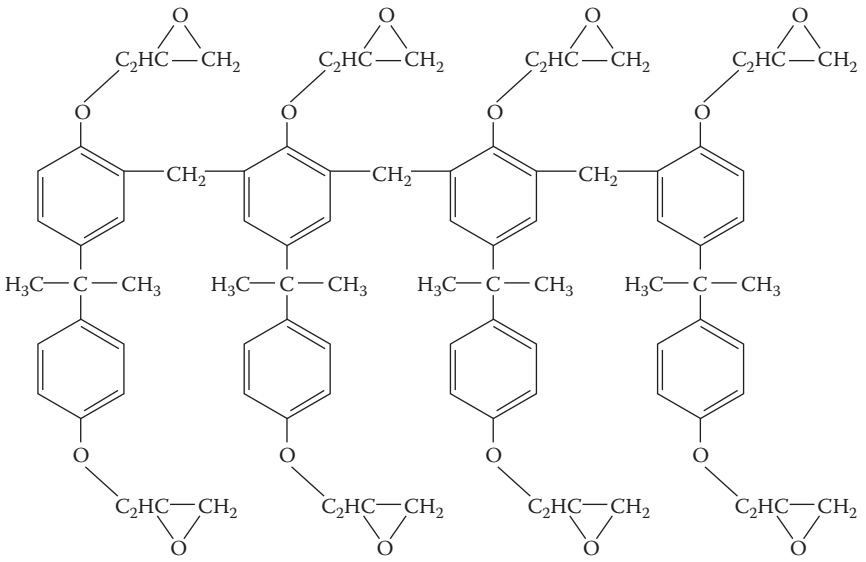


FIGURE 1.29 Glycidyl ether of bisphenol A: SU-8: cross-linking reaction of carboxy functional polymer with diepoxide.

workhorse in miniaturization science. In a chemically amplified resist like SU-8, one photon produces a photoproduct that in turn causes hundreds of reactions to change the solubility of the film. Because each photolytic reaction results in an "amplification" via catalysis, this concept is dubbed "chemical amplification".[37] We touch briefly on resist amplification here, in the context of SU-8 photoresist only, but will discuss the underlying principles in more detail in "Strategies for Improved Resolution through Improved Resist Performance" (this chapter).

Scientists at IBM discovered that certain photoinitiators, such as onium salts, polymerize low-cost epoxy resins such as EPON®-SU-8. An onium salt, on UV exposure, generates a strong Lewis acid and catalyzes the cationic polymerization of the resin. These salts are called photochemical acid generators (PAGs), and their action constitutes excellent dissolution inhibition of a phenolic resin. In contrast with conventional free-radical initiators, an onium salt cationic photoinitiator, for example, triphenylsulfonium hexafluoroantimonate, is oxygen insensitive and stable over a wide temperature range. On UV exposure of the resist, Lewis acids are released within the resist matrix, forming a latent image—a 3D distribution of the catalytic photoproduct—and image formation is realized after an activating postexposure baking (PEB), as shown in Figure 1.30. Not only UV light in the 365–436-nm range but also electrons and x-rays initiate a high level of cross-linking density, converting the SU-8 photoepoxy into a strong polymer with a T_g of more than 200°C (the T_g of the unexposed resist is 55°C).[38] During the PEB, the generated photoacid initiates the ring opening of the epoxy groups, and extensive cross-linking makes the SU-8 insoluble. An SU-8 soft bake is typically carried out at 95°C, and a hard bake at 200°C. Hot plating, on a level hot plate, is preferred over an oven heat treatment. For developing a 130-µm-thick SU-8 film, a 5-minute dip in undiluted propylene glycol methyl ether acetate, a rinse with isopropyl alcohol, and repeat of this procedure until all the unpolymerized material is dissolved works best.

SU-8 resist, patented by IBM in 1992, was originally developed for e-beam lithography[39] and became commercially available in 1996. Because of its aromatic functionality and highly cross-linked matrix, the SU-8 resist is thermally stable and chemically very inert. After a hard bake, it withstands nitric acid, acetone, and even NaOH at 90°C, and it is more resistant to prolonged plasma etching and better suited as a mold for electroplating than PMMA.* The low molecular weight [~7000 ± (1000)] and multifunctional nature of the epoxy gives it the high cross-linking propensity, which

* PMMA or poly(methyl 2-methylpropenoate) is the synthetic polymer of methyl methacrylate. This thermoplastic and transparent plastic is sold by the trade names Plexiglas, Limacryl, R-Cast, Perspex, Plazcryl, Acrylex, Acrylite, Acrylplast, Altuglas, Polycast, and Lucite and is commonly called acrylic glass or simply acrylic

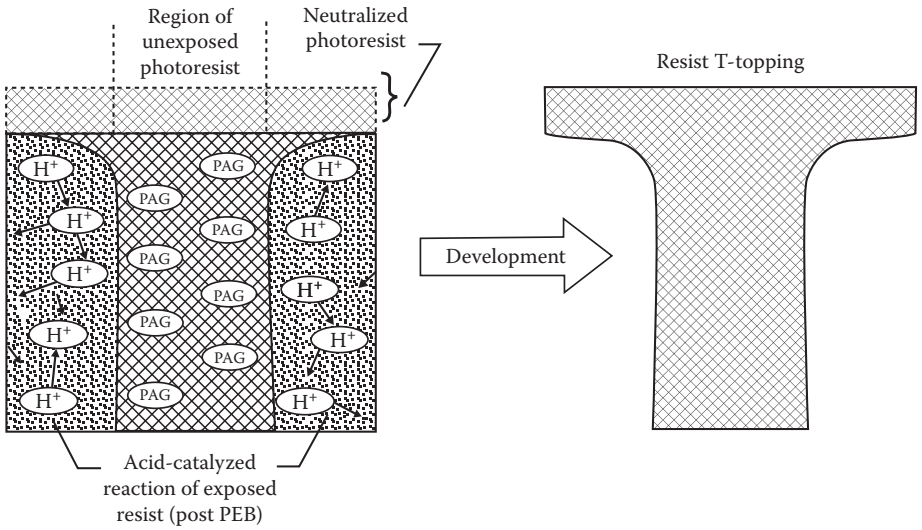


FIGURE 1.30 SU-8 is a negative-type amplified photoresist. Where light strikes the photochemical acid generators (PAGs), a Lewis acid is released within the resist matrix, forming a latent image—a 3D distribution of the catalytic photoproduct—and image formation is realized after an activating postexposure bake (PEB). After development a T-topped structure often results (see text).

also reduces the solvent-induced swelling typically associated with negative resists. As a result, very fine feature resolution, unprecedented for negative resists, has been obtained, and epoxy-based formulations are now used in high-resolution semiconductor devices, such as 0.35 μm CMOS logic (e.g., with CGR* chemically amplified negative resist). Low molecular weight characteristics also translate into high contrast and high solubility. Because of their high solubility (up to 85% solids by weight in solvents such as methyl-iso-butyl ketone), very concentrated resist casting formulations can be prepared. The increased concentration benefits thick film deposition (up to 500 μm in one coat) and planarization of extreme topographies. The high epoxy content promotes strong SU-8 adhesion to many types of substrates and makes the material highly sensitive to UV exposure. On the negative side, strong adhesion makes stripping of the exposed SU-8 material currently one of the most problematic aspects. Stripping of SU-8 may be carried out with hot 1-methyl-2-pyrrolidon, plasma, or laser ablation. There are other issues to be resolved with this resist; for example, thermal mismatch of SU-8 on a Si substrate (the thermal expansion coefficient α. for silicon is 2.361 ppm/K vs. 21–52 ppm/K for SU-8) produces stress and may cause film cracking. Moreover, the absorption spectrum of SU-8 shows much higher absorption coefficients at shorter wavelengths. As a result, lithography using a broadband light source tends to result in overexposure at the surface of the resist layer and underexposure at the bottom. The resulting developed photoresist tends to have a negative slope, which is not good for mold applications: the mold sidewall should have a positive or at least a vertical slope for easy release of the molded part from the mold. The exaggerated negative slope at the top of the resist structure surface is often called *T-topping* (see Figure 1.30). UV light shorter than 350 nm is strongly absorbed near the surface, creating locally more acid that diffuses sideways along the top surface. Selective filtration of the light source is often used to eliminate these undesirable shorter wavelengths (<350 nm) and thus obtain better lithography results. For example, Reznikova et al. used a 100-μm-thick SU-8 resist layer to filter exposure radiation at 334 nm,[40] and Lee et al.[41] reported using a Hoya UV-34 filter to eliminate the T-top (overexposed top part). Nearly vertical sidewalls can be achieved using a Hoya UV-34 filter. However, there might still be a beaked feature created at the very top of the resist features. This results from the diffraction at the interface between the

* CGR, chemically amplified resist based on poly(p-hydroxystyrene) or its copolymers with styrene or vinylcyclohexane with powderlink crosslinker and a photochemical acid generator.

mask and the photoresist, and it can be eliminated by filling this gap with index matching oil (see results by Wanjun Wan in Figure 1.31). Besides its sensitivity and relatively low cost (<$100/L), aspect ratios up to ~25 for lines and trenches have been demonstrated in SU-8-based contact lithography.[42] When patterned at 365 nm, the wavelength at which the photoresist is the most sensitive, total absorption of the incident light in SU-8 is reached at a depth of 2 mm. In principle, resist layers up to 2 mm thick can be structured.[43] Wan et al.[44] recently confirmed this astounding potential experimentally. Wan's group at Louisiana State University, using both wavelength optimization by patterning using a filtered i-line (365 nm) and air gap compensation (with glycerin or a Cargille refractive index matching fluid), demonstrated aspect ratios greater than 190 (for a feature with a 6-μm thickness and a height of 1150 μm) and structures as high as 2 mm. In the inset in the header of this chapter we reproduce one result of Wan et al., and more are displayed in Figure 1.31.

New resists like SU-8 have enabled miniaturization scientists to make LIGA (German acronym for *Lithographie*, *Galvano-formung*, *Abformung*)-like structures less expensively. LIGA is the ultimate lithography technique for making high-aspect-ratio features. In LIGA, a synchrotron x-ray source is used for exposure, and resist walls with the ultimate in wall straightness and wall smoothness are obtained (see Chapter 10). For many applications, photolithography with a much less expensive UV exposure system and a thick photoepoxy, such as SU-8, suffices. This alternative method to LIGA is sometimes referred to as *poor man's LIGA*. The example SU-8 application from Mimotec (http://www.mimotec.ch) shown in Figure 1.32 is a typical LIGA-like application, in

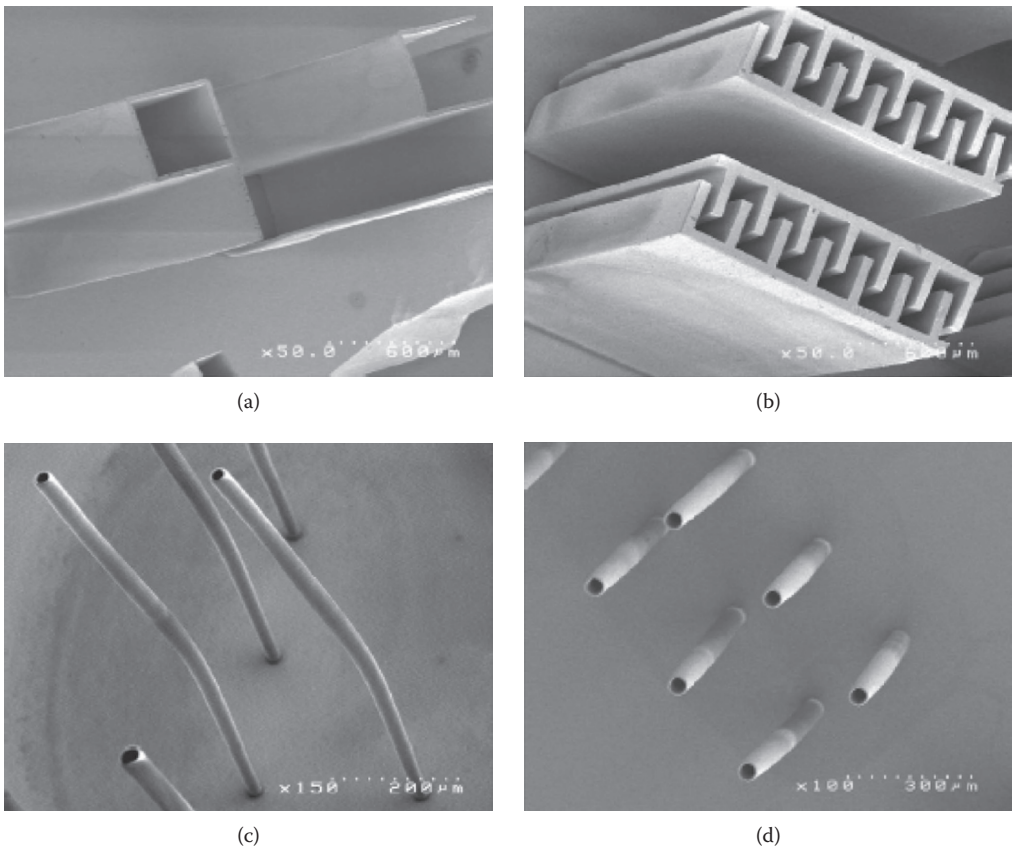


FIGURE 1.31 Ultrahigh-aspect-ratio structures obtained using a filtered light source (using a thick PMMA plate as filter) and with gap compensation (using an index matching solution). (a) Micro features with designed thickness of 7 μm and height of 2000 μm. (b) Micro comb structures with designed thickness of 40 μm and height of 2000 μm. (c) Micro cylinders 1150 μm tall and designed wall thickness of 6 μm and an internal diameter of 20 μm. As the thickness is reduced, the strength of the cylinders becomes too low to maintain their upright position. (d) Micro cylinders 1150 μm tall and designed wall thickness of 9 μm and internal diameter 27 μm.

FIGURE 1.32 Typical SU-8 application. The metal structure shown here has been made by plating in an SU-8 mold. (http://www.mimotec.ch). 1) Rake: cogwheel actuation; 2) cam reader; 3) microinscriptions; and 4) rest position.

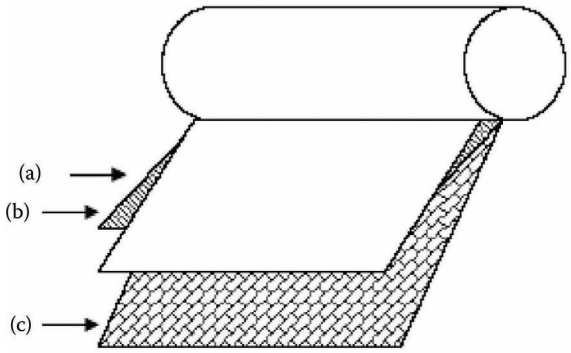
which a sacrificial cavity in SU-8 resist has been electroplated to make a metal positive replicate of the cavity. The advantages of the poor man's LIGA method are that it is inexpensive and can be quickly brought into play, all of which makes the process very competitive compared with LIGA.

To find the latest updates on SU-8 applications and processing tips, visit http://aveclafaux.freeservers.com/SU-8.html. This site also lists all the relevant physical properties of SU-8 accompanied with literature references. Another resource with plenty of SU-8 information is the MEMScyclopedia at http://memscyclopedia.org/su8.html.

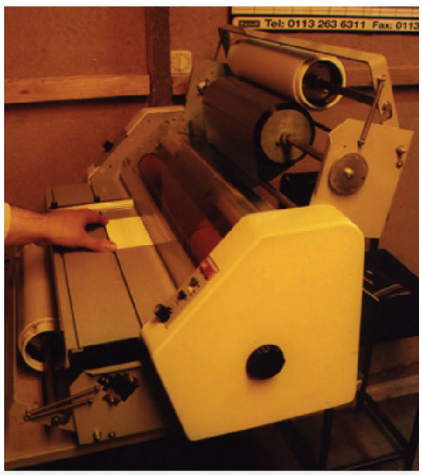
The first two companies that bought an SU-8 sales license from IBM were MicroChem Corporation[45] and Sotec Microsystems.[46] The latter company, now defunct, developed conductive and magnetic SU-8 and used dry SU-8 films to laminate over large cavities and channels[42]. MicroChem (previously named Microlithography Chemical Corp.) is the oldest manufacturer of the SU-8 photoresist, licensed by IBM. SU-8 Gerstel Engineering Solutions, Ltd. (http://www.gersteltec.ch) not only sells SU-8 under the name SU-8 GM-10## with different viscosities (e.g., GM-1040, GM-1060, and GM-1070) but also offers prototyping and mass production of SU-8 microsystems.

Dry Resist Most resists in IC and MEMS fabrication are deposited as liquids, whereas resists used in printed wiring board (PWB) manufacture are usually dry film resists that come in rolls (ranging from 2–60 in. wide and 125–1000 ft. long) and are laminated onto the substrate instead of being spin coated onto it (Figure 1.33). Dry film resist formulations are sandwiched between a polyolefin release sheet and a polyester base, rolled up on a support core (Figure 1.33A and B). These protective covers shield the film from environmental oxygen and facilitate handling. The dry film resist layer comes in thicknesses starting from 25 µm and goes up to 100 µm. Resist thicknesses of 1–1.5 mils are common for imaging purposes, and thicker resists (1.5–2.0 mils) are used as plating, rather than etch, resists. Dry film resists offer advantages such as excellent adhesion on most substrates, no liquid handling because there is no solvent, high process speed, excellent thickness uniformity over a whole substrate (even substrates with holes), simple handling, no formation of edge beads, low exposure energy, low cost, short processing time, and near vertical sidewalls. The photoresist film is applied by removing the polyolefin sheet before lamination of the resist onto the substrate. The polyester top cover sheet remains on top of the photo polymer until later in the process. This coversheet protects the resist from scratches, keeps contaminants from the surface, and prevents the exposure tool from touching and perhaps sticking to the resist. Conformation of the resist to the substrate is achieved by heating under pressure in a hot-roll or cut-sheet laminator (Figure 1.33C). The heat and pressure of the laminating rollers cause the dry film to soften and adapt to surface topologies. The resist is then exposed to a UV light source. Modern dry film resists are developed in a simple sodium carbonate solution (1–2%) after removal of the top cover layer. The biggest disadvantage of dry resist is its relatively low resolution compared with liquid resists as illustrated in Figure 1.34. Two major reasons for this poorer resolution are the many-times thicker resist coating and the fact that the mask is positioned on top of a thick protective Mylar cover

(A)



(B)



(C)

FIGURE 1.33 (A) Dry film resist roll. (B) Dry film photoresist composition; three-layer structure, (a) polyethylene separation sheet, (b) photoresist, and (c) polyester support. (C) Dry lamination equipment.

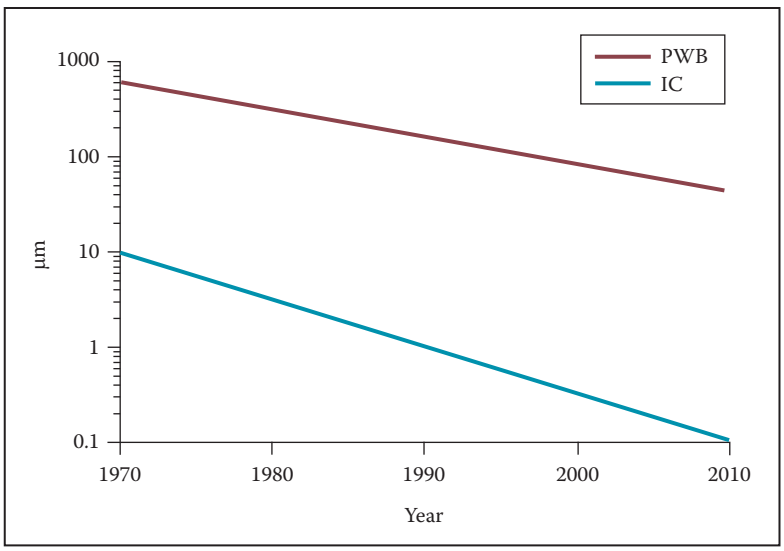


FIGURE 1.34 Comparison of resolution obtained with liquid resist compared with dry resists as a function of time. (Drawn by Mary Amasia from UC Irvine.)

film. Under highly controlled conditions, one can achieve an opening (space) of 25 µm with a 25-µm-thick dry film, but under production conditions, it is rather difficult to achieve such resolutions. By removing the top cover sheet from the photoresist before exposure, higher resolutions are possible. The removal of 25 µm of separation between the photoresist and the exposure tool decreases the chances of light diffusion. Also, foreign particles that collect on the coversheet are removed along with the coversheet.

There are a variety of dry film photoresists widely used and commercially available; examples include Riston®, Ordyl BF 410, Etertec® 5600, DF 4615, and DFR-15. All are used in the manufacture of circuit boards and can be made thick; additionally, all are candidates for broad use in MEMS as well.

Today it is difficult to obtain high-aspect-ratio devices with these resists because of their resolution limitation, high optical absorbency, and the difficulty in obtaining layers thicker than 50 µm in a single coat.[29] However, for many applications, the poorer resolution of dry resists (25–75 µm) is no obstacle for their usage in mesoscale devices (100s of micrometer to millimeter scale).

The potential benefits of using dry resist films as permanent components in the mass production of biosensors and microfluidics were recognized and described, for example, by Madou et al.[47] These authors suggest that continuous, web-based manufacturing may finally make ubiquitous, disposable miniaturized devices such as biosensors and microfluidics possible (see also Volume III, Chapter 1, Example 1.1). Dry resist film materials are less expensive than silicon and form a convenient substrate; they come in rolls so that large sheets can be processed. In the future, this author envisions a continuous lithographic process, including exposure and development, taking place between a dry resist supply roll and a pick-up reel. This would be especially attractive if higher resolution dry resist could be formulated. Dry film SU-8, as Louis Guerin suggested many years ago, holds the promise of continuous high-aspect-ratio, high-resolution lithography (http://www.geocities.com/guerinlj).

At MicroChem (http://www.microchem.com/about/index.htm), progress is being made making SU-8 dry resist films, and Dupont already offers the MX series dry film for MEMS applications. The new MX products feature an aspect ratio of 3:1. At a thickness of 50 µm, resolution is 16–17 µm, which is better than any other dry film product. Additionally, these films can be multilaminated to build up the thickness (http://tyvek.com.mx/APL/en_US/products/mx/index.html).

Nonspin liquid resist-coating methods such as roller coating or spraying (see above) also will play an important role in MEMS manufacturing of the future. Large substrate sheets can be accommodated; the throughput is higher than with dry resists; and the resolution is considerably better.

Glass Transition Temperature of a Resist (T_g)

Resists must meet several rigorous requirements: good adhesion, high sensitivity, high contrast, good etching resistance (wet or dry etching), good resolution, easy processing, high purity, long shelf life, minimal solvent use, low cost, and a high glass transition temperature, T_g. Most resists are amorphous polymers that exhibit viscous flow with considerable molecular motion of the polymer chain segments at temperatures above the glass transition. At temperatures below T_g, the motion of the segments is halted, and the polymer behaves as a glass rather than a rubber. If the T_g of a polymer is at or below room temperature, the polymer is considered a rubber; if it is above room temperature, it is considered to be a glass. Because above T_g the polymer flows easily, heating the resist film above its T_g for a reasonable amount of time enables the film to anneal into its most stable energetic state. In the rubber state, it is easy to remove the solvent from the polymer matrix; that is, soft bake the resist. Extreme attention needs to be given to the cleanliness of the working environment with the resist in this state. When softening the resist at or above T_g, it may be easier to remove solvent, but the resist tends to pick up impurities. The importance of resist reflow, as we will learn later, also lies in planarizing of undesirable topography. The glass transition temperature of a resist is influenced by baking and exposure; the T_g for SU-8, for example, is 55°C for the unexposed material but 200°C after exposure and hard baking. In the fabrication of carbon-MEMS structures from photoresist

by pyrolysis (Volume III, Chapter 5), we make sure that the temperature always stays below the T_g to avoid collapse of the resist features.

In general, polymers that crystallize are not useful as resists because the formation of crystalline segments prevents the formation of uniform high-resolution isotropic films.[19]

Wafer Priming

Resists, especially positive resists, do not adhere well to a Si wafer. The native silicon dioxide on a Si surface, typically 20–50 Å thick, forms long-range hydrogen bonds with water adsorbed from the air. When resist is spun onto such a surface, it adheres to the water molecules rather than to the surface, resulting in poor adhesion. This effect is more pronounced when the humidity is high or if the wafer has been previously immersed in water. Good humidity control (at ~40% relative humidity) and annealing are required to remove surface water and prepare a silicon wafer for resist coating. In the early 1960s, photoresist was spun directly onto the wafer surface. This worked for the large geometry circuits at the time, but when geometries reached the 20-μm range, further feature decrease became intimately linked to the quality of the resist adhesion. To faithfully reproduce the smallest features of a circuit design, a better photoresist/Si wafer contact had to be devised. To this end, surface priming with reactive silicone primers was introduced. A typical resist adhesion promoting primer is 1,1,1,3,3,3-hexamethyldisilazane (abbreviated HMDS) [formula: $(CH_3)_3SiNHSi(CH_3)_3$], developed and patented by IBM. Other primers include trichlorophenylsilane (TCPS) [formula: $C_6H_5SiCl_3$] and bistrimethylsilylacetamide (BSA) [formula: $(CH_3)_3SiNCH_3COSi(CH_3)_3$]. Reactive Si–NH–Si functional groups in HMDS react with the oxide surface in a process known as *silylation*, and a strong siloxane bond to the surface (Si-O-Si) is created (Figure 1.35). At room temperature, unhydrolyzed silane does not react with surface silanols [$Si(s)–OH + (CH_3)_3SiOCH_3 \rightarrow$ no reaction], but with some surface-adsorbed water present, a small amount of reaction occurs as a result of methoxysilane hydrolysis, followed by reaction with surface silanols [$Si(s)–OH$ + Surface $H_2O + (CH_3)_3SiOCH_3 \rightarrow (CH_3)_3SiOH + Si(s)–OH$ and $(CH_3)_3SiOH + Si(s)–OH \rightarrow Si(s)–O–Si(CH_3)_3 + H_2O$]. When silylation is carried out with HMDS, a strong catalytic effect is observed. The amine adsorbs onto the relatively acidic surface silanols, rendering the oxygen attached to the surface silicon atom highly

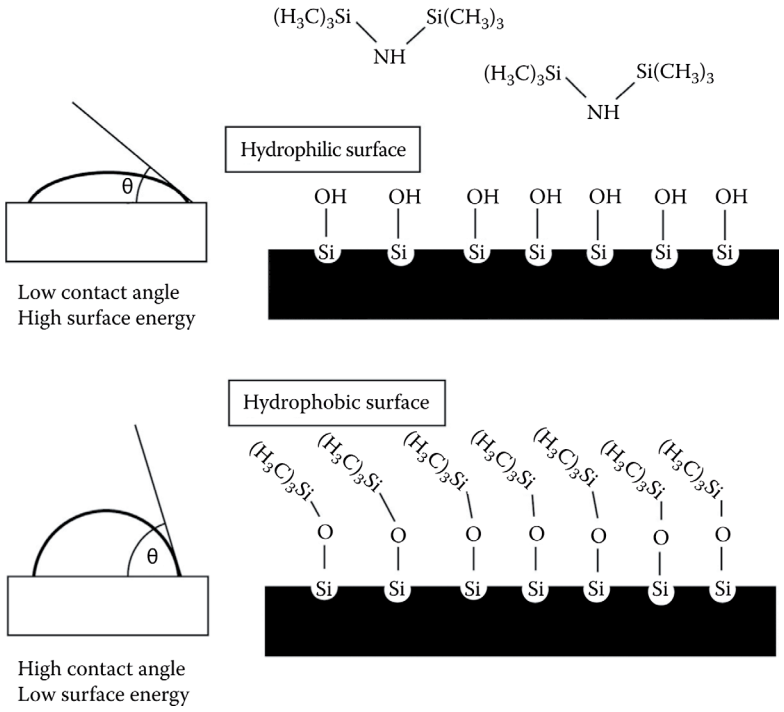


FIGURE 1.35 HMDS reaction mechanism monitored by the contact angle θ of a drop of water on the treated surface.

nucleophilic. This oxygen then attacks the relatively electrophilic silane silicon, resulting in surface silylation. The more hydrophobic the wafer surface, the better the resist adhesion, and the methyls, it is assumed, bond/adhere to the photoresist, enhancing the photoresist adhesion. The process works not only on silicon dioxide but also on other oxides (e.g., Al_2O_3). It should be noted that HMDS is extremely flammable and a suspected carcinogen; it should be handled with care. A dehydration bake of the SiO_2 surface at 200–250°C for 30 minutes (with optional vacuum) removes adsorbed water from the silanol groups at the silicon surface, which then can react with the amino groups of the HMDS vapor. The primer may then be applied by dipping the wafer in a 1–10% HMDS solution in xylene or by spin coating it onto the wafer. Baking the wafers to dehydrate and then removing them from the oven and transporting them to a dip solution or a spinner poses a serious problem, however, because on removal from the bake oven, the surface may rehydrolyze. Further shortcomings of this approach include a tendency to generate and deposit particles and HMDS wastage. Today priming of wafers is performed in a process that combines a bake and prime in the same vacuum. The wafer is exposed to HMDS vapor in a stream of dry nitrogen in a vacuum chamber while heating the wafer for approximately 40–60 seconds. This method provides the best adhesion and constitutes the best overall process as it remedies several problems: 1) no rehydrolyzation can occur because the unprimed wafers are not exposed again to the atmosphere before priming; 2) the same vacuum that is used during the bake cycle is used to draw monolayer amounts of HMDS; and 3) the interior of the chamber may be evacuated several times (with intermediary nitrogen backfills) before exposure to HMDS. Thus, at the time of HMDS exposure, the amount of oxygen present in the chamber (and hence the chance of a flammable mix forming) is very low. After treating a Si wafer with HMDS, a contact angle test with a drop of water reveals the surface to be much more hydrophobic (see Figure 1.35).

The chemical modification of oxide surfaces using organofunctional silanes is important not only in affecting better resist adhesion but also for the fabrication of water-repellent and antifouling coatings, DNA arrays, protein arrays, enzyme-based sensors, and so on.

Sputtering of the surface (Chapter 3), for some applications, presents an attractive alternative to vapor priming; the microroughness at the surface induced by sputtering provides for mechanical adhesion of the resist to the substrate. Primers for GaAs include monazoline C, trichlorobenzene, and xylene.

Resist Stripping

We now turn to the last step of the photolithographic process: photoresist stripping, as illustrated in Figure 1.1f.

Wet Stripping

Photoresist stripping, in slightly oversimplified terms, is organic polymer etching. The primary consideration is complete removal of the photoresist without damaging the device under construction. Turning back to Figure 1.1, once the exposed oxide has been etched away in (e), the remaining photoresist can be stripped off with a strong acid such as H_2SO_4 or an acid-oxidant combination such as H_2SO_4–Cr_2O_3 attacking the photoresist but not the oxide or the silicon (Figure 1.1f). Other liquid strippers include organic solvent strippers and alkaline strippers (with or without oxidants). Acetone can be used if the postbake is not too long or occurs at a low enough temperature. With a postbake of 20 minutes at 120°C, acetone is still fine. But with a postbake at 140°C, the resist develops a tough "skin" and has to be burned away in oxygen plasma. There are many commercial strippers available; some are specific for positive resists (e.g., ACT-690C from Ashland), others for negative resists, and still others are universal strippers (e.g., ACT-140 from Ashland). Other popular commercial strippers are Piranha and RCA clean. The RCA1 clean for organics was described above. To make up a Piranha solution, measure 5 parts of H_2SO_4 in a Pyrex beaker, and very slowly add 1 part of H_2O_2. Note that this mixture is exothermic. When cool, it may be refreshed by very slowly adding more H_2O_2.

The ozone-water method is a Piranha alternative. In this process, water of 25–95°C, depending on the application, is sprayed onto the substrate

as it is rotated. At the same time, dry ozone gas is injected into the reaction chamber. The ozone diffuses through the thin boundary layer of water to react with the organics on the substrate. Strip rates exceed 1.2 μm/min.[48]

The oxidized Si wafer with the etched windows in the oxide (Figure 1.1f) now awaits further processing. This might entail a wet anisotropic etch of the Si in the oxide windows with SiO_2 as the etch mask.

Dry Stripping

Dry stripping or oxygen plasma stripping, also known as ashing, has become more popular as it poses fewer disposal problems with toxic, flammable, and dangerous chemicals. Wet stripping solutions lose potency in use, causing stripping rates to change with time. Accumulated contamination in solutions can be a source of particles, and liquid phase surface tension and mass transport tend to make photoresist removal difficult and uneven. Dry stripping is more controllable than liquid stripping, less corrosive with respect to metal features on the wafer, and, more importantly, it leaves a cleaner surface under the right conditions. Finally, it does not cause the undercutting and broadening of photoresist features that can be caused by wet strippers.

In solid-gas resist stripping, a volatile product forms either through reactive plasma stripping (e.g., with oxygen), gaseous chemical reactants (e.g., ozone), and radiation (UV), or a combination thereof (e.g., UV/ozone-assisted). Plasma stripping uses a low-pressure electrical discharge to split molecular oxygen (O_2) into its more reactive atomic form (O). This atomic oxygen converts an organic photoresist into a gaseous product that may be pumped away. This type of plasma stripping belongs in the category of chemical dry stripping and is isotropic in nature (see Chapter 3). In ozone strippers, ozone, at atmospheric pressure, attacks the resist. In UV/ozone stripping, UV helps to break bonds in the resist, paving the way for a more efficient attack by the ozone. Ozone strippers have the advantage that no plasma damage can occur on the devices in the process. Reactive plasma stripping currently is the predominant commercial technology because of its high removal rate and throughput. Some different stripper configurations are barrel reactors, downstream strippers, and parallel plate systems. These prevalent stripping systems are reproduced in Figure 1.36.[49] More details will be provided in Chapter 3.

Posing new challenges to modern stripping processes are shrinking feature sizes, increasing aspect ratios, and stripping of resist over some of the newest dielectrics ($\sigma_r < 3$) and copper interconnects. Especially novel dielectric materials, such as fluorinated silicate glass, carbon-doped oxides, and fluorinated amorphous carbon, are easily attacked or degraded by stripping.[48] In practice, combinations of dry and wet etching often form the most successful strategy.

Dry stripping of resist has been so successful that it has accelerated the use of plasmas in other lithography steps such as development and deposition of resist. For example, it was the study of the attack of oxygen on polymers that led to the development of dry developed resists as used in the DESIRE process (see page 75). Such dry developed resists are vital for future submicrometer lithography where

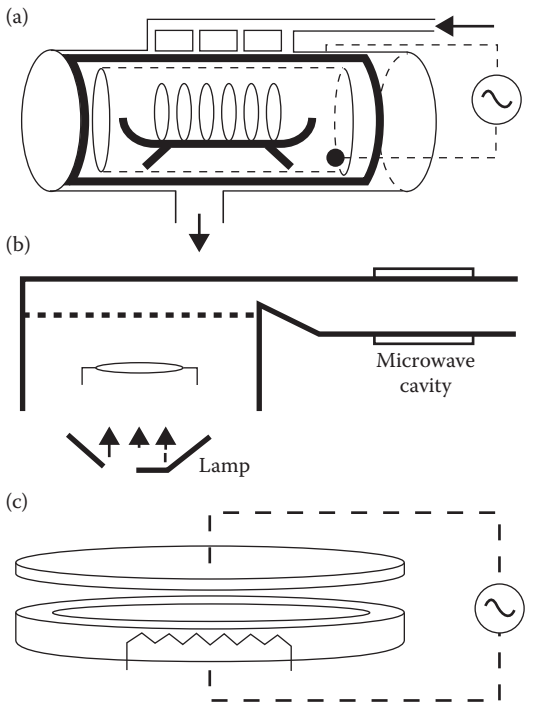


FIGURE 1.36 Various dry stripping reactors: barrel reactor (a), downstream etchers (b), parallel plate systems (c). (From Flamm, D.L. 1992. Dry plasma resist stripping. Part I: overview of equipment. *Solid State Technol* 35: 37–39. With permission.[49])

underetching and broadening of features are most critical. Although not detailed in the current short treatise, inspection and metrology techniques (see Volume III, Chapter 6) play a crucial role at various points in the lithography process.

Critical Dimension, Overall Resolution, Line Width Metrology

The absolute size of a minimum feature in an IC or a miniature device, whether it involves a line width, spacing, or contact dimension, is called the *critical dimension* (CD). The overall resolution of a process describes the consistent ability to print a minimum size image, a CD, under conditions of reasonable manufacturing variation.[50] Many aspects of the process, including hardware, materials, and processing considerations, can limit the resolution of lithography. Hardware limitations include diffraction of light or scattering of charged particles (in the case of charged-particle lithography or hard x-rays), lens aberrations, mechanical stability of the system, and so on. The resist material properties that impact resolution are contrast, swelling behavior, thermal flow, and chemical etch resistance. The most important process-related resist variables include swelling (during development) and stability (during etching and baking steps). Resolution frequently is measured by line-width measurements using either transmitted or reflected light or other metrology techniques. Optical techniques perform satisfactorily for features of 1 μm and larger, providing a precision of ±0.1 μm (at 2 σ, i.e., all data points within plus and minus two standard deviations). By 1998, devices with features as small as 0.25 μm were launched and necessitated equipment requirements for line-width measurement with a precision of at least ± 0.02 μm (at 3 σ). Scanning electron microscopes (SEMs) or atomic force microscopes have come forward as the methods to reach these goals. A line width, L, is defined as the horizontal distance between the two resist-air boundaries in a given cross-section of the line at a specified height above the resist/substrate interface. Because different measurements may measure the line width of the same line at different heights of the cross-section, the measuring technology used always needs to be identified. The successful performance of devices depends on the control of the size of critical structures across the entire wafer and from one wafer to another, referred to as line-width control. A rule of thumb is that the dimensions must be controlled to tolerances of at least ± 1/5 of the minimum feature size. Typically, a series of features with known sizes across a substrate is measured and then plotted as a function of position on the wafer. The standard deviation at the 1 or 2 σ level is adopted as the line-width control capability of the particular exposure/resist technology. Plotting these data as a function of time enables line managers to maintain optimum performance on a manufacturing line.[50]

In Volume III, Chapter 6, the metrology tools used in IC, MEMS, and NEMs are discussed: laser interferometers for accurate alignment of mask to wafer (with a 5-nm "resolution"), film thickness measurement with profilometers, SEM, scanning transmission electron microscope (STEM), scanning probe microscope (SPM) with vertical resolution of 1 Å, and lateral resolution depending on tip sharpness.

Lithographic Sensitivity and Intrinsic Resist Sensitivity (Photochemical Quantum Efficiency)

Lithographic Sensitivity

A distinction must be made between the intrinsic sensitivity of a resist (that is, the resist's response to radiation) and the lithographic sensitivity when defining the measurement of the efficiency that translates resist exposure into a sharp image. In the literature, the values given for the lithographic sensitivity of a resist show a tremendous spread as a result of the complex relationship between the intrinsic resist sensitivity and the dose required to successfully process that resist. This relationship involves the intrinsic resist sensitivity and the bandwidth of the optical exposure system, baking conditions, resist thickness, developer composition, and development conditions. To reproduce a reported lithographic sensitivity, all these parameters need to be duplicated exactly. The best way to determine lithographic sensitivity is experimentally, as we will explain next.

Intrinsic Sensitivity of a Resist (Photochemical Quantum Efficiency)

A first indication of the intrinsic sensitivity of a resist to a certain wavelength can be deduced from the spectral-response curve of the resist. If the resist absorbs strongly in ranges where the radiation source shows strong emission lines, relatively short exposure times can be expected. Practical limits confine resist sensitivity: too sensitive a resist might mean an unacceptably short shelf life, and yet clearly the resist should be insensitive to the yellow and green light of the clean room.

High intrinsic resist sensitivity is a sought-after characteristic. To increase resolution of photolithography, the shortest possible wavelengths must be used. At those wavelengths exposure sources become less bright and optics absorb more. Because the total energy incident on a resist is a function of light source intensity, time, and absorption efficiency of the exposure optics, a decrease in intensity and an increase in light absorption require compensation through longer exposure times. This results in a smaller throughput of wafers per hour; conversely, a more sensitive resist decreases the exposure time, resulting in a higher throughput.

The intrinsic sensitivity or photochemical quantum efficiency, Φ, of a resist is defined as the number of photoinduced events divided by the number of photons required to accomplish that number of events:

$$\Phi = \frac{\text{Number of photon-induced events}}{\text{Number of photons absorbed}} \qquad (1.4)$$

For polymer resins where the polymer undergoes scission or cross-linking without the need for light-absorbing sensitizers (one-component resists), a *G*-value is introduced. The *G*-value corresponds to the number of scissions or cross-links produced per 100 eV of absorbed energy. For scission reactions, the symbol $G(s)$ is used; for cross-linking, the symbol $G(x)$ is used. In contrast to lithographic sensitivity, the measurement of intrinsic radiation sensitivity as expressed through Φ, $G(s)$, or $G(x)$ is reliable, and values from different sources agree relatively well.

The experimental determination of the quantum efficiency of a one-component resist is a complex undertaking. Samples of the polymer must be exposed to a known dose of gamma radiation, and the molecular weight of the irradiated samples must be measured either by membrane osmometry or gel permeation chromatography. Quantitative analysis of the molecular weight versus dose in polymers that undergo scission or cross-linking leads to an important relationship for a better understanding of resist exposure. We will use such a relationship when exploring x-ray lithography for the creation of "high-rise" PMMA resist structures (»10 µm high; see Chapter 10 on LIGA). Let us consider a positive resist here. For a positive resist sample of weight w (in grams) containing N_0 molecules before exposure, the definition of average molecular weight M_n^0 is given by[28,51]:

$$M_n^0 = \frac{wN_A}{N_0} \qquad (1.5)$$

where N_A is Avogadro's number. Expressing the dose, D, in eV/g, the total number of scissions produced in the sample, N^*, is proportional to the absorbed dose, or:

$$N^* = KDw \qquad (1.6)$$

where K is a constant dependent on the polymer structure, generally expressed in terms of a *G*-value. $G(s)$ (for positive resists) and $G(x)$ (for negative resists), like Φ, are figures of merit used to compare one resist material with another. With K expressed in terms of $G(s)$, Equation 1.6 can be rewritten as:

$$N^* = \left[\frac{G(s)}{100}\right]Dw \qquad (1.7)$$

in which we divide by 100 to express the number of events per 100 eV. As scission occurs, the number of molecules increases, and the new average molecular weight after exposure to dose D is then given by:

$$M_n' = \frac{wN_A}{N_0 + N^*} \qquad (1.8)$$

where the total mass of the polymer is assumed to remain constant during exposure.

By substituting Equations 1.5 and 1.7 into Equation 1.8, we obtain:

$$M_n' = \frac{N_A}{\frac{N_A}{M_n^0} + \left[\frac{G(s)}{100}\right]D} \tag{1.9}$$

which is independent of the sample mass *w*. Rearranging Equation 1.9 we obtain:

$$\frac{1}{M_n'} = \frac{1}{M_n^0} + \left[\frac{G(s)}{100}\right]D \tag{1.10}$$

From Equation 1.10, we conclude that a linear relationship exists between the inverse of the average molecular weight and the exposure dose *D*. The intercept on the *y*-axis yields $1/M_n^0$, and the slope allows one to calculate $G(s)$. There is a very high correlation between $G(s)$ values for gamma radiation [the radiation commonly used to determine $G(s)$] and sensitivity for electrons, ions, and x-rays.

The $G(s)$ of polymers commonly used as one-component, positive resist systems ranges from 1.3 for some PMMAs to approximately 10 for certain poly(olefin sulfones). A PMMA with a $G(s)$ value of 1 has a corresponding photochemical quantum yield for scission, Φ (Equation 1.4), of 0.02.[28] For some polymers, both scissioning and cross-linking events occur simultaneously on exposure. It is possible, even in the latter case, to uniquely determine both scission efficiency $G(s)$ and cross-linking efficiency $G(x)$.[51]

For one-component negative resists the figure of merit for intrinsic sensitivity, $G(x)$, expressed as number of cross-links per 100 eV absorbed dose, ranges from 0.1 for poly(ethylene) to approximately 10 for polymers containing oxirane groups (epoxy groups) in their side chains.

For a two-component positive system such as DNQ, Φ in Equation 1.4 corresponds to the quantum efficiency, that is, the number of sensitizer molecules converted to photoproduct, divided by the number of absorbed photons required to accomplish that conversion. Quantum efficiency in this case can easily be measured by using a narrow-bandwidth radiation source and a UV-Vis spectrophotometer. The quantum efficiency, Φ, of typical DNQ sensitizers ranges from 0.2–0.3 (compared with 0.02 for PMMA). Because of the high opacity (i.e., high nonbleachable absorption) of novolak resins in the deep UV region (200–300 nm) region, other resists like PMMA are used for shorter wavelength exposures (e.g., for deep UV, e-beam, and x-ray lithographies). The quantum efficiency of the bis(aryl) azide sensitizers in negative resist systems ranges from 0.5–1, making negative resists more sensitive than positive resists.

Resist Profiles

Overview of Profile Types

Manipulation of resist profiles is one of the most important concerns of a lithography engineer. Depending on the final objective, one of the three resist profiles shown in Figure 1.37 is attempted. A re-entrant, undercut, or a reverse resist profile (resist sidewall >90°) is required for metal liftoff (see below for more details). Some authors confusingly call slopes >90° *overcut* [28]; most, including this author, refer to this type of resist profile as an *undercut* (Figure 1.37a). Shallow resist angles (<90°) enable continuous deposition of thin

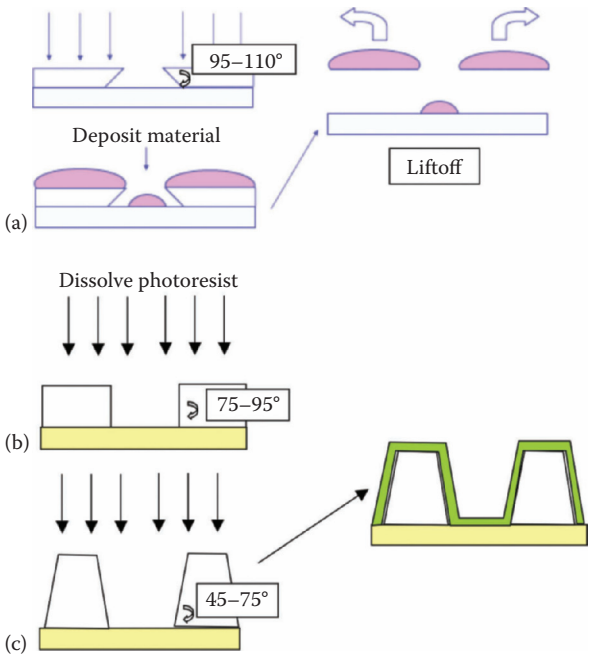
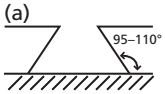
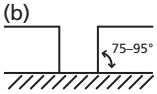
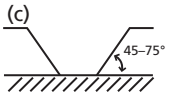
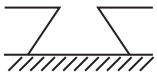


FIGURE 1.37 The three important resist profiles. (a) Re-entrant, undercut, or a reverse resist profile (resist sidewall >90°) is required for metal liftoff. (b) A vertical (75–90° resist sidewall angle) slope is desirable for a perfect fidelity transfer of the image on the mask to the resist. (c) Shallow resist angles (45–75° < 90°) enable continuous deposition of thin films over the resist sidewalls.

films over the resist sidewalls (Figure 1.37c). A vertical (90° resist sidewall angle) slope is desirable for a perfect fidelity transfer of the image on the mask to the resist (Figure 1.37b). If resist sidewalls are perpendicular, or if the resist layer is undercut, a deposited metal is likely to be noncontinuous: there is a separation between the desired and the undesired metal for a liftoff process. In the case of a shallow slope, the deposit is continuous.

In general, after development of negative and positive resists, three different photoresist wall profiles may be obtained as summarized in Figure 1.37 and tabulated in Table 1.9. In the table, R is the development rate of the exposed region, R_0 is the development rate of the unexposed region, and γ is the resist contrast. The resist contrast will be explained in greater detail later in this chapter. This table also lists typical applications for each resist profile. First, we will concentrate on the dependence of the resist profile of a positive resist on exposure dose and development mode. In a developer-dominated process, "force" developed with $R/R^0 < 5$, a shallow outward sloping resist profile results, and thinning of the entire resist layer occurs. For positive resists, the shallow angle is the most typical profile with a 45–75° resist wall angle. With a quenched developer, $R/R_0 = 5–10$, and a moderate dose, a straight resist wall profile results (~90°, e.g., 75–95°). In the vertical wall case, the removal of the laterally exposed region has been inhibited, and a perfect pattern transfer of the mask features onto the resist is obtained. An undercut profile is difficult to achieve in positive resists because the optical exposure dose (and hence the development rate of the system) is greater at the surface than at the resist/substrate interface, resulting in a normal profile with shallow resist angles. An undercut profile is desirable for liftoff processes in which deposited layers are lifted from the substrate

Table 1.9 Photoresist Profiles Overview

Profile	Dose	Developer Influence	R/R_0	γ	Uses
(A) Positive resists undercut (a) 	High (often with backscatter radiation)	Low	>10	>6	Ion implant; liftoff; not good for plasma etching; often only obtained through image reversal
Vertical (b) 	Normal dose	Moderate	5–10	<4	Liftoff; reactive ion etch; wet etch; ion beam etch; perfect fidelity
Normal or overcut (c) 	Low	Dominant	<5	<3	Typical for positive resists; wet etch; metallization; <20% resist loss
(B) Negative resists undercut 	Dominant	Little influence	<0.1	<3	Permanent resists; larger devices; MEMS

(A) Positive resist. (a) Desired resist profile for liftoff, that is, exposure-controlled profile also called *undercut*. (b) Perfect fidelity image transfer by applying a normal exposure dose and relying moderately on the developer. (c) Receding photoresist structure with thinning of the resist layer, that is, developer control also called *overcut*, the normal profile for positive resists. (B) Negative resist. Profile is mainly determined by the exposure. Development swells the resist slightly but otherwise has no influence on the wall profile for the older types of resists. The undercut profile is the normal profile for the newer types (aqueous developable) of negative photoresists.
Source: Based on Moreau, M.W. 1988. *Semiconductor lithography*. New York: Plenum Press. With permission.[28]

by dissolving the underlying resist structure (see next section). In the extreme case of overexposure and a fast developer, the result is a lip, retrograde or inward wall-angle undercut (>90, e.g., 95–110°). This profile might come about if there is a lot of light back scatter from shiny surfaces at the interface of the substrate and the bottom of the resist layer. A ratio of $R/R_0 > 10$ qualifies as a fast developer. However, liftoff profiles with positive resists are more readily formed with multilayer resist (MLR) systems or with a postexposure soaking procedure (see below). With negative resists, forming more insoluble products at the resist surface than at the resist/substrate interface, one more easily obtains an undercut profile; a single-layer resist (SLR) will do, and no complicated MLR systems are required in this case.

A more rigorous, mathematical treatment of resist profiles is presented under Mathematical Expression for Resist Profiles.

We now detail the more unusual profiles for positive and negative resists a bit further. Not all photons strike the resist/substrate interface at the same angle; especially with a high overdose, scattering at a reflective interface may cause broadening of the radiation profile at the substrate/resist interface. The scattered radiation profiles for overexposed positive and negative resists are shown in Figure 1.38.

Time-independent organic solvent development (with a strong developer) of overexposed negative resists from outside the scattered region reveals the scattered radiation zone because cross-linking and further swelling inhibit its removal. On the other hand, development of overexposed positive resists using aqueous alkaline is time dependent and rapidly removes the exposed region, edge-scattered radiation zones, and part of the photoresist top unless the developer is quenched. The time-dependent aspect of this process enables the operator to tailor positive resist profiles. With negative resists, the exposed regions remain, as they are rendered insoluble; with positive resists, the exposed region develops, and the unexposed regions usually remain soluble. Swelling in traditional negative resists is one of the reasons these resists used to be limited to the manufacture of devices with minimum feature size of about 3 μm. Scattered radiation and swelling result in a broadening of the remaining resist features. Positive resists do not exhibit this swelling because of a different dissolution mechanism. The oxygen effect, quenching the cross-linking of negative resists as discussed earlier, usually means a disadvantage for negative resists but can be turned into an advantage to improve resolution. We already know that oxygen can scavenge the photogenerated reactive nitrene species and that this reaction eliminates the precursors for cross-linking and insolubilization. Excluding oxygen from the top surface of the resist by flushing with an inert gas or blocking with a polymer topcoat causes the oxygen dissolved in the polymer film to move laterally from the unexposed dark areas into the light zone. The light zone

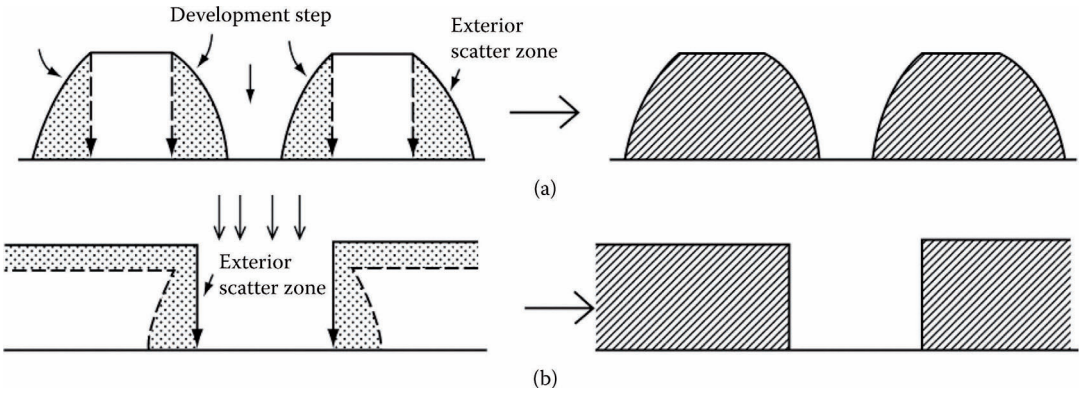


FIGURE 1.38 Edge-scattered radiation profile for negative and positive resists. (a) Negative resist image; time-independent development of cross-linked negative resist fails to remove light scatter zone. (b) Positive resist image; development of positive resist rapidly removes exposed region and can be quenched to inhibit removal of lateral scattered exposed resist region. (Based on Brodie, I., and J.J. Muray. 1982. *The physics of microfabrication*. New York: Plenum Press. With permission.[52])

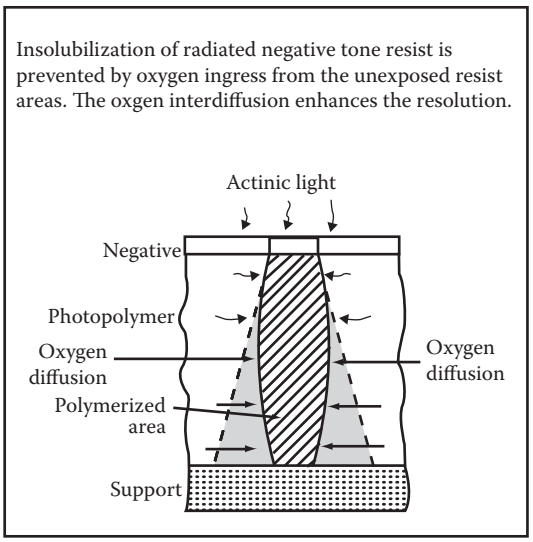


FIGURE 1.39 Lightly scattered zones do not insolubilize. (From Moreau, W.M. 1988. *Semiconductor lithography*. New York: Plenum Press. With permission.[28])

becomes an oxygen sink, and the lightly scattered zones (Figure 1.39) do not insolubilize, leading to better resolution.[28]

Liftoff Profile

The creation of a straight photoresist wall or, better yet, that of a lip or undercut can be taken advantage of in a so-called lift-off process. Lift-off is important, for example, for patterning catalytic metals such as Pt, frequently used in chemical sensors but not easily patterned directly. In the process sequence, shown in Figure 1.40, a solvent dissolves the positive photoresist underneath the deposited metal, starting at the edge of the unexposed photoresist, and lifts off the metal. It is important that there be a discontinuity or gap in the metal deposit so that solvent can get at the uncoated resist wall. This is accomplished by depositing the metal with a line-of-sight–type technique such as thermal evaporation, which is described in Chapter 7. In a line-of-sight deposition technique, a vertical or inward sloping wall will receive little or no metal deposit, leaving a gap for the resist solvent to dissolve the unexposed resist and lift off the metal on top of it. One disadvantage of this technique is the rounded profile, a result of shadowing, associated with deposited features and temperature limitations (see Figure 1.40). A more desirable profile for a conductor line has a rectangular cross-section minimizing electrical resistance. The latter is one reason why liftoff in IC fabrication, where contact resistance is of prime concern, is used with discretion. Also with liftoff,

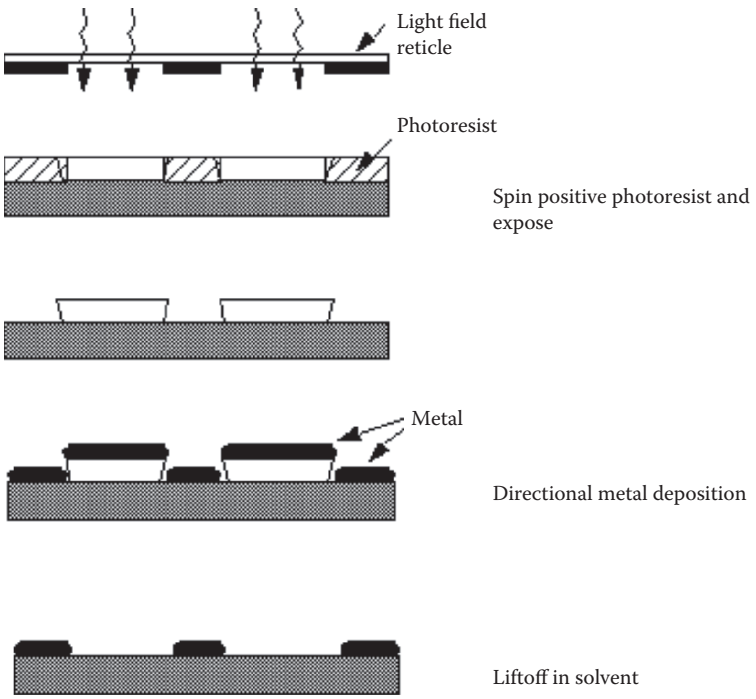


FIGURE 1.40 Example of liftoff sequence with positive resist for the construction of a Pt-based electrochemical sensor electrode. Rounding of deposited features through shadowing is observed (see text).

the metal deposition technique is limited to temperatures less than 200–300°C, where resist begins to degrade.[53]

In a liftoff process, one can either use a negative resist, such as the alkaline-developing ZPN1100, or modify a positive resist to exhibit the desired undercut. We know already that an undercut does not readily form with a positive resist: therefore, some "tricks" are in order. The first trick is to "presoak" the positive resist surface with an aromatic solvent (e.g., chlorobenzene) to convert a surface layer. This layer will develop at a much slower rate than the bulk of the resist film, thereby providing an undercut during development in alkaline solution. The second method, image reversal, follows further below.

A presoak process to develop an undercut on a positive resist (Shipley 1827) involves the following steps—with typical example materials, equipment, and process parameters:

1. Dehydration bake of Si wafer to remove all moisture. Typically this is done in a natural convection oven at 200°C for 30 minutes.
2. Wafer priming using HMDS.
3. Spin coat resist using a Solitec 5100 spinner. Spin to 1000 rpm for 5 seconds to spread the resist and then at 7000 rpm for 35 seconds to complete the spin cycle.
4. Soft bake at 90°C for 20 minutes in a natural convection oven.
5. Expose the wafer for 15 seconds in a Kasper Contact Mask Aligner.
6. A 5-minute chlorobenzene soak. Chlorobenzene diffuses into the photoresist top layer, causing it to swell. A gel is formed to the depth of the diffusion, which develops much slower than the bulk of the resist. This causes the developer to undercut the photoresist structures and produces the desired profile. After the soak, blow-dry the wafers with a nitrogen gun.
7. Develop the wafers using Microposit MF319 from Shipley. Development rates are increased with exposure time and decreased by soaking time. The exposed regions develop faster; the undercut of the structure is formed when the development front passes the gel layer, and fast lateral development of the unexposed regions begins. Use a development time of 5 minutes with mild agitation. Then place the wafers in DI water for approximately 2 minutes, and blow-dry with nitrogen.
8. Deposit desired material.
9. Liftoff. Dissolution of the photoresist is the final step in the liftoff process. Positive resist is very soluble in acetone, which has been used traditionally. Soak the wafers in acetone for 5 minutes with mild agitation.

There are many different techniques to make undercut resist profiles with a positive tone resist, including:

1. Chlorobenzene induced lip in single-layer photoresist (see above)
2. Multilayer resists (e.g., a bilayer) where top layer develops more slowly
3. Special undercoatings that develop faster than resist in developer
4. Trilayer methods
5. Image reversal resists (see next section)

Image Reversal

Because swelling does not take place during the development of positive photoresists, several process variations aim at reversing the tone of the image so that the resists can act as high-resolution negative resists. Examples of positive photoresists that may be used for this purpose are OCG 895i and Shipley 1800. One purpose of reversal processes is for the generation of an undercut profile for liftoff. As we have seen, positive resists more readily lead to overcut resist profiles as shown in Table 1.9Ac, whereas with a negative resist an undercut is more readily generated (Table 1.9B). A typical image-reversal process is demonstrated in Figure 1.41a. After a diazopositive resist has been patterned, an amine vapor (such as imidazole or triethanolamine, or, more generally, a base) is diffused into the exposed areas. The amine neutralizes the byproduct of the photodecomposition (a carboxylic acid in this case) and makes these exposed areas highly resistant to further change by exposure to light and highly insensitive to further development. The base inducing the

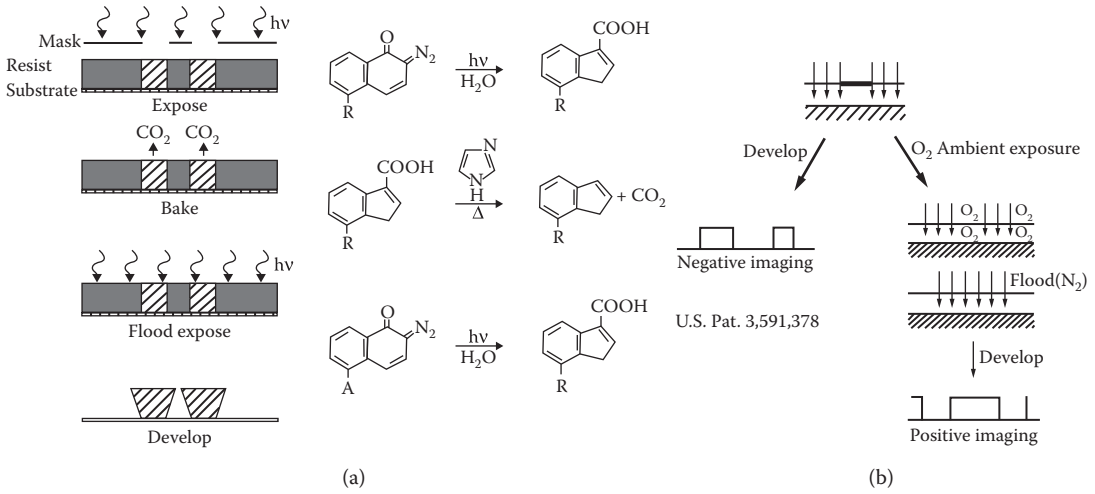


FIGURE 1.41 Image reversal process sequence. (a) Positive photoresist. (b) Negative photoresist. [(a) Based on Alling, E., and C. Stauffer. 1988. Image reversal photoresist. *Solid State Technol* 31:37–43;[54] (b) based on Moreau, W.M. 1988. *Semiconductor lithography.* New York: Plenum Press. With permission.[28]]

image reversal may also be added to the resist formulation before coating. A subsequent flood exposure makes the areas adjacent to the neutralized image soluble in conventional positive photoresist developers. The net result is a negative image of the mask with an undercut profile and improved resolution because the resist does not swell as negative resists do.

Image reversal with a negative resist, such as KTFR, is possible as well (see Figure 1.41b). We have already discussed that remaining oxygen in a resist can improve resolution by preventing insolubilization in light-scattered zones outside the exposed area (Figure 1.39). Image reversal is an extension of this insolubilization. By working at low intensity, oxygen flooding can scavenge all photogenerated azide polymerization precursors, making the exposed areas soluble in a developer as they are prevented from polymerizing. A subsequent flood exposure under a nitrogen blanket, at higher intensity, initiates the polymerization in the previously nonexposed areas.[28]

A dual-tone resist chemistry, based on a phenolic resin with diazonaphthoquinone as the acid generator and hydroxymethylmelamine as the hardening agent, was developed in 1986 by Feely and coworkers at Rohm and Haas (Figure 1.42).[10] The increase in molecular weight for the exposed dual-tone Rohm and Haas photoresist is the result of a condensation reaction between phenol formaldehyde resins and amino-based cross-linkers. Initially, the Rohm and Haas process drew a lot of attention because it promised to have many miniaturization science applications. Earlier, in Figure 1.8a, we illustrated the process flow options for this resist.[9]

Exposure produces the latent image, which results in the "normal" positive resist relief image when developed at this stage (right-hand path in Figure 1.8a). If, on the other hand, after exposure, the film is heated to induce the acid-catalyzed reaction of the melamine with the phenolic resin, the exposed areas become insoluble, and image reversal is initiated. The reversal is completed by a subsequent flood exposure, solubilizing the nonexposed areas and rendering a negative-tone image of the mask (left-hand path in Figure 1.8a). The cross-linked nature of the image imparts resistance to swelling and dissolution by aqueous and organic solvents, as well as dimensional and thermal stability to temperatures >300°C. By using three-tone masks that have opaque, transparent, and partial transmission (50%), stepped (positive-tone process) and cantilevered (negative-tone process) structures are produced (Figure 1.8b). By appropriate processing of image reversal systems, positive, negative, and vertical walls become possible as well. Figure 1.8c showcases an example of the types of microstructures that can be printed with the Rohm and Haas system. In principle, this resist could be the basis of a complete microstructural universe.

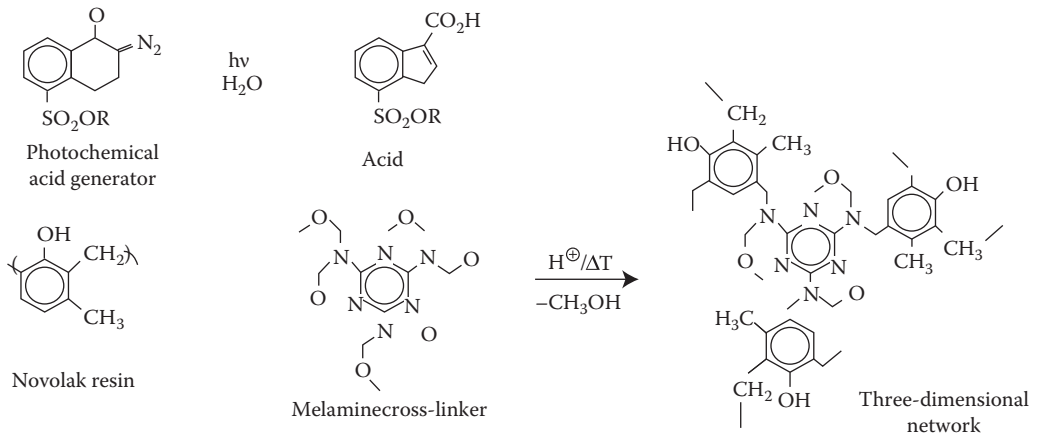


FIGURE 1.42 Cross-linking of a novolak resin through acid-catalyzed condensation with a melamine for negative aqueous-base development.

Contrast and Experimental Determination of Lithographic Sensitivity

The resolution capability of a resist, defined as the smallest line width to be consistently patterned, is directly related to resist contrast γ. For positive resists, the contrast is related to the rate of chain scission and the rate of change of solubility with molecular weight. The latter is very solvent dependent. After development, the thickness of the exposed resist layer decreases until, at a critical dose D_P, the film is completely removed. Lithographic sensitivity, D_P, and contrast can be obtained from the response curve—a plot of normalized film thickness versus log D (dose) (Figure 1.43a). To construct a curve as shown in Figure 1.43a, a series of positive resist pads of known area are subjected to varying doses and developed in a solvent that does not attack the unexposed film. The thickness of the remaining film in the exposed area is then measured and normalized to the original thickness and plotted as a function of cumulative dosage. Contrast γ_p is determined from the slope of this sensitivity or exposure response curve as:

$$\gamma_p = \frac{1}{(\log D_p - \log D_p^0)} = \left[\log \frac{D_p}{D_p^0}\right]^{-1} \qquad (1.11)$$

And the lithographic sensitivity D_p is the x-axis intersection. For a given developer, D_p corresponds

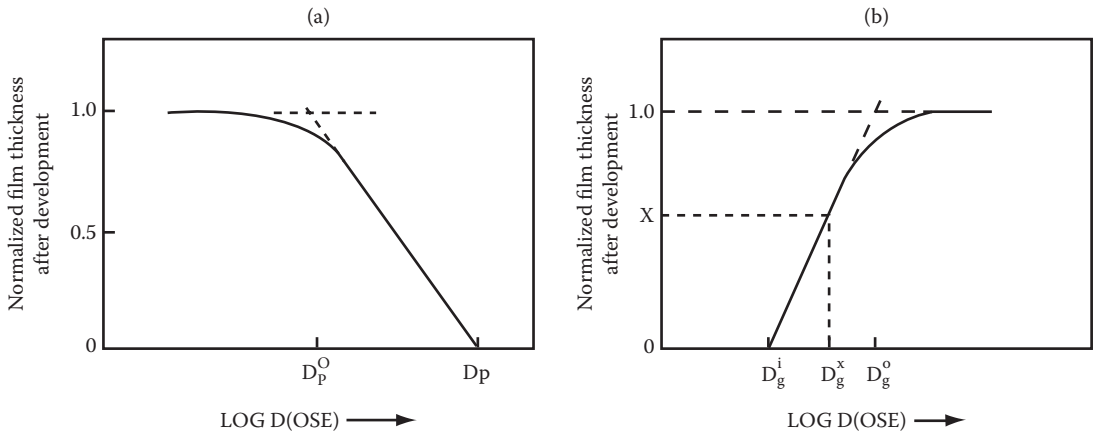


FIGURE 1.43 Typical response curves or sensitivity curves. (a) For a positive resist. Contrast γ_p is determined from the slope. The contrast for a positive resist is markedly solvent dependent. A typical contrast value for a positive optical resist is 2.2. (b) For a negative resist. The value of $D\gamma^x$ usually occurs at 0.5–0.7 normalized thickness as shown in Figure 1.44. The slope determines the contrast, γ_n. A typical value for the contrast of a negative optical resist is 1.5. (From Thompson, L.F., C.G. Willson, and M.J. Bowden, eds. 1994. *Introduction to microlithography.* Washington, DC: American Chemical Society. With permission.[55])

to the dose required to produce complete solubility in the exposed region while not affecting the unexposed resist. D_p^0 is the dose at which the developer first begins to attack the irradiated film. For a dose less than D_p but higher than D_p^0, another developer could "force develop" the resist. In force developing, the developer attacks or thins the original, unexposed resist (Table 1.9Ac). The above describes how the profile of a positive resist can be manipulated by the operator.

For a negative resist, contrast relates to the rate of cross-linked network formation at a constant input dose. This is simpler than in the case of a positive resist, where contrast is also very solvent dependent. Consequently, if one negative resist has a higher cross-linking rate compared with another, it also possesses the higher contrast of the two. With negative resists, the onset of cross-linking, as evidenced by gel formation, is not observed until a critical dose D_g^i (also called the *interface gel dose*) has been reached (see Figure 1.43b). In Figure 1.43b, we show the response or sensitivity curve: the normalized developed film thickness versus log dose. Below the interface gel dose, no image can form, as the film thickness is insufficient to serve as an etching mask. At higher doses, the image thickness increases until the thickness of the image equals that of the resist before exposure (in reality, it remains thinner as the film shrinks as a result of cross-linking). The latter dose is shown in Figure 1.43b as D_g^0, the dose required to reach 100% polymerization of initial film thickness (before exposure). Contrast, γ_n, is obtained from the slope of this curve as:

$$\gamma_n = \frac{1}{(\log D_g^0 - \log D_g^i)} = \left[\log \frac{D_g^0}{D_g^i}\right]^{-1} \qquad (1.12)$$

Lithography sensitivity (D_g^x) defines the dose cross-linking the film to the required thickness for optimal resolution. The required dose is sometimes defined as the dose resulting in dimensional equality of clear and opaque features (corresponding to nominally equal structures on the mask) imaged in the resist. The so-defined lithographic sensitivity can be determined separately from a plot of feature size versus dose for an opaque and a clear feature of equal size (Figure 1.44). This dose, D_g^x, corresponding to the lithographic sensitivity transposed on the x-axis of Figure 1.43b, fixes the required cross-linked film thickness after development on the γ-axis (usually 0.5–0.7 times the normalized thickness). The lithographic sensitivity also may be taken as the dose $D_{0.7}^x$, at which 70% of the original film is retained after development.[27]

Resists with higher contrast result in better resolution than those with lower contrast, and this can be explained as follows. In an exposure, energy is delivered in a diffused manner because of diffraction and scattering effects. Some areas outside the mask-defined pattern will receive an unintended dose higher than D_g^i but lower than D_g^x. The resulting resist profile will exhibit some slope after development. The higher the contrast of the resist, the more vertical the resist profile. Because line width is measured at a specified height above the resist/substrate surface, as the resist profile becomes less vertical, the resist line width represents the original mask dimension less accurately.

Values of D_p and D_g^x are figures of merit used only to compare different resists. For lithographic sensitivity numbers to have any value they must be accompanied by a detailed description of the conditions under which they were measured.

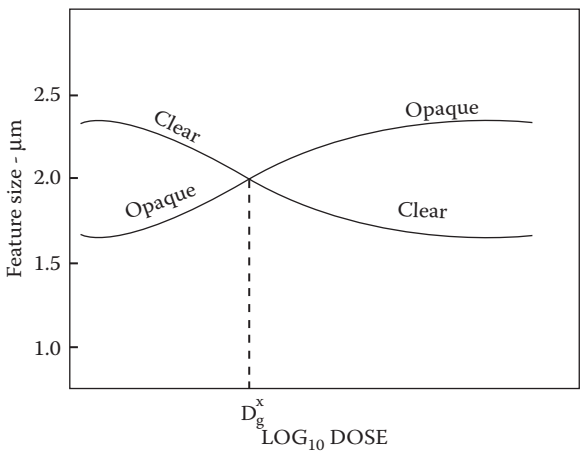


FIGURE 1.44 Size of a clear and opaque 2.0-µm feature (on the mask) as a function of the exposure dose for a negative resist. The dose (D_g^x) resulting in the correct feature size (same size as on the mask) is called the *lithographic sensitivity*. (From Thompson, L.F., C.G. Willson, and M.J. Bowden, eds. 1994. *Introduction to microlithography.* Washington, DC: American Chemical Society. With permission.[55])

Resolution in Photolithography

Introduction

A line-width measurement made with a SEM or another metrology tool from the many options reviewed in Volume III, Chapter 6 is used to determine the resolution of a lithographic system. Correct feature size must be maintained within a wafer and from wafer to wafer, as device performance depends on the absolute size of the patterned structures. The term *critical dimension* (CD) refers to a specific minimal feature size and is a measure of the resolution of a lithographic process. We first consider the theoretical resolution limits of different photolithography printing techniques; in the subsequent section, we review how one can go beyond those conventional limits by using resolution-enhancing techniques (RETs). With reference to Figure 1.6 where we compare contact, proximity, and projection printing, we now delve into a mathematical formulation of the expected resolution for each of these printing techniques.

Resolution in Shadow Printing and Self-Aligned Masks

Introduction

In the shadow printing mode, including contact and proximity arrangements of mask and wafer, optical lithography has a resolution with limits set by a variety of factors. These include diffraction of light at the edge of an opaque feature in the mask as the light passes through an adjacent clear area, alignment of wafer to mask, nonuniformities in wafer flatness, and debris between mask and wafer. Figure 1.45 illustrates a typical intensity distribution of light incident on a photoresist surface after passing through a mask containing a periodic grating consisting of opaque and transparent spaces of equal width, b.[51] Diffraction causes the image of a perfectly delineated edge to become blurred or diffused. The theoretical resolution, R, that is, the minimum resolved dimension in a grating mask (b_{min} for a line or a space) as illustrated in Figure 1.45 and using a conventional resist is given by:

$$R = b_{min} = k\sqrt{\lambda\left(s+\frac{z}{2}\right)} \qquad (1.13)$$

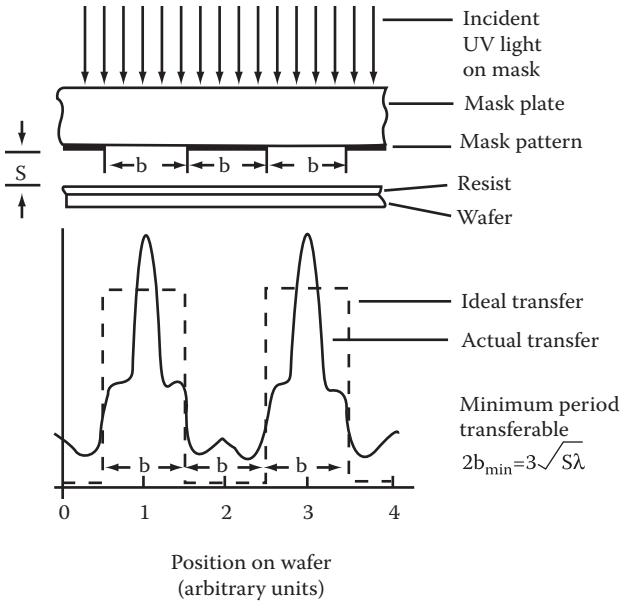


FIGURE 1.45 Light distribution profiles on a photoresist surface after light passes through a mask containing an equal line and space grating. (From Thompson, L.F., C.G. Willson, and M.J. Bowden, eds. 1994. *Introduction to microlithography*. Washington, DC: American Chemical Society. With permission.[55])

where

b_{min} = half the grating period and the minimum feature size transferable
s = the gap between the mask and the photoresist surface
λ = the wavelength of the exposing radiation
z = the photoresist thickness
k = a constant that theoretically is ~1.5

Depending on the relative magnitude of s and z, one distinguishes between contact and proximity printing. The square root relation in contact and proximity printing is a consequence of the near-field or Fresnel diffraction theory valid in the near-field region just below the mask openings; this contrasts with the far-field behavior (Fraunhofer diffraction) of projection lithography (see below). These three types of lithography are compared in Figure 1.46 (see also Volume I, Chapter 5 on far-field and near-field imaging).

Self-aligned printing can be regarded as an extreme form of contact printing where there is no gap between the mask and the substrate.

Contact Printing

In contact printing, a photomask is pressed against the resist-covered wafer with pressures in the range

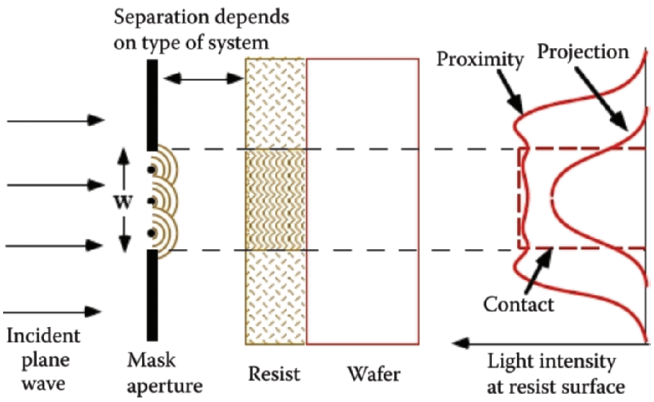


FIGURE 1.46 Proximity and contact printing are near-field, whereas projection is far-field or Fraunhofer diffraction (see Volume I, Chapter 5).

of 0.05–0.3 atm, and the s in Equation 1.13 is zero. In this case, Equation 1.13 reduces to:

$$R = b_{min} = k\sqrt{\lambda \frac{z}{2}} \qquad (1.14)$$

With $\lambda = 400$ nm and a 1-µm-thick resist, we conclude that a resolution higher than 1 µm is possible. For thinner resist layers, that is, z very small and with shorter wavelengths (e.g., 248 nm), the resolution capability of contact printing further improves. Equation 1.14 clarifies the need to use shorter wavelength and thinner resist layers to achieve higher resolution. However, the theoretical maximum resolution is seldom achieved because only diffraction effects were taken into account to derive Equation 1.13. The other factors mentioned above (e.g., wafer flatness, mask alignment) and shadowing effect (penumbral blur) usually conspire to make the resolution worse. Typical contact printers are the Kasper and the Cobilt 800. The required contact between mask and wafer also causes mask damage and contamination, rendering the method unsuitable for most modern microcircuit fabrication.

Proximity Printing

In proximity printing, spacing of the mask removed from the substrate (by at least 10 µm) minimizes defects that result from contact. On the other hand, diffraction of the transmitted light reduces the resolution. The degree of reduction in resolution and image distortion depends on the wafer-to-substrate distance, which may vary across the wafer. For proximity printing, Equation 1.13, with $s \gg z$, can be rewritten as:

$$R = b_{min} = k\sqrt{\lambda s} \qquad (1.15)$$

The technology parameter k theoretically is 1.5. Based on Equation 1.15, for a gap of 10 µm using 400-nm exposing radiation and a k value of 1.5, the resolution limit is about 3 µm.[56] More typical mask and wafer separations are in the range of 20–50 µm. The smallest features resolvable in a practical UV proximity exposure measure about 2–3 µm for most processes. A typical instrument used for proximity printing is the Canon PLA-600FA (the same setup can also be used for contact printing). Resolution is not as good as in contact printing (see above) or projection printing (see below), and for dimensions less than 2 µm optical projection methods are used.

The main disadvantage of proximity printing is a severe reduction in resolution as a result of diffraction spreading. Resolution can be enhanced by either decreasing the gap at the risk of contact and defect generation or by reducing the wavelength. Using wavelengths in the deep UV or even extreme UV range will not suffice for optical proximity printing to compete with projection printing. However, using x-rays with a wavelength of about 1 nm, feature sizes less than 0.2 µm can be produced with proximity methods. This makes 1× proximity x-ray a promising candidate for NGL.

Resolution with Self-Aligned Masks

The most desirable fabrication processes involve in situ deposited masks, also called *self-aligned* or *conformable* masks. These masks, forming molecular contact with the machine under construction, offer superior resolution as light has no chance to diffract between mask and substrate. They may be regarded as an extreme case of contact printing—that is, s is zero with the mask in atomic contact over the whole device under construction. Conformable masks comprise either permanent or sacrificial layers produced in intermediate process steps on the device itself rather than on separate quartz plates.

Figure 1.47 presents an important example of self-alignment with the so-called SALICIDE process, i.e., the **s**elf-**al**igned sili**cide** process, a process in which

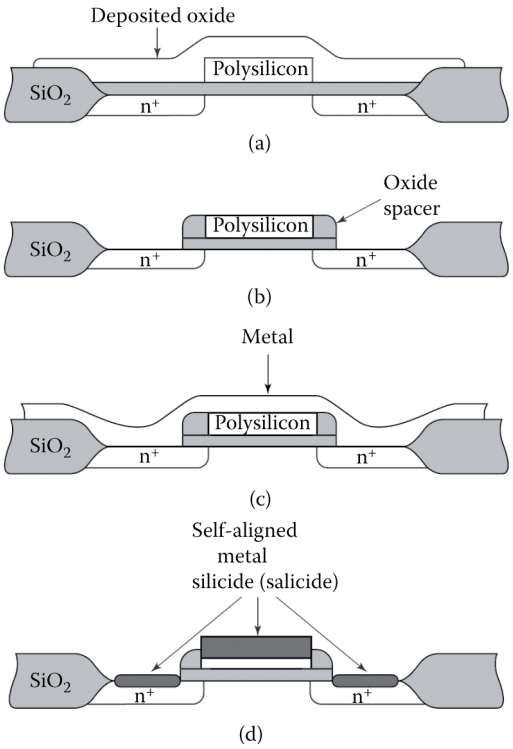


FIGURE 1.47 Example of a self-aligned silicide (SALICIDE) process in the formation of a MOS transistor. (a) An oxide is deposited over the polysilicon gate. (b) The deposited oxide is RIE patterned, leaving the sidewalls of the poly-Si covered with oxide (oxide spacer). (c) Metal is deposited over the structure and heated to form silicides where naked Si is exposed. (d) Unreacted metal is etched away, leaving silicide automatically aligned to gate, source, and drain region.

silicide contacts are formed only in those areas in which deposited metal is in direct contact with silicon. This self-aligning process is commonly implemented in MOS/CMOS processes in which ohmic contacts to the source, drain, and poly-Si gate are formed by the "salicide" process. The salicide process begins with deposition of a thin transition metal layer over fully formed and patterned semiconductor devices (e.g., transistors). The wafer is heated, allowing the transition metal to react with exposed silicon in the active regions of the semiconductor device (e.g., source, drain, gate) forming a low-resistance transition metal silicide. The transition metal does not react with the silicon oxide or nitride insulators present on the wafer. Following the reaction, any remaining unreacted transition metal is removed by chemical etching, leaving silicide contact metal automatically aligned to gate, source, and drain region.

We explained in Volume I, Chapter 4 that an important advantage of the SALICIDE process in new generation CMOS circuitry is that the entire junction area is used for contact formation, which translates into a lower overall contact resistance. The silicide contact materials on the transistors are then further contacted through chemical vapor deposition (CVD)-deposited plugs of pure W filling the contact vias (see Volume I, Chapter 4).

Projection Printing

Resolution

In projection printing, wafer contact is completely avoided; a high-resolution lens projects an image of the photomask onto the photoresist-covered wafer. Imaging in this case is far field or Fraunhofer diffraction dominated, as shown in Figure 1.46. A schematic of a projection lithography system is presented in Figure 1.48.

In Volume I, Chapter 5, we calculated (based on the Maxwell equations) the Rayleigh criterion for far-field resolution as:

$$R = 0.61\lambda/NA$$

with λ the wavelength of the light used for the pattern transfer, and NA the numerical aperture of the imaging lens system.

Diffraction in a small opening transforms normally incident light into an angular spread as

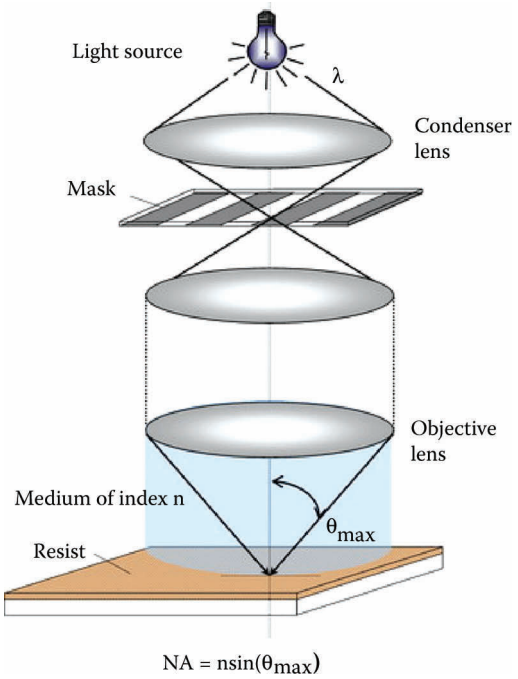


FIGURE 1.48 Optical projection printing. The basic components of a generic optical projection system.

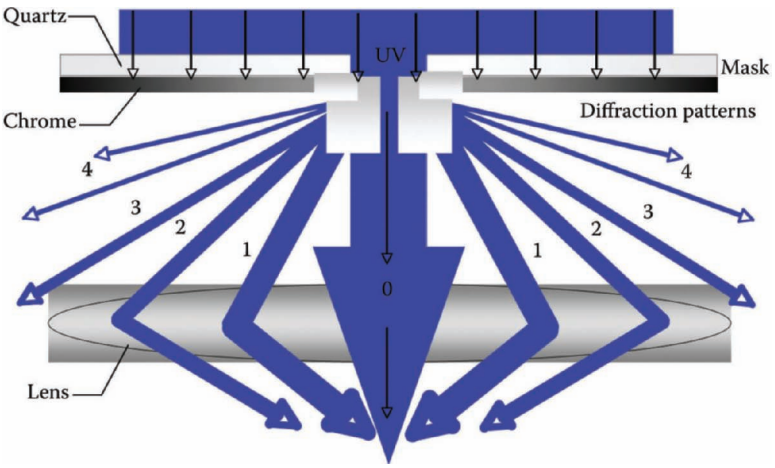


FIGURE 1.49 The numerical aperture (NA) of a lens tells us how much of this diffracted light the lens can capture and image (see Volume I, Equation 5.3: NA = nsinθ).

illustrated in Figure 1.49. The numerical aperture of a lens determines how much of this diffracted light the lens can capture and image (see Volume I, Chapter 5: NA = nsinθ). Because of this spreading effect, fewer diffracted orders are captured by the lens and form the image. This means that information about the smallest pattern is lost because the smaller the pattern the wider the spread. In Figure 1.50 we show the effect of including increasing numbers of diffracted orders on the image of a slit of width *w*. The more diffracted orders we can pick up, the higher the fidelity of the resulting image. As explained in Volume I, Chapter 5, for absolute image fidelity, we need an NA = 1. Pendry's near-field amplifying superlens (NFSL) offers that prospective. A near-field superlens illustrated in Volume I, Chapter 5 enables subdiffraction limit imaging so that even nanometer objects can be imaged with light of say 365-nm wavelength.

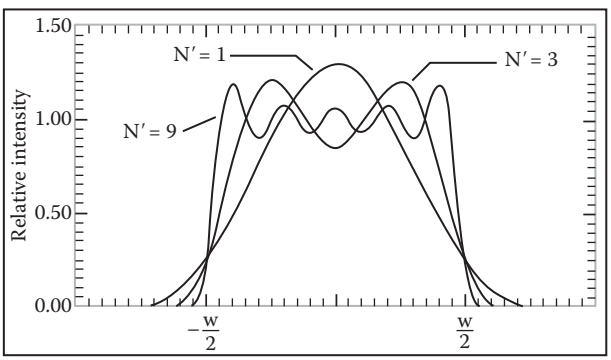


FIGURE 1.50 The higher the number of diffracted orders a lens can pick, i.e., the higher the NA of a lens, the better the resolution of the image.

The NA of lenses in projection aligners ranges from ~0.16–0.80 (see Figure 1.80). With a good lens, possessing an NA of 0.5, the Rayleigh expression suggests that the best resolution one can obtain roughly equals the size of the wavelength of the exposure system. Although the Rayleigh criterion is too simplistic—it only takes into account characteristics of the optical system—it correctly predicts that a larger NA or a shorter wavelength leads to better resolution.

The derivation of R = 0.61λ/NA (Volume I, Chapter 5) was based on the concept of a "point source" something that does not really exist. Here we want to introduce the practical resolution of a photolithography system starting from the grating equation. From the grating formula $d = \frac{N\lambda}{2n\sin\theta} = \frac{N\lambda}{2NA}$, we calculated that the resolution with incoherent illumination is better than with coherent light [R = 0.5 λ/NA for coherent light (Volume I, Chapter 5) vs. R = 0.25 λ/NA for incoherent light (Volume I, Chapter 5)]. We generalize these results using a dimensionless constant k_1 to define the practical limiting resolution *R* in projection printing as:

$$R = k_1 \frac{\lambda}{NA} \tag{1.16}$$

where k_1 is the experimentally determined dimensionless parameter that depends on resist parameters, process conditions, mask aligner optics, and so on.

To achieve higher resolutions, we want to work at shorter wavelengths λ, higher NA, and smaller k_1.

The k_1 parameter fluctuates between 0.3 for surface imaging resists to 0.5 for multilayer resists (MLRs), 0.75 for single layer resists (SLRs), and 1.1 for reflective surfaces like aluminum. Up until 1995, production engineers demanded that k_1 be about 0.7 for SLRs; today, its value is closer to 0.5 (see Figure 1.81). Technologies enabling lower k_1 values, besides surface imaging and MLRs, include phase-shift masking, antireflective coatings, and off-axis illumination. We will review these methods under resolution-enhancing technique (RET) and learn, for example, that with RET one can obtain a resolution about one-half of the exposure wavelength or even better. The better resolution afforded by a lower k_1 makes for a narrower process window and thus a more difficult process; it is expected that one cannot go much lower than a k_1 of 0.25 (see Figure 1.81).

Depth of Focus

To obtain good line-width control, the latent image must remain in focus throughout the depth of the resist layer. A certain amount of defocus tolerance is allowed. The defocus tolerance or depth of focus (DOF or δ) of an optical system is given by:

$$\text{DOF} = \pm\delta = \pm k_2 \frac{\lambda}{NA^2} \tag{1.17}$$

And by using Equation 1.16:

$$\text{DOF} = \pm\delta = \pm k_2 \frac{R^2}{k_1^2 \lambda} \tag{1.18}$$

In Equations 1.17 and 1.18, k_2 is a process-dependent constant hovering around 0.5. For a detailed mathematical derivation of the DOF equations we refer to Bowden.[56] From Equation 1.16, we know that to achieve better resolution we must reduce λ and increase NA. As deduced from the DOF equations, the penalty is a reduction in DOF, perhaps becoming so small that a focused image through the total depth of a typical 1.0–1.5-μm-thick photoresist cannot be achieved anymore. A good NA of a lens for a g-line (436 nm) lithography system is 0.54. With a k_1 factor of 0.8, this leads to a resolution of 0.65 μm (based on Equation 1.16). With i-line (365 nm) lithography, a resolution of 0.65 μm can be achieved with a 0.45-NA lens while exhibiting a superior DOF of 0.9 μm compared with 0.7 μm for the g-line. Wide-

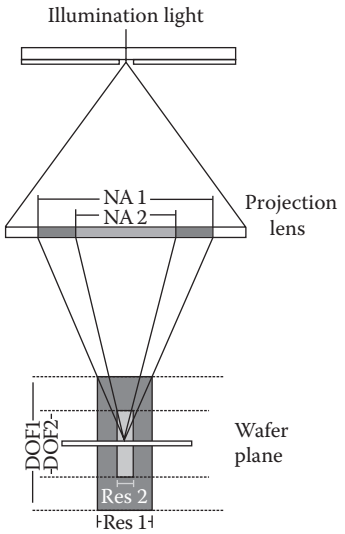


FIGURE 1.51 Variable numeric aperture stepper.

field i-line steppers with variable numeric apertures (Figure 1.51) are available with a 22 × 22-mm field, a 5× reduction ratio, and a resolution of better than 0.35 μm (e.g., the Nikon NSR-2205i14E, http://www.nikon.co.jp/main/eng/news/dec14e_97.htm). This type of equipment allows one to balance resolution, DOF, and wafer throughput for different applications. Because the patterns on wafers have their own topology and wafers may not be perfectly flat, a large DOF is needed to ensure that the image stays in focus over the entire field. If even a small part is out of focus, the final product will be ruined.

Photography enthusiasts easily relate to this discussion (see Figure 1.52): a high aperture setting on a camera leads to sharp images for objects that are close to the camera (see rose), whereas a small aperture keeps objects in focus all the way to infinity (see stones).

Microlithographic (or practical) DOF is defined as the total defocus allowable for a desired tolerance on a minimum feature size. It may be different from the values estimated from Equations 1.17 and 1.18. The practical DOF must encompass device topography, resist thickness, wafer flatness, and focus tilt errors. For sub-0.5-μm lithography, only a small amount of residual nonplanarity in device topography can be tolerated without negatively affecting critical dimension control. Small DOF values require expensive planarization processes to bring all IC features in focus within the DOF of the optical system. Some of these planarization processes will be reviewed

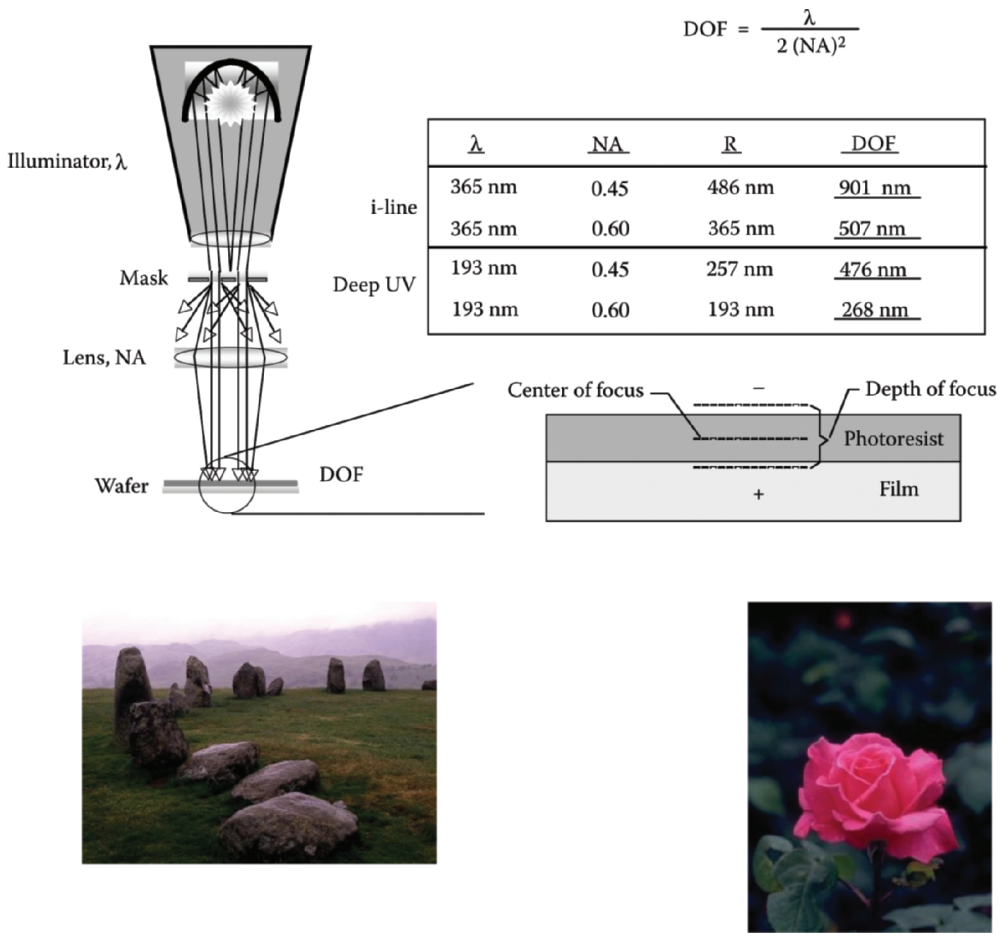


FIGURE 1.52 The concept of depth of focus (DOF) in lithography and photography.

further below. Thin film imaging (TFI) methods, reviewed below, also allow one to work around DOF problems by imaging in thin film resist layers and then transferring that image to underlying thicker resist layers. Miniaturized structures often possess more extreme topologies than ICs, making planarization an even bigger challenge. During the past 20 years, progress in expanding the DOF of semiconductor equipment has not kept pace with the decreasing CDs. For miniaturization science, the former is more important. We will learn how topographical masks and using x-ray and e-beam lithography have enabled the higher DOFs needed for high-aspect-ratio miniaturized 3D machines.

Types of Projection Methods

Since 1973, when Perkin-Elmer (http://instruments.perkinelmer.com) first introduced its scanning projection system for lithography, optical projection of mask patterns has become the standard lithography method. Some of the many projection systems in use include projection scanners (e.g., Perkin-Elmer's Micralign 700 series), 1:1 and reduction (e.g., 5:1 and 10:1 times, often denoted as 5× and 10×), step-and-repeat projection systems (e.g., 10× Electro-Mask), step and scan systems [e.g., ASML's (http://www.asml.com) PAS 5500/900], and double-sided mask aligners [e.g., the SUSS MicroTec MA/BA6 Mask Aligner (http://www.suss.com/products/mask-aligner/ma-ba6-gen2.html)].

A scanning projection system exposes the wafer in a single scan without the benefit of midscan realignment to local alignment marks. The wafer and the mask move simultaneously and continually on an air-bearing carriage through a light arc covering the whole mask and wafer at once. With a deep UV light source, resolution of 1 μm can be obtained, DOF of ±6 μm is possible, and an overlay accuracy of ±0.25 μm (1 σ) has been reached. In practice, these scanners are mainly used for alignment of patterns with CDs in the 3-μm range and for high throughput applications (e.g., 100 wafers/hour).

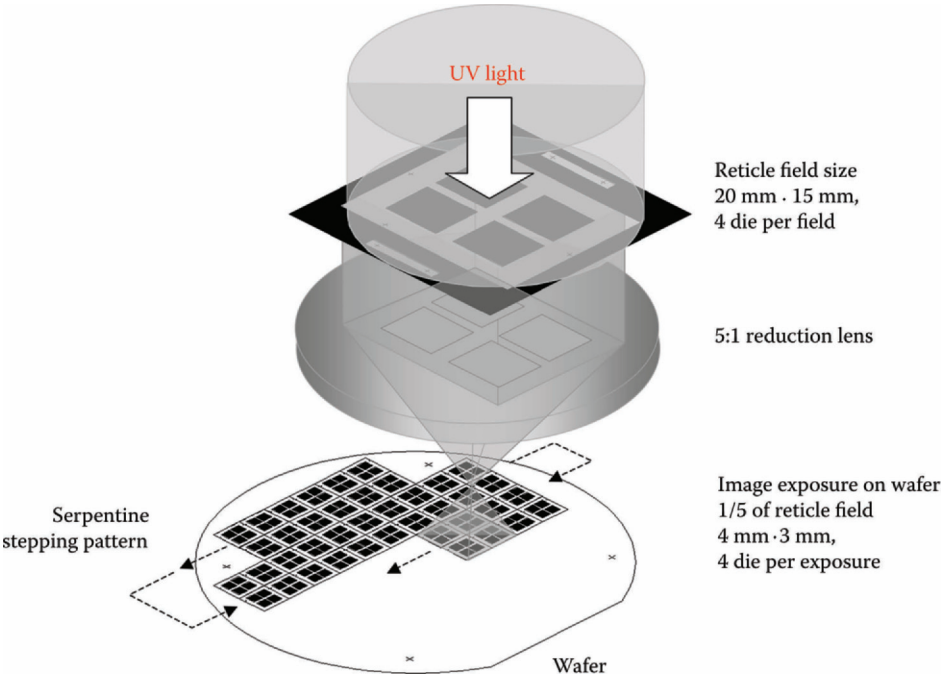


FIGURE 1.53 Stepper exposure field.

A stepper system, also step-and-repeat, exposes one small part of the wafer followed by a new exposure at the next position (Figure 1.53). By stepping and repeating, the entire wafer is covered with the reticle pattern. Steppers have the ability to align to each field and make adjustments to x, y, rotation and focus, and tilt. They offer higher alignment accuracies but are slower than scanners. Most steppers use reduction lenses (e.g., with a 10:1 or a 5:1 reduction) rather than one-to-one projection printing. The reduction printer exposes part of the wafer to a pattern from a mask 5–10 times larger than the projected image. The reduction process makes reticle inaccuracies less significant, consequently improving the resolution and resulting in easier mask making. The only drawback pertains to the size of the image field: the higher the reduction ratio, the smaller the image field. Because of lens imperfections and diffraction considerations, projection techniques have a lower resolution for pattern transfer than that provided by a contact or self-aligned mask; however, CDs of 65 nm have already been achieved today with projection lithography.

With steppers, the full exposure field (pretty much the whole lens) is illuminated, and the reticle image is projected in one flash at each location across the wafer. In newer step and scan systems (Figure 1.54), used in 193-nm lithography, only a narrow slit of the wafer is illuminated. In tools of this type, the wafer is stepped to a new field, which is then scanned; this sequence continues until all the fields have been scanned. This method projects a part of the image from the reticle onto the wafer and uses only a small fraction of the lens area. By synchronously moving the reticle stage (up to 1 m/s) and the wafer stage (up to 25 cm/s) in opposite directions, the whole reticle pattern is imaged onto the wafer. With step and scan systems, larger image fields can be exposed (see Figure 1.54), resulting in higher productivity; because a smaller area of the lens is used (only a slit over the center of the lens), each lens can be better optimized for higher resolution and lower distortion.

A mask aligner of particular interest to the non-IC miniaturization engineer is a double-sided mask aligner. With this equipment, features can be aligned on opposite sides of the wafer. An example is the SUSS MA-150 RH with a top-to-bottom precision alignment accuracy of 1 μm (1σ) in production. Another double-sided top and bottom aligner is the 600 series EV640 by Electronic Visions Co. (http://www.elvisions.com). Double-sided mask alignment is discussed in more detail under "Mask Alignment in Projection Printing," this chapter.

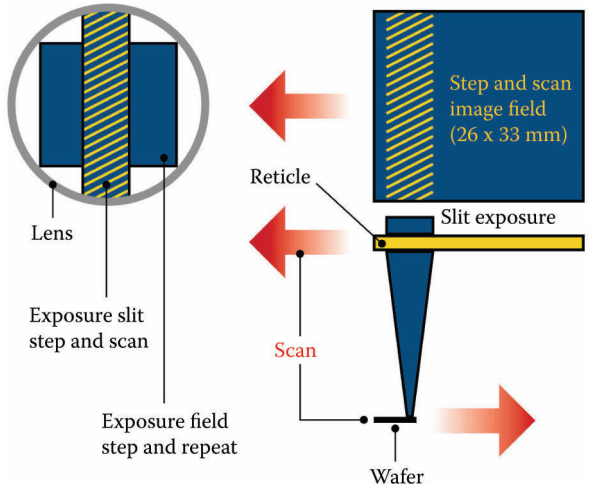


FIGURE 1.54 Step and scan systems. With wafer steppers, the full image is exposed in one flash, whereas with the step and scan systems, the wafer stage and reticle stage move simultaneously in opposite directions during the exposure to scan the image onto the wafer.

Modulation Transfer Function

In the projection printing of a grating, a series of light intensity maxima and minima are produced, as illustrated in Figure 1.55. Because of mutual interference, dark regions in the image never reach complete darkness, and the maximum brightness never corresponds to 100% transmission. The quality of the image transfer can be expressed by the modulation index (also simply modulation) M defined as:

$$M = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \qquad (1.19)$$

In this equation, I_{max} and I_{min} are the peak and trough intensities, respectively. M reveals the degree to which diffraction effects cause incident radiation to fall on the resist between the images of two slits in a mask. Ideal optics would give a modulation M equal to 1. In practice, exposure systems behave less ideally with an $M < 1$.

The optical imaging quality (that is, the capability of reproducing a mask feature on a wafer surface) for a given projection system can be characterized in terms of the modulation transfer function (MTF) curve. The MTF of an exposure system is defined as the ratio of the modulation in the image plane to that in the object or mask plane (that is, M_{im}/M_{mask}) as a function of spatial frequency (ν) or number of line pairs per millimeter on the mask. Because the intensity in the mask plane at the center of an opaque feature is essentially zero, we can equate $M_{mask} = 1$. Consequently, we can also equate MTF and M_{im} as shown in Figure 1.55 where the curve is normalized to 100%. An MTF curve is obtained by imaging a series of gratings with different spatial frequencies placed in the object plane, and the corresponding modulation $M_{im}(\nu)$ in the image plane is measured.[52]

The MTF of an exposure system depends on NA, λ, mask feature size, and the degree of spatial coherency of the illuminating system. Because NA and λ are fixed by the system hardware design, a plot of MTF versus feature size, parametrically changing the spatial coherency, often is used to compare a system's imaging capability. Coherence (σ) varies from $\sigma = 0$ for coherent radiation to $\sigma = \infty$ for fully incoherent illumination. Only point sources are completely coherent with all light waves impinging perpendicular on the mask. In reality, we deal with light sources of finite size as illustrated in Figure 1.56, resulting in partially coherent light. The degree of coherence of the light on the resist plane is dependent on that size, and the resulting light is usually only partially coherent. The definition of coherence is:

$$S = NA_c/NA_o \text{ or also } S = s/d \qquad (1.20)$$

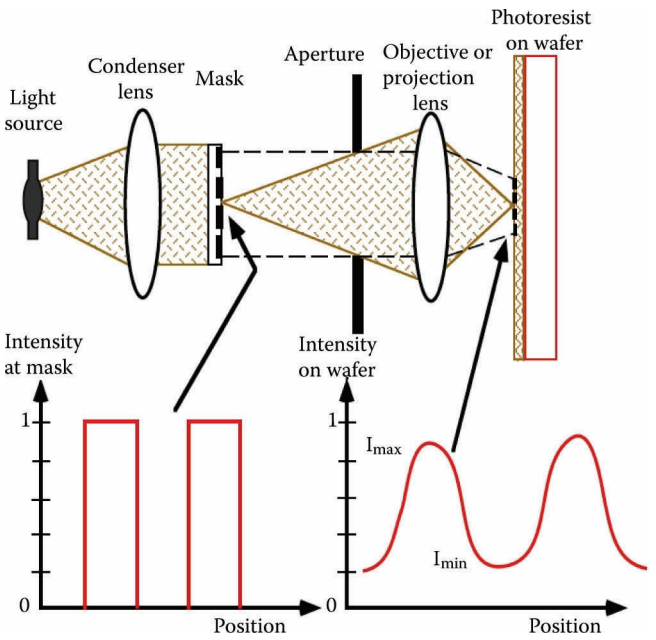


FIGURE 1.55 Modulation and modulation transfer function (MTF).

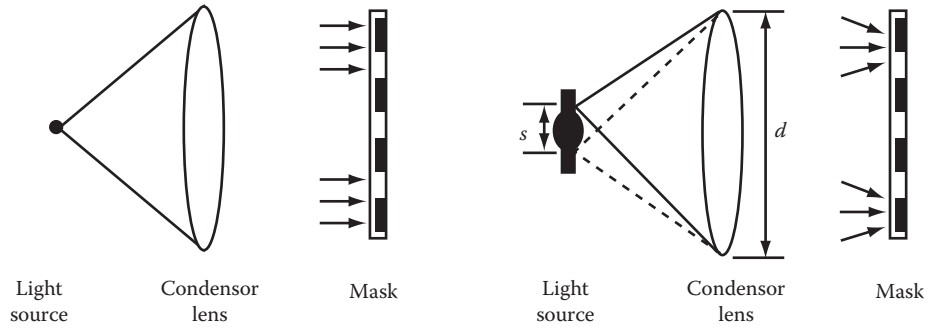


FIGURE 1.56 Coherent and incoherent light.

where NA_c and NA_o are the NA of condenser and objective lens, respectively, s is the size of the light source, and d is the diameter of the lens (see Figure 1.56).

MTF curves (also the modulation M_{im} in the image plane) for coherent, partially coherent, and incoherent illumination are shown in Figure 1.57. Contrary to the coherent case in which MTF remains constant up to the cutoff frequency $\nu_{max} = NA/\lambda$, the MTF curve for the incoherent case decreases monotonically as the frequency increases up to a maximum value ν_{max} of $2NA/\lambda$. Because $\nu_{max} = 1/2b_{min}$ and $R = b_{min}$, the maximum resolution for coherent light is predicted to be $0.5\lambda/NA$; for incoherent light it is $0.25\lambda/NA$ (see also the derivation of Equations 5.6 and 5.9 in Volume I). Completely incoherent light produces an MTF curve that has a greater resolution limit, but the MTF value only slowly increases as the feature size increases. For these reasons, partially coherent light often is preferred.[56] Based on various tradeoffs, a coherence value of ~0.7 typically is selected for a practical exposure. The cutoff frequency shown in Figure 1.57 defines the resolution limit of the exposure system. It is proportional to the NA of the lens and inversely proportional to the wavelength. Good sources for a more detailed treatise on the MTF are Bowden,[56] Sheaths,[57] and Wolf.[50]

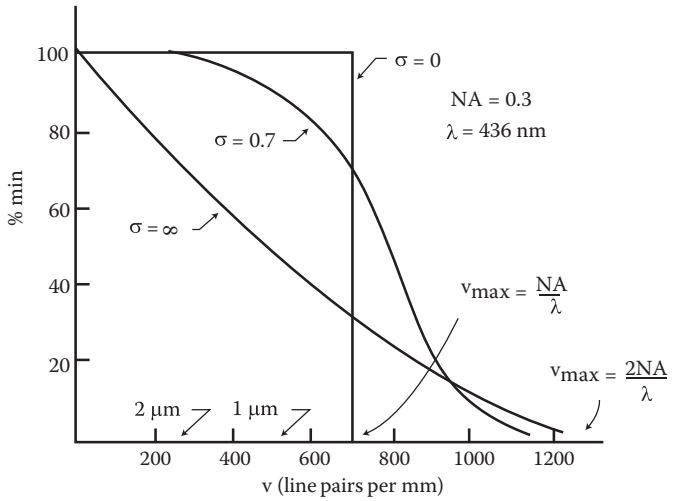


FIGURE 1.57 Percent modulation in the image plane, % M_{im}, modulation of an image as a function of spatial frequency ν for different coherency factors. σ = 0 (coherent), σ = 0.7 (partially coherent), and σ = infinite (incoherent). (From Brodie, I., and J.J. Muray. 1982. *The physics of microfabrication*. New York: Plenum Press. With permission.[52])

Critical MTF Values

For the practical resolution of a resist to approach the Rayleigh limit of the optical system, the resist should have an infinite contrast. Because a resist always has a finite contrast value, a greater degree of modulation is needed before an adequate image can be formed. Thus, the minimum MTF of an optical system to adequately define an image in a resist depends on the resist contrast value γ and is defined as the critical MTF, or $CMTF_{resist}$. The relationship between the resist contrast and CMTF is given by:

$$CMTF_{Resist} = \frac{10^{\frac{1}{\gamma}} - 1}{10^{\frac{1}{\gamma}} + 1} \qquad (1.21)$$

A resist material with $\gamma = 2$ results in a CMTF value of 0.52. For an exposure system to adequately print a given feature size, the MTF of the feature size must be larger than or equal to the $CMTF_{resist}$ of the resist used. If the MTF of an exposure system is known for various feature sizes, knowledge of the resist contrast and Equation 1.21 will allow prediction of

the smallest features printable when applying that system.

Mask Alignment in Projection Printing

Fiducial Marks Thus far we have not dwelled much on the alignment or registration of the wafer to the mask in so-called aligners or printers. In projection lithography, the mask, or reticle, is held in place above the projection lens by a reticle stage. The projection lens focuses the high-resolution images of the reticle patterns precisely onto a wafer that is positioned and held in place by the wafer stage. During this critical alignment process, the wafer stepper aligns the mask image to alignment or fiducial marks on the wafer. This ensures that the pattern pictured by each mask layer is precisely aligned and "overlays" the previous layer. In the process of manufacturing a miniaturized device, several thin films are stacked on top of one another; each is patterned differently; and each wafer goes through the wafer stepper 20–30 times. For a detailed description of mask alignment equipment, we refer the reader to the specialized literature.[58] We discuss the types of errors that can be expected from mask alignment below.

Mask alignment errors can have catastrophic effects on the performance of integrated circuits and miniature machines, often rendering them incapable of operating. For example, a device that has a minimum line width between 1 and 1.5 μm can only tolerate a variation of ±0.25 μm with respect to the alignment of masks without encountering greatly increased device failure rates. Product failure is caused mainly by poor alignment between the image being projected and the pre-existing patterns on the wafer. To align one layer with a previously fabricated layer when performing photolithography, appropriate fiducial marks are created on the mask. Different laboratories and foundries use different marks, and these may need to be placed in a specific position on the design. Additional requirements include a scribe lane around each chip to indicate where the wafer is to be cut when it is diced. The inclusion of a unique mask number and an indication of layer names on the mask make it possible to tell by looking how far through the fabrication process a wafer has progressed (http://www.dbanks.demon.co.uk/ueng). In Figure 1.58, we illustrate a

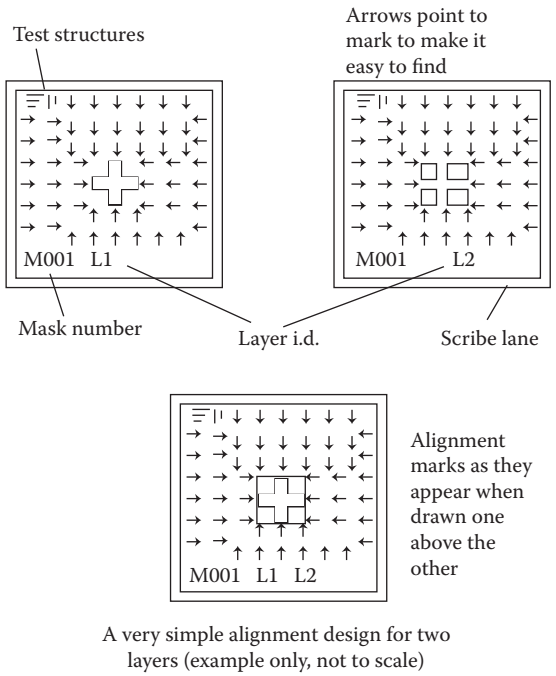


FIGURE 1.58 A simple set of fiducial marks and other important guiding features often found on masks. (Based on Banks at http://www.dbanks.demon.co.uk/ueng.)

mask set with many of the common features: fiducial marks, arrows to point to the marks, test structures, and mask and layer numbers. Optical vernier (Figure 1.59) patterns created on the different levels to be aligned are another common and useful feature. Vernier scales allow us to resolve alignment errors more accurately than the minimum feature size for a given process. To gain this fine resolution, vernier scales depend on accurately spaced lines. A pictorial tutorial on how to use vernier scales in lithography can be found on the Internet at http://www.schlenkent.com/vernier.htm.

Types of errors encountered in projection printing are illustrated in Figure 1.60. They range from misalignment to mask error, optical distortion, wafer or

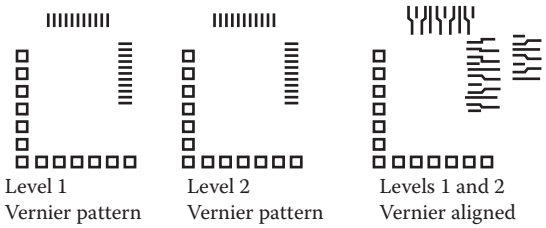


FIGURE 1.59 Standard optical vernier test pattern for overlay evaluation. (From Moreau, W.H. 1988. *Semiconductor lithography.* New York: Plenum Press. With permission.[28])

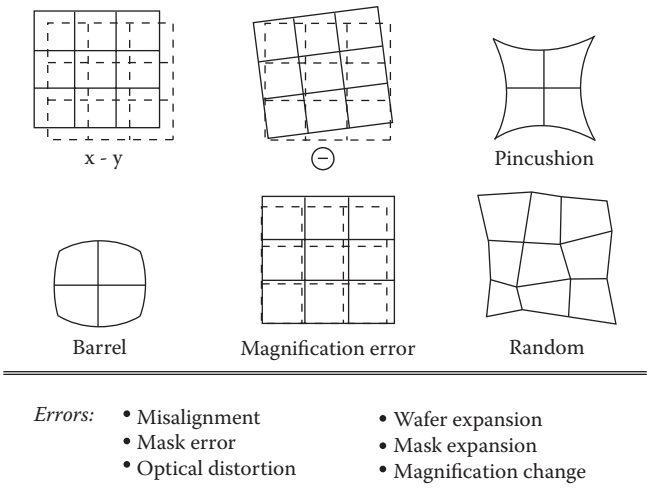


FIGURE 1.60 Sources of errors in projection printing.

mask expansion, and magnification change. Pattern registration capability is the degree to which the pattern being printed can "fit" mask alignment marks and previously printed patterns. With an optical stepper, the step-and-repeat operation is performed by laser interferometer-controlled stages to a positioning accuracy of <40 nm, thereby allowing the registration of successive layers of a semiconductor device with similar precision. Because interferometric schemes enable sensitivities at less than 10 nm, it appears feasible that aligners compatible with line widths of 0.1 µm or even 50 nm can be developed. For CDs of 0.25 µm and even down to 0.20 µm, conventional wafer steppers can be used. However, at design rules of 0.18 µm, step and scan technology becomes more important (see above under Types of Projection Methods).

Alignment in MEMS Alignment is more complex in 3D miniaturized machines than that in IC manufacture. Not only does one deal with high-aspect-ratio 3D features, which can cause problems for alignment systems with low DOF, but one also frequently needs to align 3D features on both sides of the wafer. The objective is to position alignment marks opposite each other on the two surfaces of the same wafer, allowing accurate positioning of all later feature-defining photolithographic patterns. There are several options for front-to-back alignment. One not too elegant or desirable way is by etching holes through the silicon wafer. Another better way is to use infrared to see through the silicon or, with visible light, to use a double-sided mask alignment system defining marks on opposite sides of the wafer with a set of mirrors, two light sources, and two masks. In the latter double-sided mask aligner, mask 2 is first aligned to mask 1; then the wafer is inserted and aligned, and UV exposure of both sides can begin. Wafers polished on both sides are used to minimize light scattering. Two popular commercial double-sided alignment and exposure systems are by SUSS, Germany, and Electronic Visions, Austria. The operation of the SUSS MA-150 is illustrated in Figure 1.61. Cross-hair marks on the mask are aligned to cross-hair marks on the back of the wafer. In the alignment process the alignment marks on the mask are first viewed by two microscope objectives, and the image is electronically stored. The wafer is then loaded with the back-side alignment marks facing the microscope objectives, and the wafer stage is moved and adjusted until the marks are aligned to

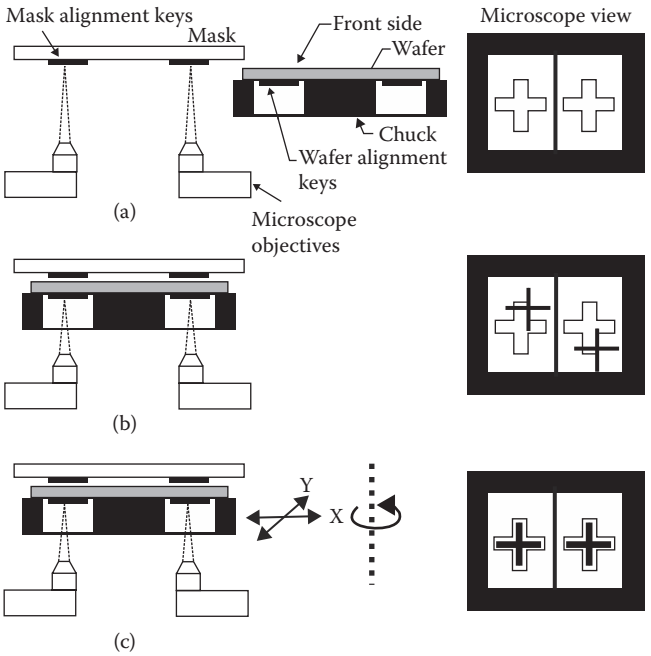


FIGURE 1.61 Double-sided alignment scheme for the SUSS MA-150 production mode system: the image of mask alignment marks is electronically stored (a); the alignment marks on the back side of the wafer are brought in focus (b); the position of the wafer is adjusted by translation and rotation to align the marks to the stored image (c). The right side illustrates the view on the computer screen as the targets are brought into alignment. [Adapted from product technical sheet (SUSS, Munich, Germany) and Maluf, N. 2000. *An introduction to microelectrochemical systems engineering*. Boston: Artech House.[59]]

the electronically stored image. With the alignment finished, exposure of the mask on the front side of the wafer is completed either in contact or proximity mode. On a SUSS MA-150 RH, with robotic bottom-side alignment (BSA), a standard deviation less than 0.33 µm on front-to-back alignment has been demonstrated; of course, the alignment accuracy depends on a multitude of process parameters, but in a typical production mode this mask aligner can achieve an accuracy of better than 1 µm (3σ).

Because two-sided mask aligners and infrared microscopes are costly, researchers have been looking into less expensive but still accurate alternatives. In the simplest, least-accurate approach, wafer flats can be used as reference for the double-sided alignment with, at best, 5-µm accuracy. In a slightly more sophisticated procedure, White and Wenzel[60] use a simple laboratory jig as shown in Figure 1.62. In this approach, mask 1, containing only alignment marks, is contact printed onto photoresist-coated mask 2 while both are positioned snugly against the three pins on the jig. After developing and etching the alignment marks on mask 2, the individual alignment patterns from the two masks are transferred onto the opposite faces of a semiconductor wafer coated on both sides with photoresist. This is accomplished by sandwiching the resist-coated Si wafer between the two alignment masks (again set snugly against the three pins of the jig), exposing each wafer surface (directly for one side and through the large hole in the jig for the other side). The alignment patterns then are etched into the wafer and used in a conventional one-sided mask aligner. The authors estimate the predictable alignment errors to be less than 1 µm across a 250-µm-thick, 2-in. wafer. Kim et al.[61] describe a different front-to-back alignment technique involving the visual alignment of stepper crossbars (alignment keys) to the center of a transparent thin diaphragm etched from the back of the wafer. These authors claim a <1-µm alignment error, but no experimental evidence is presented. Tatic-Lucic et al.[62] elaborated on the latter type of double-sided alignment method and demonstrated a <2-µm error across a 4-in. wafer. In this upgrade, only alignment marks on the front of the wafer formed in a low-stress insulating layer are required. This insulating layer with the alignment marks is later etched from the back so that a freestanding diaphragm with a cavity underneath it results. This cavity makes the front-side alignment marks visible from the back. This allows visual alignment from both sides of the wafer with a traditional GCA 4800 stepper. Yet another alternative to facilitate double-sided alignment is the pickup of capacitive signals between conductive metal fingers on the mask and ridges on a small area of the Si wafer. This technique, requiring close proximity between mask and wafer, is only applicable for 1:1 proximity lithography as used, for example, in LIGA (see Chapter 10).

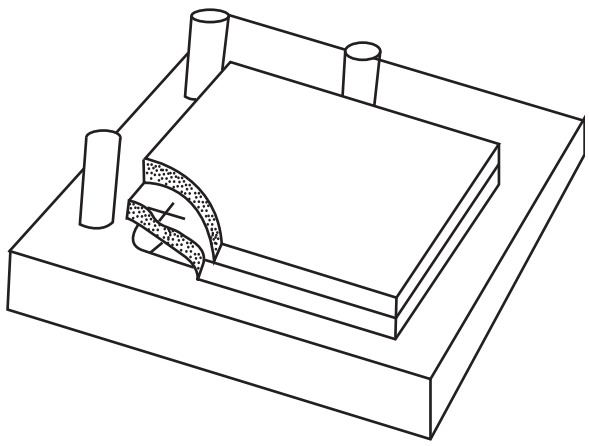


FIGURE 1.62 Sketch of two-sided alignment jig with the two alignment masks in place for contact printing marks onto mask 2 (top mask). The cross represents an alignment mark on mask 1. (Courtesy of Dr. Richard White, University of California, Berkeley.)

There are numerous problems associated with using classical exposure tools for fabricating highly nonplanar miniature machines. Therefore, maskless exposures to photons (in air) or electrons (in vacuum) with high-precision linear and rotary positioning stages and numerically controlled beam direction and stage position are gaining popularity with non-IC machinists. Nonplanar lithography methods are discussed in Chapter 2 on next generation lithographies.

Mathematical Expression for Resist Profiles

The slope of the resist edge (dz/dx) of an image of a mask feature of width W (see Figure 1.63), for a case where the dose D is low and the developer influence dominant, can be written as:

$$\frac{dz}{dx} = \left(\frac{dz}{dD}\right)\left(\frac{dD}{dx}\right) \qquad (1.22)$$

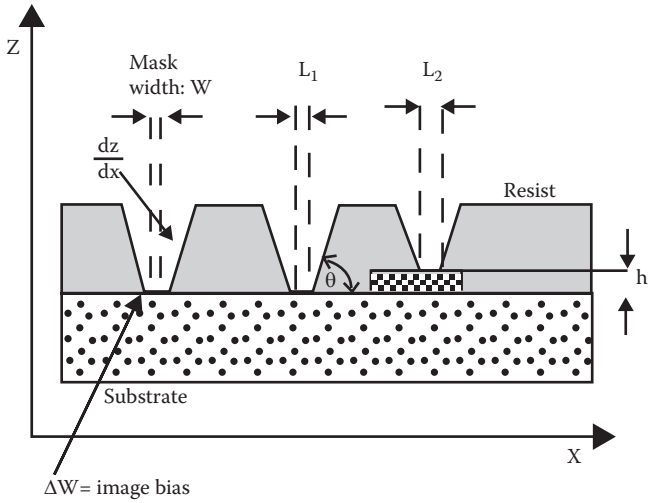


FIGURE 1.63 Edge slope of a positive resist, *dz/dx* primarily determines the image bias, Δ*W*. The change of line width of a feature developed in resist with a topographic feature of height *h* is given by $L_2 - L_1 = \Delta CD$.

where *z* is the thickness dimension and *x* is the lateral dimension in the resist.

The term *dz/dD* (developer elution term) is only resist- and resist-processing dependent. Additionally, the term *dD/dx*, the derivative of the energy or dose absorbed, also called the *intensity profile*, is a term dependent only on the exposure system and the object that is being imaged. The response of a positive resist to dose *D* is given by γ_p, the contrast of the resist, and the developer-process term in Equation 1.22 can be approximated by[63]:

$$\frac{dz}{dD} = -\frac{\gamma_p}{D_p} \tag{1.23}$$

where γ_p is the contrast of a positive resist as given by Equation 1.11, that is, the slope of the curve of thickness remaining versus log of exposure dose. In Equation 1.11, D_p^0 represents the exposure energy below which no resist removal by the solvent takes place, and D_p embodies the sensitivity of the resist or the exposure energy at which no resist remains after development. Exposure energy D_p^0 is independent of the thickness *t* of the resist. Resist sensitivity D_p, on the other hand, will vary strongly with the thickness, predominantly through the absorbance of the film:

$$D_p \approx D_T . 10^{\alpha t} \tag{1.24}$$

with α the resist absorbance per unit thickness and D_T a threshold exposure that depends on resist type and is equal to or smaller than D_p^0. With these further simplifications we can rewrite Equation 1.23 as:

$$\frac{dz}{dD} \approx \frac{1}{(a+\alpha t)D_p} \tag{1.25}$$

where *a* is a constant for the resist and α is the absorption coefficient. The *dD/dx* term in Equation 1.22, or the intensity profile, depends on the object and the imaging system. The intensity profile of the image is affected by the wavelength of exposure λ, the numerical aperture NA, the depth of focus of the exposure tool DOF, and the uniformity of illumination and can be written as:

$$\frac{dD}{dx} = \frac{2NA}{\lambda\left[1-k\left(\frac{DOF \times NA^2}{\lambda}\right)\right]^2} \tag{1.26}$$

where *k* is a parameter depending on the coherence of the light source. Based on Equations 1.25 and 1.26, we can rewrite Equation 1.22, describing the resist profile, in identifiable parameters, as:

$$\frac{dz}{dx} = \frac{2NA}{\lambda(a+\alpha t)\left[1-k\left(\frac{DOF \times NA^2}{\lambda}\right)\right]^2} \tag{1.27}$$

It follows that, for a steeper edge slope, *dz/dx*, the thickness and the absorption of the resist should be reduced, and the resist contrast γ_p should be as high as possible; finally the shorter the exposure wavelength λ and the higher the NA, the sharper the image profile. Because a higher NA means a lower DOF, which deteriorates image profiles at larger depths, there is an optimum NA for each exposure/resist system.

When imaging the grating of Figure 1.55 with a period p = 2b, the image intensity will be sinusoidal [$D \approx 1/2\ (1 + M \cos 2\pi x/p)$], where *x* is the lateral dimension, and *M* is the normalized modulation. The maximum image edge slope is then simply $dD/dx = \pi M/p$, and Equation 1.27 may then

be written with the maximum of the *dz*/*dx* term expressed as a function of the normalized modulation *M*, or:

$$\frac{dz}{dx} = \frac{\pi M}{2(a+\alpha t)D_p b} \qquad (1.28)$$

From this expression, we would want to work with the largest possible image modulation *M*.[63] Equation 1.27 also implies that a high-contrast resist with a low absorbency will have wider exposure latitude and can tolerate a larger intensity variation in exposure system output (D_p).

The image bias or error, ΔW, shown in Figure 1.63, is primarily a function of the resist edge slope and grows smaller as *dz*/*dx* becomes steeper. At a high radiation dose (>100 mJ/cm^2), the profile of a positive resist is dominated by the absorption of radiation, the reflected photons, and the quantum yield [Φ or $G(s)$] of the photochemical reaction (see Table 1.9Aa). The absorbed energy per depth *z* dominates the rate of dissolution. In a developer-dominated (also force-developed) resist profile (also overcut), the profile recedes, and thinning of the whole resist layer occurs (see Table 1.9Ac). This thinning does not occur in the case of insolubilized traditional negative resists, rendering those resists less susceptible to overdevelopment (see Table 1.9B). For traditional negative resists, γ_n is not influenced by the solvent, and apart from some uncontrollable swelling of negative resists during development, the resist profile cannot be manipulated in the development step. By using a normal exposure dose with moderate developer influence, a "perfect" image transfer can be accomplished with a positive resist as shown in Table 1.9 Ab. At those moderate doses, both the developer elution term and the energy absorption term contribute to the formation of the resist profile.

From Figure 1.63, we can see how the exposure gradient results in the developed positive resist profile having a larger opening on top, forming a so-called normal or overcut profile. This profile results in a difference of dimension in subsequent patterning steps over features with different topography. The change in CDs, ΔCD, relates to profile angle θ and topography height *h*, and can be expressed as:

$$\Delta CD = L_1 - L_2 = 2h\,(\tan\theta)^{-1} \qquad (1.29)$$

A lithography simulator such as PROLITH/2 (http://www.finle.com/index.asp) or Sample (http://cuervo.eecs.berkeley.edu/Volcano/docs/sample3D.html) may be used to perform resist profile simulations.[64] Experimentally determined parameters are fed into the simulated models to make the predicted resist profiles more accurate. An example is the measurement of deprotection reaction parameters of chemically amplified (CA) resists.[65] The concentration of the protection groups determines the solubility of the resist, and consequently the simulation parameters related to the deprotection reaction will have a significant influence on the accuracy of profile simulations of CA resists. The measurement is carried out by incorporating a PEB system inside a Fourier transform infrared (FTIR) spectrometer.

Planarization

Introduction

Microlithography process latitude can become severely limited by nonsmooth topography. Resist films crossing over steps have their local thickness altered; thinning occurs over high features, and pileup happens in low-lying regions. During exposure, the thin resist may get overexposed and the thick regions underexposed. Moreover, resist pileup regions may exhibit/standing wave effects (see below), leading to resolution losses. Reflective notching on buried topographical features could decrease resolution even further (see Figure 1.64). This rough topography creates problems for uniformity of subsequent layers, creating a need to planarize surfaces after each new layer. As seen above, practical DOF should encompass device topography, resist thickness, wafer flatness, focus, and tilt errors for the projected image to remain sharp over the whole wafer. Shallow DOF of high NA-optical-exposure tools, standing waves, and reflective notching can be solved by using very thin resist layers on planar substrates or by planarizing resist layers on nonplanar substrates. Chip makers stack metal lines four layers high, and five- and six-level devices are possible. These "high-rise" chips need rigorous interlevel planarization. The SEM image in Figure 1.65a is a cross-section of a modern IC device showing the various metallization lines (yellow) embedded

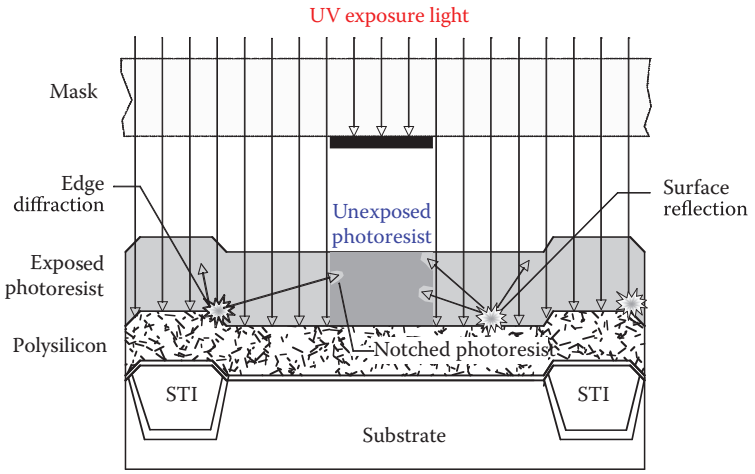


FIGURE 1.64 Reflective notching degrades resist profile. STI, shallow trench isolation.

in passivation material (e.g., silicon nitride). In Figure 1.65b we show an SEM section through 4-MB dynamic random access memory (DRAM) cells. The section was made by using a focused ion beam microscope a technique capable of performing corrective microsurgery on IC chips. The extremely high features often encountered in micromachining pose further challenging planarization problems. Strategies often rely on an interplay of deposition and removal of sacrificial layers (e.g., resist); some of the terminology used in this section will become better understood after reading Chapter 3 on dry etching and Chapter 7 that covers additive processes.

Planarization Strategies

Planarization provides a means of having very flat wafers even after previous patterning steps, and allows for very high-resolution (high R) lithography with limited depth of field (low DOF). Planarization methods provide degrees of smoothing rather

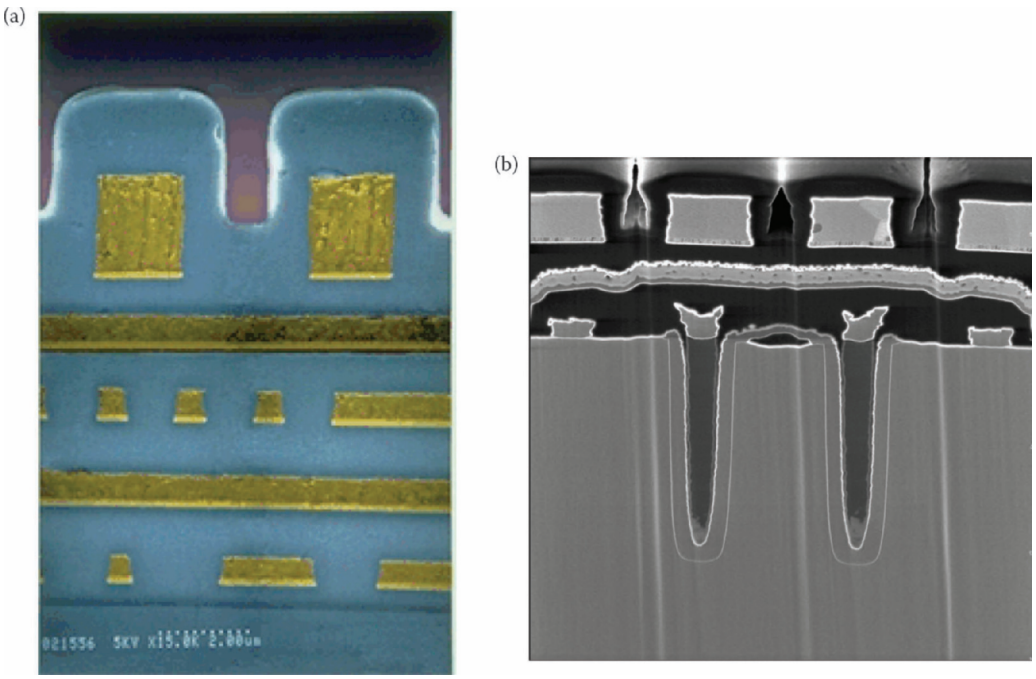


FIGURE 1.65 (a) SEM image of a cross-section of a modern IC device showing the various metallization lines (yellow) embedded in passivation material (silicon nitride, for example. (Courtesy of IBM.) (b) SEM section through 4-MB DRAM cells. The section was made using a focused ion beam microscope. (Courtesy of FIB International.)

than true absolute flatness. In IC fabrication, the chemical-mechanical polishing (CMP) process is used for interlevel dielectric layers (IDL), shallow trench isolation (STI—see Figure 1.64), and copper metallization (the "damascene" process).

A major cause for nonsmooth topography is chemical vapor deposition (CVD) of dielectric layers. CVD methods experience difficulty depositing thin layers conformably. More film tends to build on flat than vertical surfaces, and the pronounced tendency for narrow spaces to be covered (shallow trench isolation) before being completely filled in creates voids. Smoothing in this case should entail two components: filling in gaps and planarizing a film's top surface.

In the planarizing etch-back process, hilly contours left behind by, for example, a CVD oxide deposition technique, are planarized by spin-on glasses (SOGs) or resist sacrificial layers, after which both sacrificial layer and oxide are etched back simultaneously. As its name implies, a SOG dielectric is spun onto a wafer as a liquid and then cured at elevated temperatures. The etch-back process is adjusted, usually by adding oxygen to the etch gases to cause the sacrificial layer and the underlying film, principally oxide or other dielectrics such as PECVD silicon nitride, to etch at equal rates (see Chapter 3). In the etch-back planarization, photoresists are more popular than SOG because they tend to provide a more planar result.[66] SOG often is used in the dielectric sandwich approach for narrow gap filling. Using many thin dielectric layers piled on top of one another until the narrow spaces between metal lines are completely filled averts voids. In this sandwich, layers of CVD oxide alternate with coats of a SOG dielectric. Etching back with photoresist may then planarize the top surface of the film.

There are two types of SOGs: silicates (inorganic) and siloxanes (organic). The film properties of both resemble those of low-temperature CVD glasses. Silicate films tend to absorb water, resulting in a higher relative dielectric constant. In both cases, films tend to crack because of tensile strength as they are densified at 300–500°C. Film stress may be as high as 200 MPa in tension, but annealing lowers tension considerably. Silicate glasses are more prone to cracking, but doping with phosphorus makes them more crack resistant. Even phosphosilicate glasses can only be applied in layers of 1000 Å; thus, often two or three coats are required. It is easier to apply siloxane in a thicker film, for example, greater than 3000 Å. Hitachi Chemical Company America, Ltd. (http://www.hitachi.com) has developed an organosiloxane SOG able to deposit films thicker than 1 mm. As gaps in shallow trench isolation become increasingly narrow, the gap-filling to planarize a surface becomes more difficult. At gap spacings less than 0.3 μm, current SOGs fail to fill the gap completely and are prone to leave a small circular or teardrop-shaped void. Such voids appear with increasing frequency as the gap size decreases and as the gap aspect ratio increases. The effect is most pronounced with re-entrant angles.[67]

Replacing the oxides with a type of dielectric that fills and planarizes much the same as a liquid film may help to eliminate the severe topologies that CVD oxides create. Polyimides fit that dielectric by forming crack-free films that have virtually no pinholes, absorb stress, and generally exhibit a much lower dielectric constant than oxides. Currently, polyimides that will planarize a large pitch geometry to a 90% level with a single coat are available. Reluctance to expand the use of polyimide at this point is linked to its poor barrier properties toward moisture and ions. Also, the prevailing tendency is to trust inorganic CVD oxides more than organic polyimides.[66]

Because the electrical properties of SOGs are poor compared with CVD silicon oxides, an ideal solution would be if the CVD process itself filled the spaces between metal lines and planarized the top surface without breaking the vacuum within a single processing chamber. One example of this type of processing is electron cyclotron resonance CVD (ECRCVD) (see Figure 7.45 illustrating planarization using ECRCVD). ECRCVD uses ECR to generate high-density plasma that can deposit CVD SiO_2 at high rates while temperatures and pressures remain low. As deposition proceeds, the wafer is radio frequency-biased so that argon atoms from the plasma can simultaneously sputter-etch the substrate. The sputtering keeps submicrometer spaces open until they are filled. By carefully balancing the

two processes, one can deposit a film and sputter it back in such a way that high-aspect-ratio spaces are completely filled while the oxide surface planarizes. If planarization needs improvement, ECRCVD can be combined with an etch back.

CMP combines mechanical action with chemical etching; it typifies a planarization technique where a wafer with an uneven surface is polished in an abrasive suspended in an alkaline solution on a polishing pad (see Figure 1.66). Material removal rate is controlled by the slurry flow and pH, applied pressure on the polishing head, rotational speed, and operating temperature. Although slow (with removal rates less than 100 nm/min), CMP results in large surfaces with a roughness less than 1 nm.

The CMP process has been around for a long time but became popular only in the 1990s when it became one of the crucial steps in the manufacturing of IC chips. The reason CMP became crucial is because of its ability to achieve defect-free global planarization of surfaces. The chemicals in the slurry have a "softening" effect on the surface to be polished, thus reducing damage when the abrasives abrade the surface. Oxides, polysilicon, and metal topography can be planarized this way. For oxide or silicon polishing, an alkaline slurry of colloidal silica (SiO_2 particles in a KOH solution or NH_4OH) is continuously fed to the pad/wafer interface. CMP is also popular with micromachinists. For example, Sniegowski[68] is using CMP to enhance the manufacturability of polysilicon surface micromachinery.

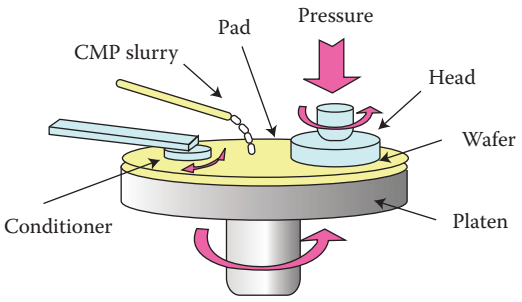


FIGURE 1.66 Schematic of chemomechanical-mechanical polishing (CMP). The CMP process involves pressing the face of the wafer to be polished against a compliant polymeric polishing pad and generating relative motion between the two surfaces. A slurry consisting of abrasives and chemicals is fed in between the interface between the wafer and the pad.

CMP planarization alleviates processing problems associated with the fabrication of multilevel structures, eliminates design constraints linked with nonplanar topography, and provides an avenue for integrating different process technologies. A typical CMP setup looks very similar to a lapping machine, but the precision is much higher.

The damascene process (named after an art form from the Damascus, Syria area) for inlaid metal objects (see Chapter 8, Figures 8.41 and 8.42) used for copper metallization is also dependent on CMP as one of its key steps.

A thin film of Au, Al, or Cu can effectively planarize when melted briefly with a laser. Planarization occurs rapidly because of the high surface tension and low viscosity of clean liquid metals.[69]

Photolithography Resolution Enhancement Technology

Introduction

Since the mid-1980s, the demise of optical lithography has been predicted as being only a few years away, but each time a new resolution limit approaches, some new method extends the useful life of the technology. The three main technologies involved in printing ICs and other miniaturized devices are resist technology, mask technology, and the exposure tool, all of which need to be addressed to optimize lithography resolution.

From the previous section, we learned that optimized production lithography uses projection printing operated close to the conventional Rayleigh diffraction limit with features about the size of the wavelength of the exposure source printed. In this section, we learn about resolution enhancement technology (RET) enabling subdiffraction printing through improvements in resist strategies, mask engineering, and exposure tools. However, the cost of RET is often prohibitive, and in the past, printing with shorter wavelengths has proven more economical.[25] But with CDs of 65 nm, one is forced to rely on RET. Miniaturization science is well suited to contribute to the development of new RET techniques, especially in the area of new mask engineering.

Strategies for Improved Resolution through Improved Resist Performance

Chemically Amplified Resists

Quantum yields for typical positive-tone resists are 0.2–0.3; for negative-tone resists they are 0.5–1 (see Table 1.6). As several photons are required for one useful scission (positive resist) or polymerization (negative) event, this places a fundamental limit on the photosensitivity of these resists. The situation is worse at shorter exposing wavelengths. Typical near-UV positive resists (e.g., DNQ-sensitized novolak resins) are not useful for deep UV lithography because of the strong unbleachable absorption at less than 300 nm (reminder: bleaching is the decrease in optical density during exposure enabling further penetration of the resist as the top layers are being cleared). The short wavelength absorption coefficients of the base resin and the added sensitizers are too high to allow uniform imaging through practical resist thicknesses (0.5–1 μm for ICs).

In the late 1970s, commercially available resists performed poorly at short wavelengths; those with a good plasma etching resistance, frequently based on novolak resins, had excessive unbleachable absorption, whereas those with an acceptable transparency, such as methacrylate-based resists, had insufficient etching resistance. This is of even greater concern in miniaturization science where often much thicker resist layers are used. Short wavelengths do not penetrate thick resist films completely, and very long exposure to dry or wet etchants, needed to create the desired features, degrades the resist. An additional problem, touched on before, is that deep UV optics absorb more light, thus making less light available at the resist plane.

To circumvent the intrinsic sensitivity limitation of low quantum yield resists, chemical amplification was proposed in 1980 and developed at IBM (http://www.ibm.com) in the early 1980s.[37,70–72] In this approach, a single photon initiates a cascade of chemical reactions of the sort that characterizes a silver halide photographic emulsion system. This amplification or gain is based on the photogeneration of a catalytic photoproduct, often an acidic species that catalyzes the scission of the base resin for a positive-tone resist or the cross-linking of the resin for a negative-tone resist. The overall quantum efficiency of the catalyzed reaction is higher than the efficiency for the initial acid generation.[73] Within a resist, which has been sensitized with an onium salt such as diphenyliodonium hexafluoroarsenate, a Lewis acid is released on photolysis. Compared with conventional free-radical initiators, the onium salts have excellent thermal stability and are not sensitive to oxygen.[29] In the case of a positive resist, the released acid, on baking at 100°C, may catalyze the cleavage or scission of the resist, making it more soluble. This type of imaging involves the usual formation of a latent image during exposure. However, the latent image, a 3D distribution of the catalytic photoproduct, does not immediately generate a concomitant change in dissolution rate, as do regular resists. Image formation only takes place after an activating thermal step—to diffuse the photogenerated acid (PGA) and complete the reaction—in a postexposure bake (PEB) at 100°C. A typical turnover rate for one acid catalyst molecule is in the range of 800–1200 cleavages. Resists amplified this way may attain a photosensitivity of 5 mJ/cm^2 or better. Tenfold sensitivity increases are common. Both the acid-catalyzed scission and cross-linking reactions are highly dependent on the PEB temperature, time, and method. The control of these parameters represents a significant difference with conventional resists. In Figure 1.67, a schematic representation of a generalized chemically amplified resist process is shown, as well as typical chemicals. Depending on the developer being polar or nonpolar, a positive or a negative image can be generated. Their dual-tone nature makes these resist systems more flexible. Chemical amplification resists are continually improving and today constitute an important foundation for the design of advanced resist systems for use in short-wavelength (<300 nm) lithographic IC technologies and miniaturization science applications. It was the advent of chemical amplification that made 248-nm lithography finally possible.

The pioneering chemical amplification work involved a catalytic deprotection scheme based on poly(tert-butoxycarbonyloxystyrene) and a triphenylsulfonium hexafluoroantimonate onium salt (Figure 1.68), in which a thermally stable, acid-labile

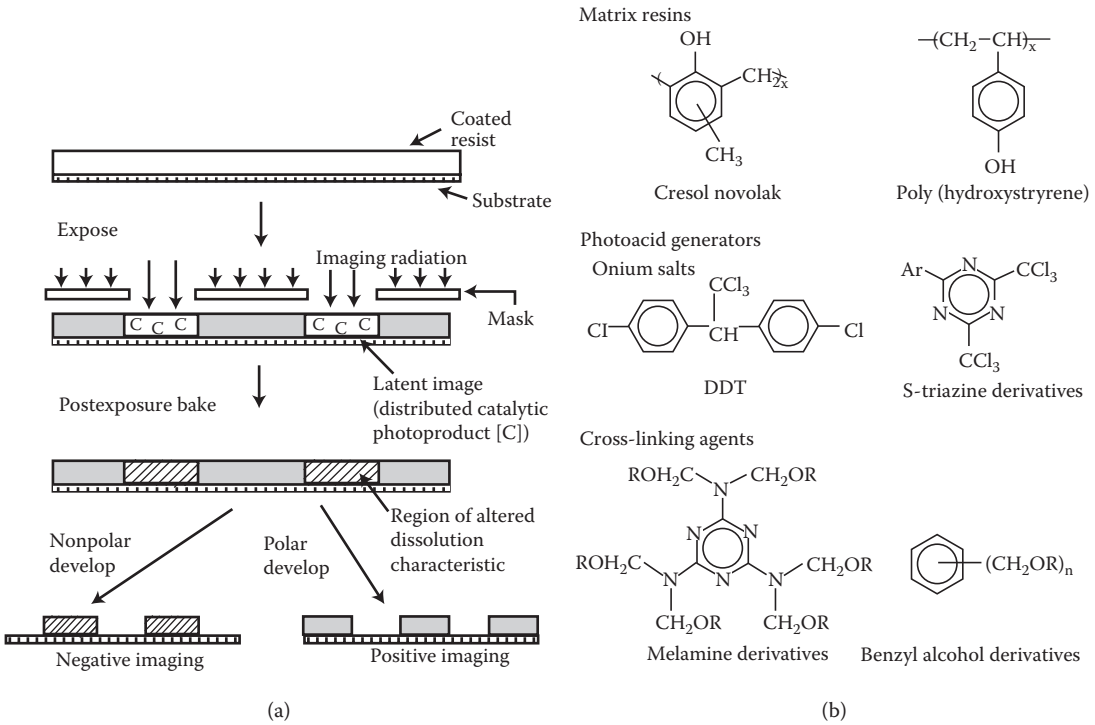
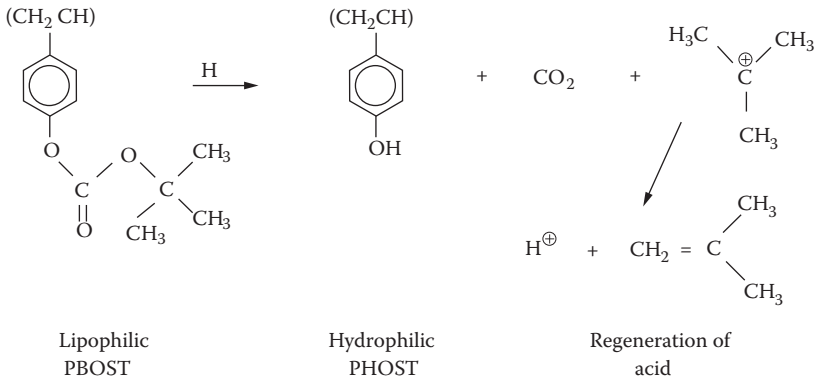


FIGURE 1.67 (a) Schematic representation of a generalized chemically amplified resist process. Depending on the developer being polar or nonpolar a positive or negative image can be created. (b) Typical chemicals used in amplification type resists. [Part (a) based on Reichmanis, E., L.F. Thompson, O. Nalamasu, A. Blakeney, and S. Slater. 1992. Chemically amplified resists for deep-UV lithography: a new processing paradigm. *Microlithogr World* November/December:7–14;[74] Part (b) based on Lamola, A.A., C.R. Szmanda, and J.W. Thackeray. 1991. Chemically amplified resists. *Solid State Technol* 53:53–60. With permission.[75]]

tert-butoxycarbonyl group (tBOC) is used to mask the hydroxyl functionality of poly(vinylphenol).[24]

In this amplification scheme, development is based on polarity changes rather than molecular weight changes through scission and polymerization. On exposure to deep UV and subsequent baking to diffuse the photogenerated acid (PGA) and complete the reaction, the acid cleaves the labile tBOC protecting group to form a polar polyvinyl phenolic polymer. Deprotection (cleavage) of pendant groups



Acid-catalyzed deprotection for polarity change (tBOC resist); cleavage of a pendant group to convert a lipophilic polymer (PBOCST) to a hydrophilic polymer (PHOST)

FIGURE 1.68 Chemical amplification with a catalytic deprotection scheme based on poly(tert-butoxycarbonyloxystyrene) (PBOCST) and a triphenylsulfonium hexafluoroantimonate onium salt, in which a thermally stable, acid-labile tert-butoxycarbonyl group (tBOC) is used to mask the hydroxyl functionality of poly(vinylphenol). (Reichmanis, E., O. Nalamasu, F.M. Houlihan, and A.E. Novembre. 1999. Radiation chemistry of polymeric materials. *Polym Int* 48:1053–59.[24])

in tBOC induces a polarity change and allows for dual-tone (positive/negative) imaging. By using a nonpolar solvent, the unexposed resist is removed and a negative image formed. A polar solvent is used to remove the exposed area, resulting in a positive image. By the mid-1980s, tBOC was used in the production of millions of 1-MB DRAM devices at IBM.[37]

We encountered dual-tone resists earlier when reviewing the novolak/DNQ system by Rohm and Haas (see Figure 1.8).

The chemical amplification resist family today has become prolific and can be grouped into categories according to imaging mechanisms: deprotection (see tBOC), depolymerization, rearrangement, intramolecular dehydration, condensation, and cationic polymerization (SU-8). Cationically polymerizing monomers such as epoxies and vinyl compounds and condensation reactions between phenol formaldehyde resins and amino-based crosslinkers increase the molecular weight after exposure. Changes in polarity are achieved through the acid-catalyzed deprotection of a variety of esters.

Although chemically amplified resists provide reasonable throughput for e-beam and deep UV lithographies (short wavelengths), they often suffer from poor environmental stability and too high a sensitivity to processing conditions. Environmental effects (oxygen) cause surface inhibition by neutralization of photogenerated acid in the top resist layer. A resist topcoat can be used to minimize this effect. Unfortunately, such a multilayer resist further complicates the photoresist process.

Antireflection Coatings: Thin Film Interference Effects

When using single-layer resists, the deleterious effects of thin film interference effects need to be overcome for good CD control, especially of submicrometer features. The most important thin film interference effects are schematically illustrated in Figure 1.66. The effects manifest themselves as nonvertical resist profiles, line-width variations, reflective notching, scumming (underexposed resist leaving organics behind after development), and alignment inaccuracies. Solutions to thin film interference effects involve the use of antireflective coatings and the use of MLRs (see below). Antireflective coatings overcome interference effects caused by reflections from either the top or bottom of the resist (Figure 1.69a and b). Example antireflective coatings are the ARC coatings (trademark from Brewer; http://www.brewerscience.com), liquid resins that are cast into a film before coating the photosensitive resist layer. A key design element is for the cast film to absorb light very efficiently at specific wavelengths. In designing the light-absorbing properties of ARC coats, two strategies were pursued. One is the mixing of dyes with a polymer resin. The second is to design the polymer with molecular chains that intrinsically absorb light.

When a resist is exposed to monochromatic radiation, standing waves are formed in the resist as a consequence of coherent interference from reflecting substrates creating a periodic intensity distribution in the direction perpendicular to the surface (see Figures 1.69c and 1.70). The standing wave effect leads to light intensity variations, perpendicular to the resist film, of as much as a factor of three from one position to another. The standing wave effect is a strong function of resist thickness, and exposure variations resulting from variation in resist thickness in the vicinity of steps result in changes in

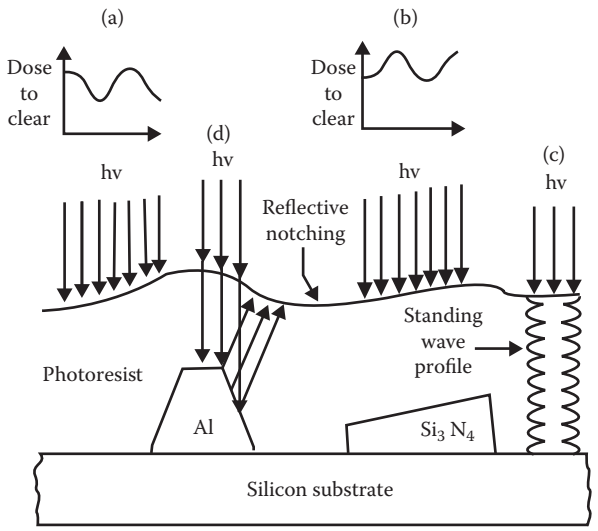


FIGURE 1.69 Thin film interference effects. Light reflected from resist/Si (a) or Si_3N_4/Si (b) interfaces through various resist or nitride thicknesses changes the dose necessary to clear the resist. Also, standing-wave profile (c) and reflective notching effects (d) are shown. (From Horn, M.W. 1991. Antireflection layers and planarization for microlithography. *Solid State Technol* 31:57–62. With permission.[76])

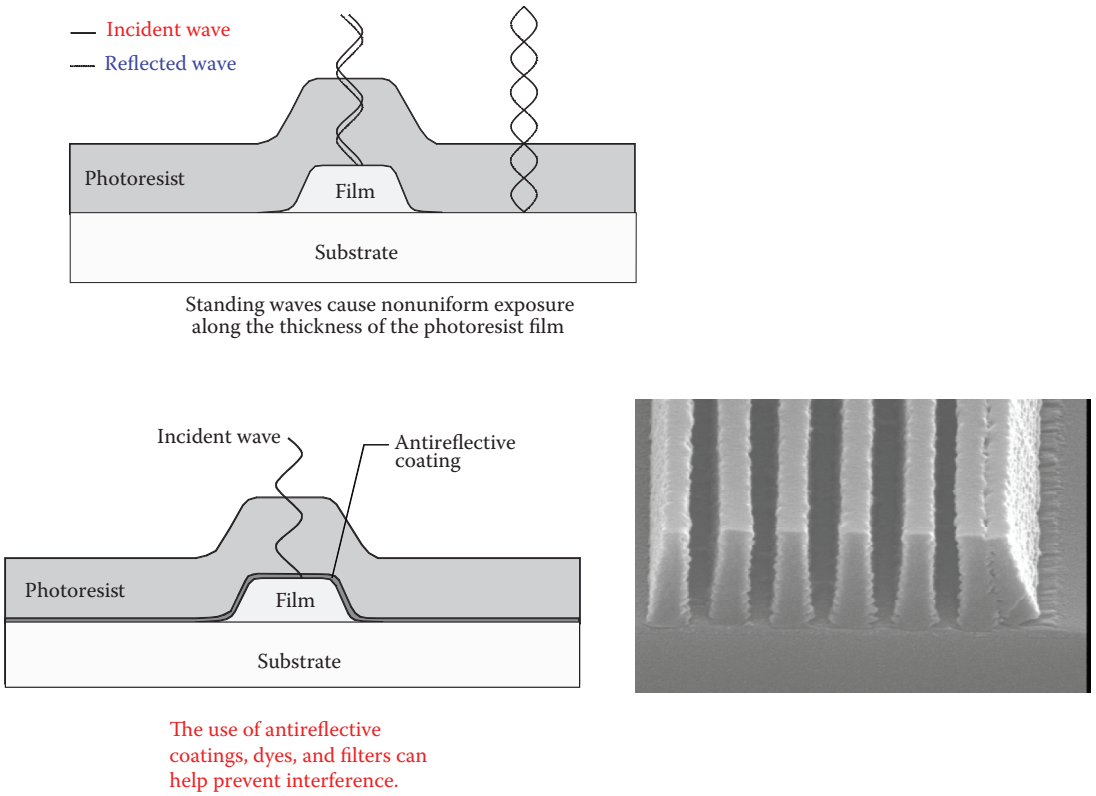


FIGURE 1.70 Standing waves and antireflective coatings.

line width. Because standing waves are the result of coherent interference of monochromatic light, using broadband illumination can decrease the effect. To minimize standing wave effects, one also can use thinner resists (<0.3 μm), but these cause a degradation in line-width control over varying topography. Reflective notching (Figures 1.69d and 1.70) comes about when light is reflected from the more reflective topological features buried in the resist.

Thin Film Imaging

Introduction Variations in resist thickness over steps lead to line-width control problems (see Figure 1.71). This lack of line-width control and the limitations of low source intensity and shallow DOF are overcome in thin film imaging (TFI). In TFI, the projected image is confined to a thin layer near the surface of the resist. Because the thin exposed layer is optically isolated from the substrate, the previously mentioned deleterious effects such as standing waves and scattering from underlying topography are avoided. There is often a need for a thick planarizing resist underneath because of topography and because a very thin film by itself usually does not have the required etch resistance. As a consequence, TFI usually involves a surface modified thick resist in the case of top surface imaging (TSI) or a multilevel resist stack (bilayer, trilayer, or multilayer). After the image has been transferred onto the thin top surface layer, it is transferred to the underlying resist layer(s). Pattern transfer often involves dry etching in oxygen plasma, that is, "dry development." We will discuss single-layer resist TSI first, and then discuss multilayer resist methods.

Thin Film Imaging in Single-Layer Resist—Top Surface Imaging Top surface imaging (TSI) is one approach to thin film imaging (TFI). In TSI, only the top surface of an SLR is modified under light exposure. A single layer of a thick, opaque resist is used, and exposure leads to differential diffusion/reaction rates of a silicon-containing compound into either exposed or unexposed areas (depending on the resist tone). It was Taylor et al.[77] who discovered that small amounts of Si (about 10%) drastically lower the etching rates of organic polymers. Areas containing Si convert to SiO_x in oxygen plasma and thus become plasma resistant.

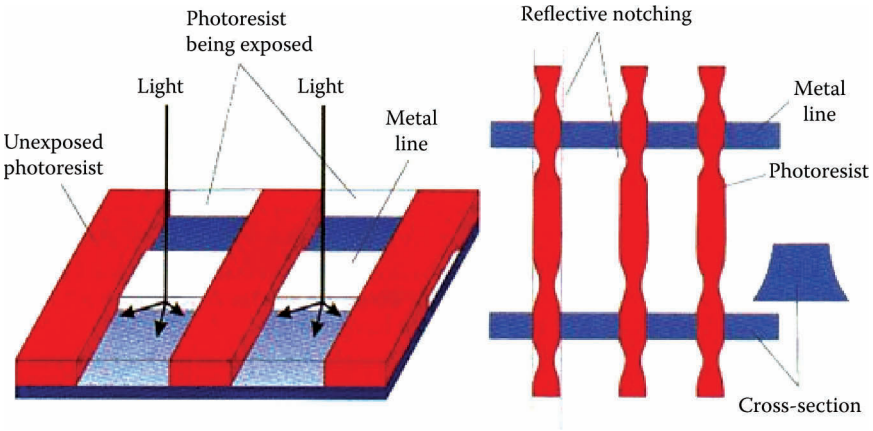


FIGURE 1.71 The dimensional variation of resist lines going over steps is a well-known phenomenon: reflection of light by the nonvertical sidewalls of the step causes increased incoupling of light into the resist adjacent to the step, leading to a reduction in line width. Similar effects may be observed on reflective materials with a particularly coarse-grained structure.

The DESIRE (dry etching of silylated image resist) process by UCB (Belgium, http://www.ucb.be), illustrated in Figure 1.72, exemplifies a single-layer resist TSI scheme. In this case, a short exposure of a thick layer of planarizing positive DQN-novolak resist produces latent images of indenecarboxylic acid. In a postexposure bake, in a vacuum, the layer is converted into a film enabling selective silylating. In the silylating process, Si atoms only diffuse/react on/into the surface of the exposed resist, locally forming a thin Si-containing resist layer. The polymer film is typically exposed and silylated to a depth of at least 1000 Å to provide sufficient differential plasma etch resistance for pattern transfer. An oxygen plasma etch subsequently removes the unexposed, unreacted resist, and the silylated mask remains because of the formation of a protecting silicon oxide layer; a negative-tone image of the mask results. The DESIRE process in Figure 1.72 illustrates not only dry resist development but also image reversal.

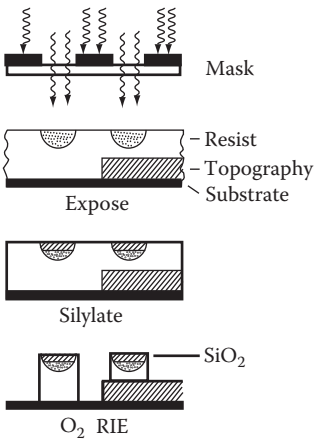


FIGURE 1.72 The DESIRE (dry etching of silylated image resist) process developed by UCB Chemical Company. Exposure and silylation are confined to the upper layers of the resist coating.

A major advantage of TFI techniques is that only the surface of the resist requires exposure, which, coupled with the anisotropic nature of plasma etching, allows for the use of thick resist layers that planarize the underlying topography to a higher degree than resists with more typical thicknesses could. The lithography process also remains unaffected by substrate reflection, and resist thickness variation as light does not have to penetrate the thickness of the resist. Image quality with DESIRE is excellent. An advantage of TSI over bilayer or multilayer TFI techniques is that only one single layer is spun on before silylation. DESIRE made it from research and development into manufacturing in the late 1980s at TI (http://www.ti.com).[78] It was used for the top metal level patterning in DRAM manufacture. This level had the most reflectivity and most topography. The resist that TI applied, using a standard coater, was the Plasmask® 200-g from Japan Synthetic Rubber (JSR, http://necsv01.keidanren.or.jp/A2J/data/06025.html). Two additional machines, a silylation tool and a dry develop etcher, were developed for the TSI process. Both liquid and vapor silylation are feasible and have been compared at IBM.[78] The use of a dry etch in development leads occasionally to

so-called *grass residue* in open areas (see Chapter 3, Figure 3.23). The grass is assumed to stem from sputtered silicon from silylated areas. With higher plasma density systems, less grass forms. This can be explained by the lower energy of the ions attacking the resist surface.

Pure poly(hydroxystyrene) (PHOST) (sensitivity 22 mJ/cm^2) has been studied extensively as a model TSI resist for use with 193-nm optical projection lithography.[79]

Multilayer Resist Thin Film Imaging In the bilayer TFI approach shown in Figure 1.73, the resist stack consists of a thick, planarizing bottom layer and a thin, top imaging overlayer. These layers function synergistically to achieve higher resolution otherwise impossible to obtain with thick, planarizing single-layer resists. After exposure and development of the thin imaging overlayer, the pattern is transferred to the planarizing underlayer. An intermediate isolation layer sometimes separates the planarizing and imaging layers, preventing their mixing. Depending on the exposure dose and development condition (wet etching) of the planarizing resist layer, an undercut or overcut may be generated. Alternatively, if good fidelity of the mask pattern is desired, dry directional etching (RIE) may be used (see Figure 1.73).

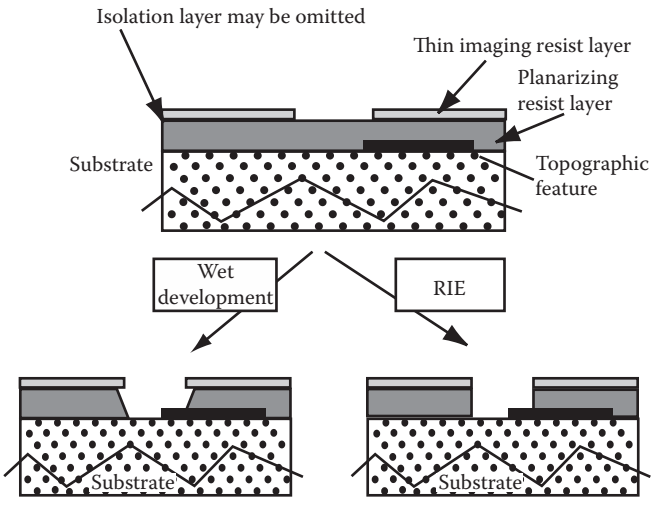


FIGURE 1.73 Multilayer resist scheme. Thick planarizing underlayer for planarizing of wafer topography, optional intermediate isolation layer, and thin imaging overlayer for optimum resolution. Wet etching (development) and dry etching (RIE) produce different resist profiles.

In a trilayer system, a thin imaging layer is spun onto a thin oxygen dry etch-resistant hard mask coated on top of a thick planarizing layer. The hard mask may be a plasma-deposited or spin-on silicon oxide.

With the advent of 193-nm lithography, TFI will become increasingly popular as radiation is readily absorbed by most photoresists at this short wavelength. Once extreme UV is used as the exposure radiation of choice (see below), widespread use of TFI might be unavoidable.[78]

Strategies for Improved Resolution through Improved Mask Technology

Background

Besides improved resists, mask engineering also provides important contributions to RET. Interestingly, Richard Feynman in the second of his now famous lectures on micromachining imagined a micromachined mask with miniature levers and valves that could be used over and over to generate different patterns.[80] The mask techniques described above are timid attempts to use micromachining technology for mask engineering. In one scenario, by controlling the phase and the amplitude of the light at the image plane in a phase-shifting mask, resolution and depth of focus are greatly improved.

In miniaturization science, where we often print high-aspect-ratio features, masks with an expanded DOF are favored—but masks with increased resolution remain of greater importance in IC manufacture.

Phase-Shifting Masks

One method that allows further improvement of photolithography resolution and DOF at a given wavelength (λ) and NA is to carefully control light diffraction using constructive and destructive interference to help create a circuit pattern. The idea of selectively altering the phase of the light passing through certain areas of a photomask to take advantage of destructive interference to improve resolution and DOF in optical lithography was first proposed in 1982.[81] The method is based on building PSMs. Using such a mask one can control both the amplitude and the phase of the light

and, in particular, arrange the mask so that light with the opposite phase emerges from adjoining mask features. Destructive interference can be used to cancel some of the image-spreading effects of diffraction.

In Figure 1.74, we illustrate the effect for three nearby mask apertures; a classical transmission mask is added for comparison.[82] The amplitude profile at the plane of the classical transmission mask consists of three square-cornered features with positive amplitude. The amplitude at the wafer level is broadened and rounded but does not change sign. With a PSM, a shifter layer results in a reversal of the phase of the light at the mask. As the photoresist only shows sensitivity toward the intensity of the light and not to the sign, the three bright features develop identically, and the areas where the light of opposite phases causes destructive interference form very dark contrast lines.

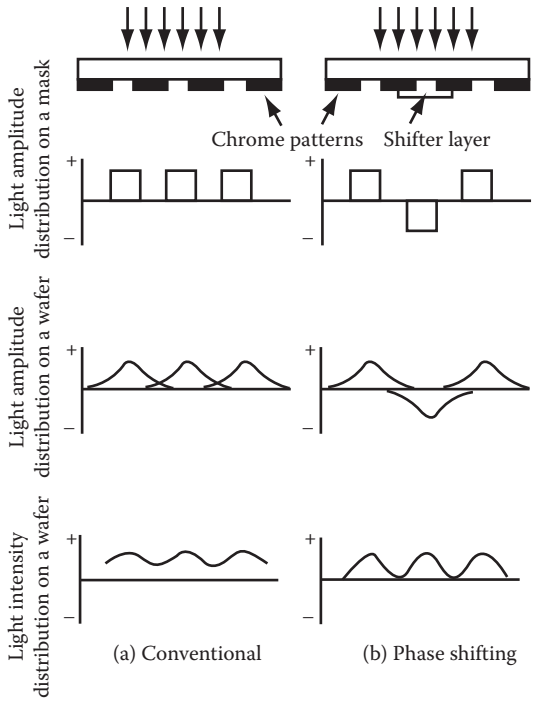


FIGURE 1.74 Comparison of a conventional transmission mask (a) and a "Levenson-type" phase-shifting mask (b). Both contain opaque chrome (Cr) regions, but the light passing through some of the apertures of the phase-shifting mask also passes through the transparent material of the phase shifter. This reverses the phase of the light at the mask, giving rise to destructive interference at the photoresist plane. (From Levenson, M.D. 1992. Phase-shifting mask strategies: isolated dark lines. *Microlithogr World* 6:6–12. With permission.[82])

One type of PSM is an alternating aperture PSM (AAPSM), also called a *hard* shifter. Going beyond the traditional chrome-on-quartz approach, AAPSMs accomplish the task of shifting the phase of the light by etching regions of the quartz substrate to the precise depth (which depends on the wavelength of the light to be used to expose the wafer) where shifting is desired and have become known as Levenson-type PSMs. The etched areas cause the light to become 180° out of phase with the light passing through the unetched regions (Figure 1.75).

An important alternative PSM is the embedded attenuated PSM (EAPSM). EAPSMs, or *soft* shifter masks, are similar to ordinary (binary) masks in that a quartz substrate is coated with a material in which the design is etched. The most common material used today is molybdenum silicide. Unlike chrome, molybdenum silicide allows a small percentage of the light to pass through; however, the amount passing through is "soft" and does not expose the resist on the wafer. Where it does pass through, the light is 180° out of phase compared with light passing through the quartz alone. Where the coating and quartz meet, light interferes in such a way as to sharpen the edges of the design. The most difficult task in the case of EAPSMs is to create a thin film structure with the desired transmission (4–8%) and phase (180°) at a specific wavelength (i-line or deep UV) that can be patterned, inspected, and repaired using currently available tools.[83] It has been

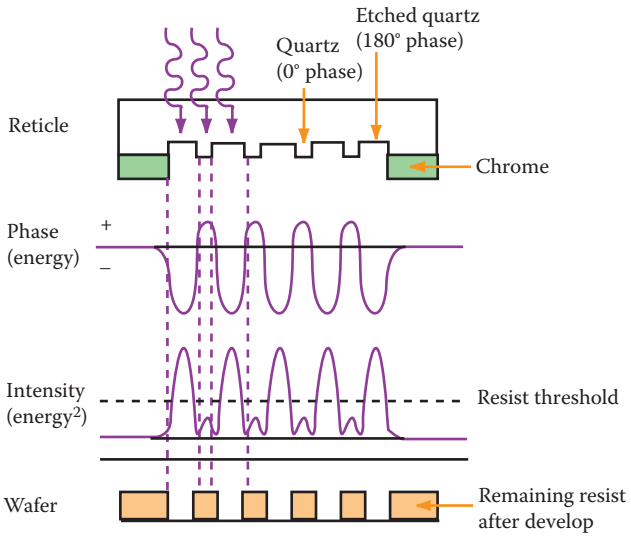


FIGURE 1.75 A hard shifter alternating aperture PSM or AAPSM.

estimated that a PSM mask for 100-nm work could cost up to $50,000. This makes RET often synonymous for Really Expensive Technology!

How narrow a feature can one make using PSMs? Oki Electric Industries (http://www.oki.de) sells a metal semiconductor field effect transistor (MESFET) device with a 0.18-µm gate patterned by i-line lithography, ordinarily considered capable of only 0.5 µm resolution![82] Thus, PSMs can easily improve resolution by 50–100%. Using 248-nm deep UV lithography researchers at MIT's Lincoln Laboratory have reported sub-100-nm features using phase-shifting masks. Because the same method is applicable to 193-nm radiation, the approach appears promising for further extending photolithography.[84]

The use of TSI and PSM techniques can reduce k_1 in Equation 1.16 to 0.4, well below the 0.61 expected from Rayleigh's criterion.

Optical Proximity Correction

Optical proximity correction (OPC) printing was first developed in the early 1970s as a means of addressing lithography distortions in IC manufacturing. The goal is to produce smaller features in an IC using current lithography equipment by enhancing the "printability" of a given wafer pattern. OPC applies systematic changes to photomask geometries to compensate for nonlinear image distortions. These distortions include line-width variations (nonuniform CDs), line-end shortening, and rounding of sharp corners. Sharp features, in general, are lost because higher spatial frequencies are lost as a result of diffraction. Causes of these distortions include reticle pattern fidelity, optical proximity effects, and diffusion and loading effects during resist and etch processing. Thus, a mask incorporating OPC is a system that negates undesirable distortion effects during pattern transfer. In OPC, the traditional (WYSIWYG—what you see is what you get) relationship between layout, mask, and silicon is no longer valid. OPC works by making small changes to the IC layout that anticipate these distortions. To compensate for line-end shortening, for example, the line is extended using a hammerhead shape that results in a line in the resist that is much closer to the original intended layout. To compensate for corner rounding, serif shapes are added to (or subtracted from) corners to produce corners on the silicon that are closer to the ideal layout (see Figure 1.76). Determining the optimal type, size, and symmetry (or lack thereof) is very complex and depends on neighboring geometries and process parameters. A sophisticated computer program is necessary to properly implement OPC. OPC improves the resolution by decreasing k_1.

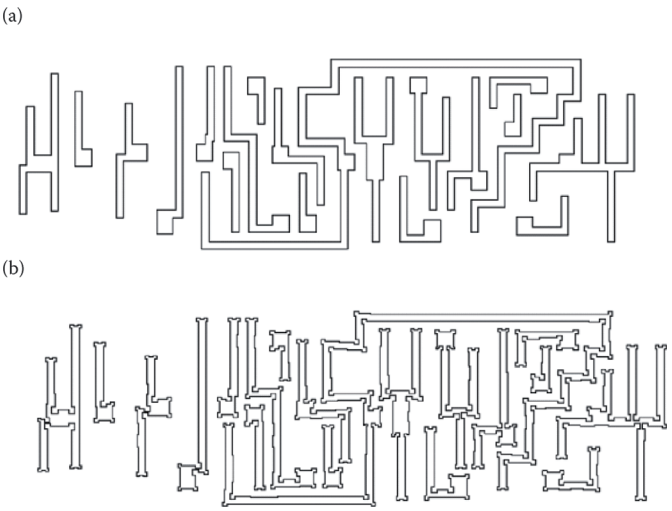


FIGURE 1.76 Optical proximity correction (OPC). (a) Mask of a standard cell before OPC. (b) Mask after OPC.

Strategies for Improved Resolution through Improved Exposure Technology

Background

Under "Improved Exposure Equipment," we describe progress in decreasing k_1 and NA in Equation 1.16 to further enhance resolution R by modifying the exposure equipment. Off-axis illumination (OAI) is one of the three major resolution enhancement technologies that have enabled optical lithography to push practical resolution limits far beyond what was once thought possible (the others being phase-shifting masks and optical proximity corrections; see above). We will briefly review OAI and Köhler illumination and finish this chapter by introducing immersion lithography.

Off-Axis Illumination and Köhler Illumination

With OAI, diffraction information is distributed selectively over the lens pupil, influencing the impact

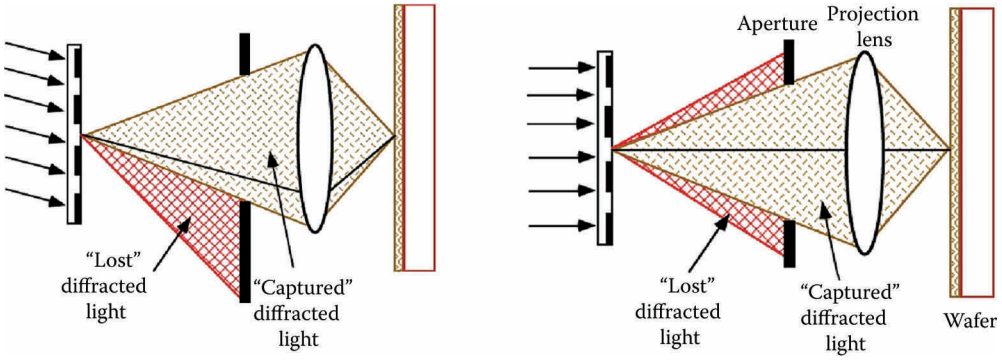


FIGURE 1.77 Off-axis-illumination (OAI).

of aberrations on imaging. Off-axis illumination allows some of the higher order diffracted light to be captured and hence can improve resolution (by decreasing k_1). OAI refers to any illumination shape that significantly reduces or eliminates the "on-axis" component of the illumination, that is, the light striking the mask at near-normal incidence. By tilting the illumination away from normal incidence, the diffraction pattern of the mask is shifted within the objective lens. This enables higher diffracted orders to be included in the lens aperture, which increases the resolution (see Figure 1.77). The increased focal depth for OAI can be understood in terms of the path difference, which is now shorter for all the waves arriving at the resist surface. OAI can improve imaging resolution by 20% and DOF by 40%.

Köhler illumination, named after August Köhler, the man who invented it in 1983, is also known as double diaphragm illumination because it uses both a field and an aperture iris diaphragm to set up the illumination in microscopy. If the light path is set up properly, one has the advantages of an evenly illuminated field, essential to avoid shadows, glare, and incorrect contrast. Köhler illumination in lithography achieves the same goals by using a collector lens in front of the light source (see Figure 1.78). This "captures" diffracted light equally well from all positions on the mask. This improves the resolution by bringing k_1 down.

Immersion Lithography

From the discussion up until this point, the resolution limit for 193-nm exposure systems may be calculated using the Rayleigh equation with $\lambda = 193$ nm, NA = 0.93, and $k_1 = 0.25$ (see Equation 1.16) as:

$$R = 0.25 \times 193/0.93 = 52 \text{ nm} \qquad (1.30)$$

From this calculation, a highly optimized ArF exposure system has an absolute optimal

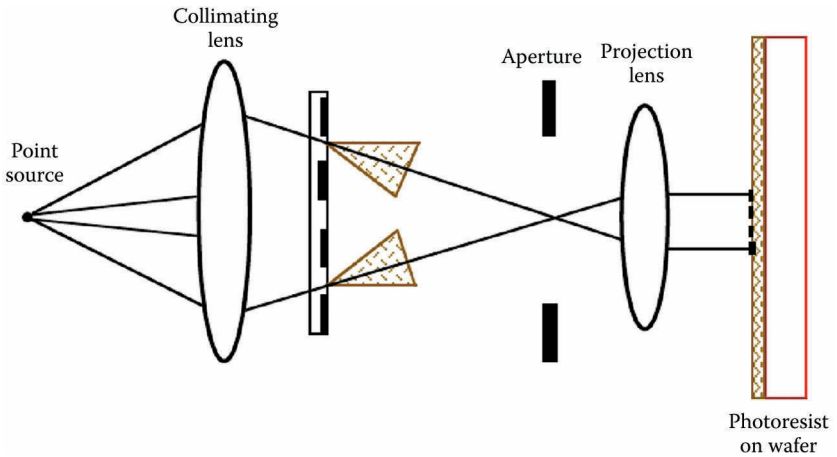


FIGURE 1.78 Köhler illumination in lithography.

resolution of 52 nm, sufficient for the 65-nm line widths introduced in 2005 but not capable of meeting the 45-nm line widths forecast in 2007. The technical challenges with 157-nm and shorter wavelength exposure systems make any technique that can improve the resolution of the 193-nm exposure systems and delay the need to move to shorter wavelengths an important development.

The NA is determined by the acceptance angle of the lens and the index of refraction of the medium surrounding the lens: NA = nsinθ (Volume I, Equation 5.3), where n is the index of refraction and θ is the acceptance angle of the lens. Because the sine of any angle is always <1 and n = 1 for air, the physical limit for an air-based system is clear, but what if a medium with a higher index of refraction is substituted for air? Microscopy has for years used oil as the medium surrounding the lens and the sample for resolution enhancement, and it is somewhat surprising that the semiconductor industry has taken this long to seriously consider the merits of replacing air with an alternative.

The medium between the lens and the wafer being exposed needs to have an index of refraction >1, have low optical absorption at 193 nm, be compatible with photoresist and the lens material, and be uniform and noncontaminating. Surprisingly, ultrapure water may meet all these requirements. Water has an index of refraction n = 1.47, absorption of <5% at working distances of up to 6 mm, is compatible with photoresist and lens, and in its ultrapure form is noncontaminating. Plugging n = 1.47 into Equation 5.3 (Volume I) and assuming sinθ can reach 0.93, then the resolution limits for 193-nm immersion lithography are:

$$R = 0.25 \times 193/1.470.93 = 35 \text{ nm} \qquad (1.31)$$

A 35-nm theoretical resolution carries 193-nm exposure beyond 2007. Similar techniques applied to 157-nm exposure could carry optical lithography even further, although it should be noted that water is not a usable medium at 157 nm and suitable mediums are still being researched (a high index fluid such as polyfluoropolyether is a candidate medium in this case). The resolution enhancement from immersion lithography is about 30–40% (depending on materials used), or about one technology node. The depth of focus is also 40–70% better (proportional to the refractive index of the imaging medium considered) than a corresponding "dry" tool at the same resolution.[85]

In Figure 1.79 we illustrate a schematic of an immersion lithography setup. It is estimated that such a setup may cost $30 million.

There are a large number of practical issues yet to resolve to implement immersion lithography. The stage on a 193-nm exposure tool steps from location to location across the wafer scanning the reticle image for each field. To achieve high throughput, the stage must accelerate rapidly, move accurately to the next field location, settle, scan the image, and then step to the next location all in a short period. Maintaining a consistent bubble-free liquid between the lens and the wafer is very difficult. One other issue that is likely to be significant for immersion lithography is temperature control. Variations in temperature cause variations in n and therefore image distortion. Maintaining temperature uniformity with a rapidly moving stage and a pulsed laser passing through the fluid will likely be a significant challenge. A more complete listing of issues includes:

- Very big lenses (hence expensive)
- Field size reduction
- Mechanical issues and hydrodynamics
- Stage vibrations transferred to lens
- Heating of immersion liquid on exposure
- New defect mechanisms at wafer level
- Interaction of photoresist with immersion liquid
- Fluid contamination
- Polarization effects degrading contrast

Summary Projection Lithography

As minimum line widths have shrunk, the exposing wavelength has also periodically shrunk. Table 1.10 lists the year, minimum line-width generation, and exposure wavelength for state-of-the-art ICs since the mid-1980s. At 1.2-μm and larger line widths, the g-line output of mercury lamps (λ = 436 nm) was used; at the 0.8-μm (800 nm) generation, the

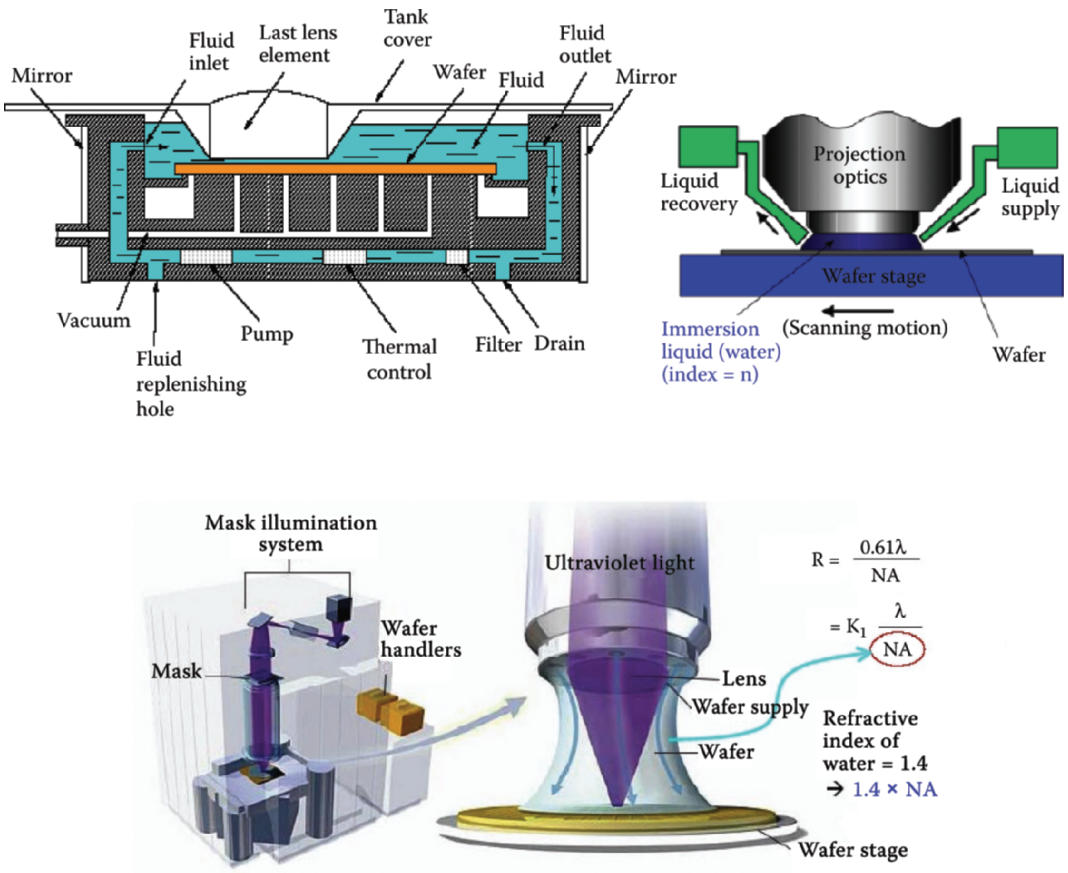


FIGURE 1.79 Schematics of an immersion lithography setup.

i-line output of mercury lamps (λ = 365 nm) was introduced for critical layers, and i-line use continued to the 350-nm line width, where early adopters began to use KrF excimer lasers (λ = 248 nm) as the exposure source. KrF use has surprised many observers by persisting through the 130-nm linewidth generation. With 90-nm line widths now entering production, KrF is finally running out of stream, and ArF excimer lasers are being introduced (λ = 193 nm). Beyond ArF, there are fluorine excimer lasers (F_2) with λ = 157 nm, but there are a number of technical challenges to overcome. At less than the 157-nm wavelength, the optical exposure systems must change to all reflecting optics because of high levels of absorption in refractive lens at shorter wavelengths. The introduction of an all-reflective lens exposure system introduces a number of technical challenges.

TABLE 1.10 Minimum Line Width and Exposure Wavelength versus Year

Year	Line Width (nm)	Wavelength (nm)
1986	1200	436
1988	800	436/365
1991	500	365
1994	350	365/248
1997	250	248
1999	180	248
2001	130	248
2003	90	248/193
2005 (fcst)	65	193
2007 (fcst)	45	193

At the same time that exposure wavelengths have been reduced, improvements in lens design have led to improvements in the NA of the exposure systems lens, as shown in Figure 1.80. In the mid-1980s an NA value of approximately 0.4 was typical; today 248-nm exposure systems are available with an NA greater than 0.8. The physical limit to NA for exposure systems using air as a medium between the lens and the wafer is 1; the practical limit is somewhere around 0.9, with recent reports suggesting that an NA as high as 0.93 may be possible for ArF systems in the future.

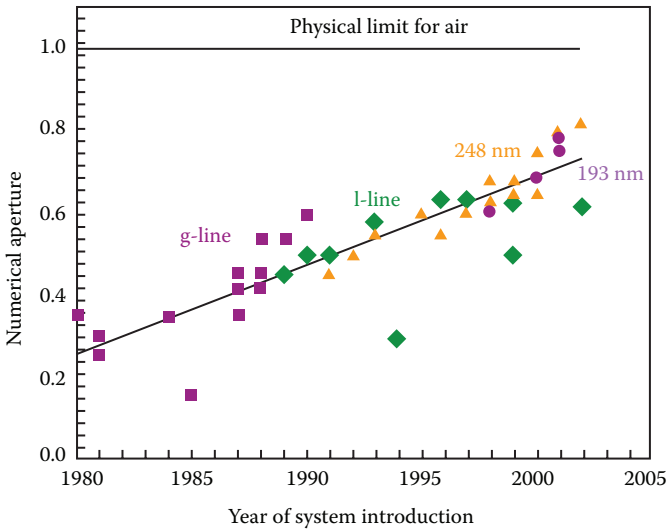


FIGURE 1.80 The evolution of the numerical aperture in lithography systems as a function of time.

The third element in the Rayleigh equation is k_1, a complex factor of several variables in the photolithography process such as the quality of the photoresist and the use of resolution enhancement techniques such as phase shift masks, OAI, and OPC. While exposure wavelengths have been decreasing and NA has been increasing, k_1 has been decreasing as well (see Figure 1.81). The practical lower limit for k_1 is thought to be about 0.25. In Figure 1.81, we plot the evolution of k_1 during the past 18 years.

From this chapter it is clear that photolithography remains the only viable IC, MEMS, and NEMS production technology. As we realize now it is very urgent that we look beyond photolithography for the next generations of ICs and NEMS. This is the problem we tackle in Chapter 2 on NGLs and lithography research.

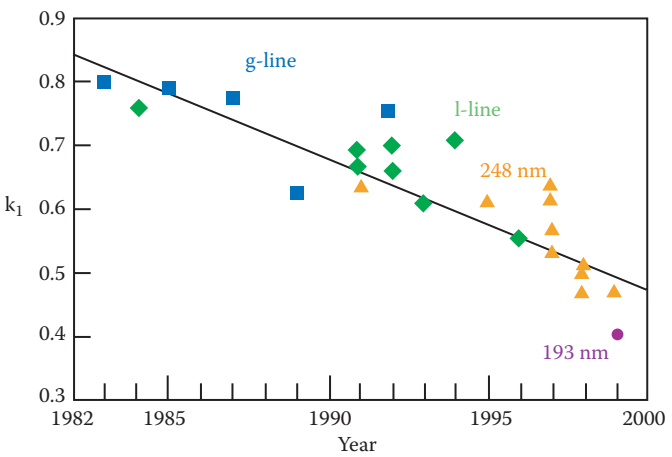


FIGURE 1.81 The evolution of k_1 versus time.

Acknowledgments

Special thanks to Maria Amasia and Genis Turon.

Appendix 1A: Spin Coating Process Troubleshooting

Film Too Thin

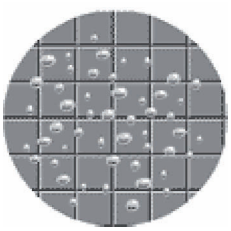
Spin speed too high: select lower speed. Spin time too long: decrease time during high-speed step. Inappropriate choice of resin.

Film Too Thick

Spin speed too low: select higher speed. Spin time too short: increase time during high-speed step. Exhaust volume too high: adjust exhaust. Inappropriate choice of resin.

Air Bubbles on Wafer Surface

Air bubbles in dispensed fluid. Dispense tip is cut unevenly or has burrs or other defects.

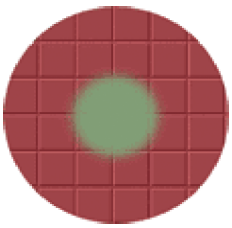


Swirl Patter

Spin bowl exhaust rate is too high. Fluid strikes substrate surface off center. Spin speed and acceleration setting is too high. Spin time is too short.

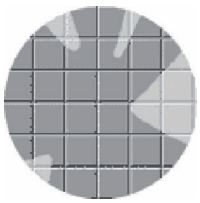
Center Circle

If the circle is the same size as the spin chuck: proper chuck selection should be based on substrate size and rigidity. The proper chuck diameter is 1/4–1-in. smaller than the substrate diameter.



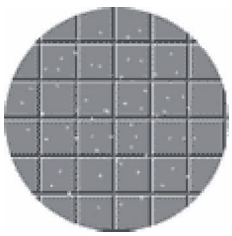
Uncoated Areas

Insufficient dispense volume.



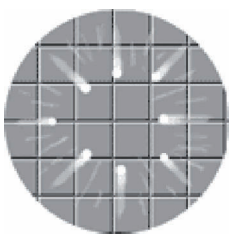
Pinholes

Air bubbles. Particles in fluid. Particles exist on substrate surface before dispense.



Comets, Streaks, or Flares

Fluid velocity (dispense rate) is too high. Spin bowl exhaust rate is too high. Resist sits on wafer too long before spin. Spin speed and acceleration setting are too high. Particles exist on substrate surface before dispense. Fluid is not dispensed at the center of the substrate surface.



Poor Reproducibility

Variable exhaust or ambient conditions: adjust exhaust lid to fully closed. Substrate not centered properly: center substrate before operation. Insufficient dispense volume: increase dispense volume. Inappropriate resin material. Unstable balance in speed/time parameters: increase speed/decrease time or vice versa.

Poor Film Quality

Exhaust volume too high: adjust exhaust lid or house exhaust damper. Acceleration too high: select lower acceleration. Unstable balance in speed/time parameters: increase speed/decrease time or visa versa. Insufficient dispense volume: increase dispense volume. Inappropriate resin material: contact resin manufacturer.

Questions

1.1: Analyze each of the expressions for photolithography resolution and explain how to improve resolution in each case. What are the advantages and disadvantages of using e-beam lithography compared to typical photolithography using UV radiation? What is the most likely next-generation lithography?

1.2: An exposure is performed with coherent light using a step and repeat projection printing system. The light source has a wavelength of 365 nm (i-line of a Hg arc lamp).
The pattern is a grating with a line-to-line spacing of 1 μm.
(a) Calculate the minimum value of the numerical aperture (NA) that will provide contrast at the image plane (the plane of the resist).
(b) What is the maximum value of the numerical aperture, above which there will be no improvement in image quality?
(c) Calculate the depth of field of the image for cases (a) and (b).

1.3: Lift-off refers to the process of exposing a pattern into a photoresist, depositing a thin film such as a metal or dielectric over the entire area, and then washing away the photoresist

to leave behind the film only in the patterned area. Why is it easier to obtain a lift-off profile with a negative resist than with a positive resist?

1.4: Which of the following statements are NOT correct?

(a) Short exposure wavelengths can create standing waves in a layer of photoresist.

(b) Regions of constructive interference create increased exposure.

(c) Standing waves can impair the structure of the resist, but they can be eliminated by use of multiple wavelength sources or postbaking.

(d) Standing waves effects are most noticeable at the center of the resist.

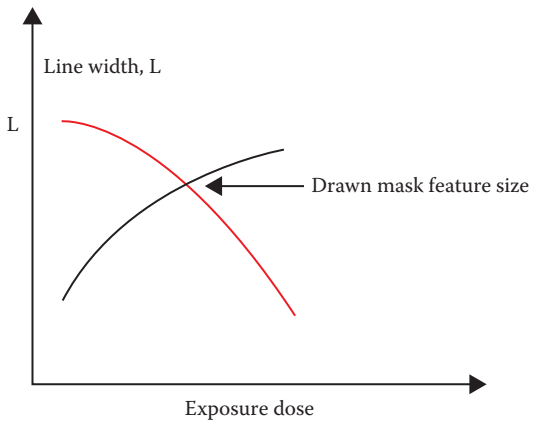
(e) The primary components of a positive photoresist are

1. Nonphotosensitive base phenolic resin
2. Photosensitive dissolution inhibitor
3. Coating solvent

(f) Projection lithography resolution is limited by exposure wavelength, resist thickness, and diffraction and dispersion of light.

(g) Proximity lithography resolution is limited only by exposure wavelength and resist thickness.

1.5: Indicate which line in this graph corresponds to **negative** photoresist and which corresponds to **positive** photoresist.



1.6: Why is it easier to find good data from the literature on resist sensitivity than on lithographic sensitivity?

1.7: Sketch the lithography steps involved in generating a staircase resist pattern that is oriented in the same direction on both sides of a Si wafer. The steps of the staircase need to be aligned within 2 μm and your laboratory cannot afford a >$250,000 double-sided mask aligner.

1.8: Present a comparison of negative and positive photo resists. Also explain what a permanent resist is. Describe what happens chemically to positive and negative resists when exposed to UV radiation.

1.9: A polyimide photoresist requires 100 mJ/cm^2 per micron of thickness to be developed properly. If a lamp provides 1000 W/m^2, how long does it take to expose a 20-μm-thick film?

1.10: List four major factors in which IC technology differs from MEMS or miniaturization science.

1.11: Why does one prefer a high T_g in a resist and why are there no crystalline materials used in lithography?

1.12: Provide a simple rule of thumb for dimensional tolerances in lithography. Why is the resolution with incoherent light larger than that for coherent illumination? How is the depth of focus (DOF) of an imaging system influenced by the numerical aperture NA of the imaging lens, the resolution of the system, and the wavelength of the exposing light?

1.13: Demonstrate with some simple sketches how you would pattern a small Pt electrode for an electrochemical sensor using wet etching. The substrate is an oxidized Si wafer. (Remember that Pt is very difficult to wet etch.)

1.14: Mark the INCORRECT statement(s).

(a) If we want to do the lift-off process, we do not need a postbake.

(b) Before we spin dry a wafer, we need to make sure the vacuum is off.

(c) To increase the pump-down speed for the metal evaporation process, we need to turn on the roughing and foreline pump at the same time.

(d) Ultrasonic vibration can help the lift-off process.

(e) Four-point resistivity measurements can be used to determine the doping concentration of the water.
(f) O_2 plasma etching can be used to remove the photoresist.
(g) After the fabrication is completed, we can deposit a thin layer of Si_3N_4 to protect the Al wires.

1.15: How is the diffraction limit expression derived? Hint: Use the mathematical expression for the intensity variation of the Fraunhofer diffraction pattern of a circular aperture (Airy disk) with radius a.

1.16: How does immersion lithography influence the value of the DOF?

1.17: What are the pros and cons of proximity, contact, and projection lithography?

1.18: Explain what happens chemically to positive and negative photoresist when exposed to UV radiation.

1.19: List the most important methods that can be employed to improve resolution in photolithography.

References

1. Harris, T. W. 1976. *Chemical milling.* Oxford, UK: Clarendon Press.
2. Cunningham, J. A. 1992. The remarkable trend in defect densities and chip yield. *Semicond Int* 15:86–90.
3. Iscoff, R. 1991. Wafer cleaning: can dry systems compete? *Semicond Int* 14:48–54.
4. Bok, E., D. Kelch, and K. S. Schumacher. 1992. Supercritical fluids for single wafer cleaning. *Solid State Technol* 35:117–19.
5. Singer, P. H. 1992. Trends in wafer cleaning. *Semicond Int* 15:34–39.
6. Tolliver, D. L. 1988. *Handbook of contamination control in microelectronics: principles, applications, and technology.* Park Ridge, NJ: Noyes Publications.
7. Deal, B. E. 1993. *Handbook of semiconductor wafer cleaning technology.* Park Ridge, NJ: Noyes Publications.
8. Sure, A., T. Dillon, J. Murakowski, C. Lin, D. Pustai, and D. W. Prather. 2003. Fabrication and characterization of three-dimensional silicon tapers. *Optics Express* 11:555–61.
9. Feely, W. E. 1988. *Technical digest: solid state sensor and actuator workshop.* Hilton Head Island, SC: IEEE.
10. Feely, W. E., J. C. Imhof, and C. M. Stein. 1986. The role of the latent image in a new dual image, aqueous developable, thermally stable photoresist. *Polym Eng Sci* 26:1101–04.
11. Six, P. 1995. Phase masks and grey-tone masks. *Semiconductor Fabtech 5*.
12. Reimer, K., R. Engelke, U. Hofmann, P. Merz, K. T. K. v. Platen, and B. Wagner. 1999. *MF-Conf., micromachine technology for diffractive and holographic optics.* Santa Clara, CA: SPIE.
13. Stauffer, J. M., Y. Oppliger, P. Regnault, L. Baraldi, and M. T. Gale. 1992. Electron beam writing of continuous resist profiles for optical applications. *J Vac Sci Technol* B10:2526.
14. Windt, D. L., and R. A. Cirelli. 1999. Amorphous carbon films for use as both variable-transmission apertures and attenuated phase shift masks for deep ultraviolet lithography. *J Vac Sci Technol* B17:930–32.
15. Arnone, C. 1992. The laser-plotter: a versatile lithographic tool for integrated optics and microelectronics. *Microelectron Eng* 17:483–86.
16. Duffy, D. C., J. C. McDonald, O. J. A. Schueller, and G. M. Whitesides. 1998. Rapid prototyping of microfluidic systems in poly(dimethylsiloxane). *Anal Chem* 70:4974–84.
17. Smith, H. I., R. Menon, A. Patel, D. Chao, M. Walsh, and G. Barbastathis. 2006. Zone-plate-array lithography: a low-cost complement or competitor to scanning-electron-beam lithography. *Microelectron Eng* 83:956–961.
18. Metz, T. E., R. N. Savage, and H. O. Simmons. 1992. In situ control of photoresist coating processes. *Semicond Int* 15:68–69.
19. Thompson, L. F. 1994. *Introduction to microlithography*. Eds. L. F. Thompson, C. G. Willson, and M. J. Bowden. Washington, DC: American Chemical Society.
20. Webster, R. J. 1998. *Thin film polymer dielectrics for high-voltage applications under severe environments.* Master's thesis, Virginia Polytechnic Institute and State University.
21. Merricks, D. 1995. *Special polymers for electronics and optoelectronics.* Ed. J. A. Chilton and M. T. Goosey. London: Chapmann & Hall.
22. Boellaard, E., P. N. Pham, L. D. M. v. d. Brekel, and G. J. Bertens. 2003. Electrodepositable photoresist processing for RF-devices in MEMS technology. *MEPTEC Rep* Quarter Two:29–31.
23. Pham, N. P., E. Boellaard, P. M. Sarro, and J. N. Burghartz. 2002. *SAFE 2002*. Veldhoven, the Netherlands: STW.
24. Reichmanis, E., O. Nalamasu, F. M. Houlihan, and A. E. Novembre. 1999. Radiation chemistry of polymeric materials. *Polym Int* 48:1053–59.
25. Harriott, L. R. 1999. Next generation lithography. *Mater Today* 2:9.
26. Ito, T., and S. Okazaki. 2000. Pushing the limits of lithography. *Nature* 406:1027–31.
27. Le Barny, P. 1987. *Molecular engineering of ultrathin polymeric films.* Eds. P. Stroeve and E. Franses. New York: Elsevier.
28. Moreau, W. M. 1988. *Semiconductor lithography.* New York: Plenum Press.
29. Shaw, J. M., J. D. Gelorme, N. C. LaBianca, W. E. Conley, and S. J. Holmes. 1996. Negative photoresists for optical lithography. *IBM J Res Dev* 41:81–94.
30. O'Brien, M. J., and D. S. Soane. 1989. *Microelectronics processing.* Eds. D. W. Hess, and K. F. Jensen. Washington, DC: American Chemical Society.
31. Ando, S., T. Matsuura, and S. Sasaki. 1994. *Polymers for microelectronics: resists and dielectrics.* Eds. L. F. Thompson, C. G. Willson, and S. Tagawa. Washington, DC: American Chemical Society.
32. Studt, T. 1992. Polyimides: hot stuff for the '90s. *Res Dev* August:30–31.
33. Furuya, A., F. Shimokawa, T. Matsuura, and R. Sawada. 1994. Fluorinated polyimide fabrication by magnetically controlled reactive ion etching (MC RIE). *J Micromechan Microeng* 4:67.

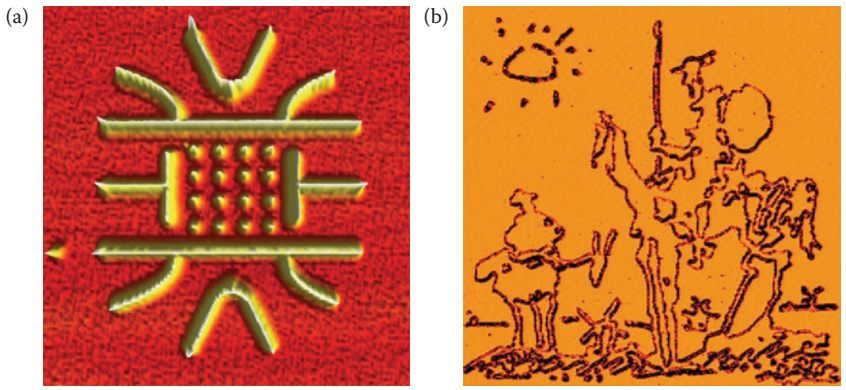
34. Suzuki, K., I. Shimoyama, and H. Miura. 1994. Insect–model based microrobot with elastic hinges. *J Microelectromech Syst* 3:4–9.
35. Shioda, T. 2007. Next generation packaging technology which supports the ubiquitous devices. Application of the flexible optical waveguide to the consumer products. *Electr Parts Mater* 46:38–42.
36. Flack, W., S. Kulas, and C. Franklin. 2000. Process characterization of an aqueous developable photosensitive polyimide on a broadband stepper. SPIE 3999-45.
37. Ito, H. 1996. Chemical amplification resists: history and development within IBM. *IBM J Res Dev* 41:69–80.
38. LaBianca, N. C., J. D. Gelorme, E. Cooper, E. O'Sullivan, and J. Shaw. 1995. *JECS 188th Meeting.* Chicago: JECS.
39. Angelo, R., J. Gelorme, J. Kuczynski, W. Lawrence, S. Pappas, and L. Simpson. 1992. U.S. Patent 5,102,772, IBM.
40. Reznikova, E. F., J. Mohr, and H. Hein. 2005. Deep photolithography characterization of SU-8 resist layers. *Microsyst Technol* 11:282–91.
41. Lee, S. J., W. Shi, P. Maciel, and S. W. Cha. 2003. *University/government/industry microelectronics symposium.* Boise, ID: IEEE.
42. Guerin, L. J. 1999. *Industrial exhibition.* Montreux, Switzerland.
43. Bertsch, A., H. Lorenz, and P. Renaud. 1999. 3D Microfabrication by combining microstereolithography and thick resist UV lithography. *Sensors Actuators* 73:14–23.
44. Yang, R., and W. Wang. 2005. A numerical and experimental study on gap compensation and wavelength selection in UV-lithography of ultra-high aspect ratio SU-8 microstructures. *Sensors Actuators Chem* 110:279–88.
45. MicroChem Corp. 2000. http://www.microchem.com.
46. Sotec Microsystems. 2000. http://mems.isi.edu/mems/yp/LOU.html.
47. Madou, M., and J. Florkey. 2000. From batch to continuous manufacturing of microbiomedical devices. *Chem Rev* 100:2679–91.
48. Braun, A. E. 1999. Photoresist stripping faces low-k challenges. *Semicond Int* October:64–74.
49. Flamm, D. L. 1992. Dry plasma resist stripping. Part I: overview of equipment. *Solid State Technol* 35:37–39.
50. Wolf, S., and R. N. Tauber. 2000. *Silicon processing for the VLSI era.* Sunset Beach, CA: Lattice Press.
51. Willson, C. G. 1994. *Introduction to microlithography.* Eds. L. F. Thompson, C. G. Willson, and M. J. Bowden. Washington, DC: American Chemical Society.
52. Brodie, I., and J. J. Muray. 1982. The physics of microfabrication. New York: Plenum Press.
53. Sze, S. M. 1988. *VLSI technology.* New York: McGraw-Hill.
54. Alling, E., and C. Stauffer. 1988. Image reversal photoresist. *Solid State Technol* 31:37–43.
55. Thompson, L. F., C. G. Willson, and M. J. Bowden, eds. 1994. *Introduction to microlithography.* Washington, DC: American Chemical Society.
56. Bowden, M. J. 1994. *Introduction to microlithography*. Eds. L. F. Thompson, C. G. Willson, and M. J. Bowden. Washington, DC: American Chemical Society.
57. Sheats, J. R., and B. W. Smith, eds. 1998. *Microlithography: science and technology.* New York: Marcel Dekker.
58. Rothschild, M., A. Forte, R. Kunz, C. Palmateer, and J. Sedlacek. 1997. Lithography at a wavelength of 193 nm. *IBM J Res Dev* 41:49–55.
59. Maluf, N. 2000. *An introduction to microelectrochemical systems engineering.* Boston: Artech House.
60. White, R. M., and S. W. Wenzel. 1988. Inexpensive and accurate two-sided semiconductor wafer alignment. *Sensors Actuators* 13:391–95.
61. Kim, E. S., R. S. Muller, and R. S. Hijab. 1992. Front-to-backside alignment using resist-patterned etch control and one etching step. *J Microelectromech Syst* 1:95–99.
62. Tatic-Lucic, S., and Y.-C. Tai. 1994. Novel extra-accurate method for two-sided alignment on silicon wafers. *Sensors Actuators* A41–42:573–77.
63. Bruning, J. H. 1980. Optical imaging for microfabrication. *J Vac Sci Technol* 17:1148–55.
64. Oldham, W. G., S. N. Nandgaonkor, A. N. Neureuther, and M. O'Toole. 1979. A general simulator for VLSI lithography and etching processes: Part I. Application to projection lithography. *IEEE Trans Electron Devices* 26:717–22.
65. Sekiguchi, A., M. Isono, and T. Matsuzawa. 1999. Measurement of parameters for simulation of 193 nm lithography using Fourier transform infrared baking system. *Jpn J Appl Phys* 38:4936–41.
66. Comello, V. 1990. Planarizing leading edge devices. *Semicond Int* November:60–66.
67. Wiesner, J. R. 1993. Gap filling of multilevel metal interconnects with 0.25-µm geometries. *Solid State Technol* October:63–64.
68. Sniegowski, J. S. 1996. *Micromachining and microfabrication process technology II*. Austin, TX: SPIE.
69. Ong, E., H. Chu, and S. Chen. 1991. Metal planarization with an excimer laser. *Solid State Technol* August:63–68.
70. Ito, H., and C. G. Willson. 1982. Chemical amplification in the design of dry developing resist materials. *Technical Papers, Regional Technical Conference, Society of Plastic Engineers* 331–53.
71. Ito, H., C. G. Willson, and J. M. J. Fréchet. 1982. New UV resists with negative or positive tone. Digest of technical papers of 1982 symposium on VLSI technology.
72. Ito, H., C. G. Willson, and J. M. J. Fréchet. 1985. Positive- and negative-working resist compositions with acid generating photoinitiator and polymer with acid labile groups pendant from polymer backbone. U.S. Patent 4,491,628, IBM.
73. Reichmanis, E., F. M. Houlihan, O. Nalamasu, and T. X. Neena. 1994. *Polymers for microelectronics: resists and dielectrics.* Eds. L. F. Thompson, C. G. Willson, and S. Tagawa. Washington, DC: American Chemical Society.
74. Reichmanis, E., L. F. Thompson, O. Nalamasu, A. Blakeney, and S. Slater. 1992. Chemically amplified resists for deep-UV lithography: a new processing paradigm. *Microlithogr World* November/December:7–14.
75. Lamola, A. A., C. R. Szmanda, and J. W. Thackeray. 1991. Chemically amplified resists. *Solid State Technol* 53:53–60.
76. Horn, M. W. 1991. Antireflection layers and planarization for microlithography. *Solid State Technol* November:57–62.
77. Taylor, G. N., T. M. Wolf, and L. E. Stillwagon. 1984. Role of inorganic materials in dry-processed resist technology. *Solid State Technol* 27:145.
78. Seeger, D. E., D. C. La Tulipe Jr., R. R. Kunz, C. M. Garza, and M. A. Hanratty. 1997. Thin-film imaging: past, present, prognosis. *IBM J Res Dev* 41:105–18.
79. (a) Rothschild, M., R. B. Goodman, M. A. Hartney, M. W. Horn, R. R. Kunz, J. H. C. Sedlacek, and D. C. Shaver. 1992. Photolithography at 193 nm. *J. Vac. Sci. Technol. B* 10:2989; (b) Hibbs, M., R. R. Kunz, and M. Rothschild. 1995. 193-nm Lithography at MIT Lincoln Lab. *Solid State Technol*. 38:69.
80. Feynman, R. P. 1993. Infinitesimal machinery. *J Microelectrochem Syst* 2:4–14.

81. Levenson, M. D., N. S. Visnawathan, and R. A. Simpson. 1982. Improving resolution in photolithography with a phase-shifting mask. *IEEE Trans Electron Devices* 29: 1828–36.
82. Levenson, M. D. 1992. Phase-shifting mask strategies: isolated dark lines. *Microlithogr World* 6:6–12.
83. Van Den Broeke, D. 1996. Transferring phase-shifting mask technology into mainstream manufacturing. *Semiconductor Fabtech* 6.
84. Fritze, M., D. Astolfi, H. Liu, C. K. Chen, V. Suntharalingam, D. Preble, and P. W. Wyatt. 1999. Sub-100 nm KrF lithography for complementary metal-oxide-semiconductor circuits. *J Vac Sci Technol* B17:345–49.
85. Switkes, M., M. Rothschild, R. R. Kunz, S-Y. Baek, D. Coles, and M. Yeung. 2003. Immersion lithography: beyond the 65 nm mode with optics. *Microlithogr World* May:4.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2

Next-Generation Lithographies and Lithography Research



(a) AFM image of quantum dot structure created on a GaAs wafer using oxidation lithography, 2.5-μm scan. (Courtesy of D. Graf and R. Shleser, Ensslin Group, ETH Zurich, http://www.asylumresearch.com.) (b) AFM image of the nanolithographically patterned polycarbonate with scratch lithography, 5-μm scan. Copy of Pablo Picasso's "Don Quixote" (http://www.asylumresearch.com).

Outline

Introduction

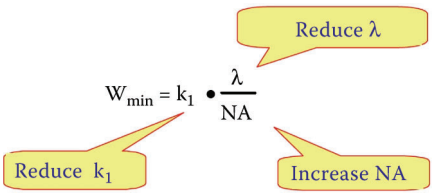
Moore's Law

Next-Generation Lithographies

Lithography Techniques in the Research and Development Stage

Questions

References



Resolution of photolithography system.

Introduction

In the integrated circuit (IC) industry, continuous improvements to deep UV (DUV) photolithography have postponed the industrial adoption of alternative next-generation lithographies (NGLs) because of the huge financial investment in existing photolithography equipment. DUV photolithography is expected to continue until the 45-nm node through resolution-enhancing techniques (RET) such as off-axis illumination (OAI), phase shift masks (PSM), and optical

proximity correction (OPC) as reviewed in Chapter 1. The 32-nm node is viewed as beyond the scope of DUV lithography. There are no clear lithography solutions listed in the International Technology Roadmap for Semiconductors (ITRS roadmap) for 2010 and beyond (Volume I, Appendix 1A).

In this chapter we introduce as possible NGLs: extreme ultraviolet lithography (EUVL), x-ray lithography, charged-particle beam lithography based on electrons and ions (such as electron and ion projection techniques), and imprint lithography [nanoimprint lithography (NIL) and step-and-flash imprint lithography (SFIL)]. These techniques are those that are regarded today as important technologies for the beyond the DUV lithography era, either for mask making or for IC production. We also review lithography approaches in the research and development stage, including lithography based on very thin resist layers and block copolymers, zone plate array lithography (ZPAL), quantum lithography (two-photon lithography), and proximal probe-based techniques such as atomic force microscopy (AFM), scanning tunneling microscopy (STM), dip-pen lithography (DPL), near-field scanning optical microscopy (NSOM), and apertureless near-field scanning optical microscopy (ANSOM). Also considered are holographic lithography and plasmonic lithography. Lithography with superlenses ("Pendry's dream") was reviewed in Volume I, Chapter 5 and will not be repeated here. As an example of a lithography method that is capable of patterning resist on nonplanar substrates (e.g., patterning a resist on the surface of a cylinder) we detail soft lithography.

It is possible that some of the alternative lithography tools we treat in the research and development category will emerge as serious NGLs in the coming years.

IC and miniaturization science are taking increasingly separate paths in adopting preferred lithography strategies. For ICs, throughput and finer geometries are needed, and batch processing is a prerequisite. In miniaturization science, modularity, good depth of focus (DOF), extension of the z-direction, that is, the height of features (skyscraper-type structures), incorporating nontraditional materials (e.g., gas-sensitive ceramic layers, polymers), and replication methods catch the spotlight, and batch fabrication is not always a prerequisite. While reading the current chapter, these important differences in characteristics of IC manufacture and miniaturization science should be kept in mind.

Moore's Law

Introduction

The dramatic increase in performance and cost reduction in the electronics industry derives from progress in IC fabrication and packaging. Speed and performance improvement of IC chips and their associated packages, and hence higher-level hardware such as computers, all are based on the minimum printable feature size [critical dimension (CD)].

Moore's Law states that the number of transistors per square centimeter roughly doubles every 18–24 months without increase in cost. Recent trends show that feature size decreases about 30% every 2 years, and one needs more and more upfront investment money to fabricate. Moore's Law—an observation, really—might hit an even bigger snag with CDs in the 32-nm mode, where DUV lithography runs out of steam and quantum effects start wreaking havoc.

After introducing Moore's Law, we analyze Ray Kurzweil's optimistic speculations about the future of Moore's Law.

Moore's Law Detailed

Control of the thickness (z) of films deposited on a substrate and control of the depth of a cavity etched in a substrate can be achieved with remarkable precision, down to 20 Å. The width and length (x, y) dimensions of a deposit or etch are more difficult to control and depend on the type of lithography used. The minimum x, y feature is determined primarily by the precise focus of the energy source that allows discrimination between areas of exposed resist and nonexposed resist. Other key factors are the tolerances on the mask, the ability to align the mask to the wafer and align subsequent layers of masks to create the proper vertical geometry, and the ability to control the rate and direction of etching and deposition.

Gordon Moore, inventor of the IC and former chairman of Intel, first described the rate of progress in reducing feature sizes for ICs when, in 1965, he observed that the surface area of a single transistor reduced by approximately 50% every 12 months. In 1975, he revised this to every 2 years, not every

Table 2.1 CDs, DRAM Bits, MPU Transistor Density, and Die (Chip) Sizes Based on the SIA Road Map*

	1997	1999	2002	2005	2008	2011
Minimum feature size: CD in nm	250	180	130	100	70	50
DRAM (bits)	256M	1G	4G	16G	64G	256G
MPU transistors/cm^2	3.7M	6.2M	18M	39M	84M	180M
DRAM chip size (mm^2)	280	400	560	790	1120	1580
MPU chip size (mm^2)	300	340	430	520	620	750

Source: Semiconductor Industry Association. 1997. *The National technology roadmap for semiconductors: technology needs*. http://www.rennes.supelec.fr/ren/perso/gtourneu/enseignement/roadmap97.pdf. For updates visit http://www.sia-online.org.[2]

*See Figure 2.1 and Volume I, Chapter 4.

18 months as is often quoted.[1] Although only an engineer's rule of thumb, this is now widely known as Moore's law. In Table 2.1, based on the Semiconductor Industry Association (SIA) road map[2] (see also Volume I, Appendix 1A), we summarize the minimum feature size (CD), dynamic random access memory (DRAM)* bits, microprocessor transistors per square centimeter, and DRAM and microprocessor die sizes as a function of the year of their introduction.[3] With decreasing CDs comes more rigid CD control: a 180-nm line width requires ±14-nm CD control, and a 50-nm line width requires ±4-nm CD control.

Historically, DRAMs have driven the progress of microlithography more than microprocessors, which exhibit more relaxed design rules (see Table 2.1). The state of the art in ICs today is a 2-GB DRAM with 60-nm features (Samsung). Intel (Integrated Electronics) introduced the Core 2 Quad "Kentsfield" chip in January 2007, a chip featuring a 65-nm technology mode.

In Figure 2.1, we illustrate Moore's law by plotting DRAM bit capacity and minimum feature size on a logarithmic y-axis with time on the x-axis (see also Volume I, Chapter 4).

Looking ahead, IC developers relying on Si complementary metal oxide semiconductor (CMOS) architecture today will switch more to alternative CMOS materials and architectures. At least five fundamental technical roadblocks need to be addressed to scale down CMOS transistors further (beyond 45 nm):

1. Carrier mobility limitations/clustering of dopant atoms: As CMOS transistors shrink, mobility of carriers decreases, and to compensate, smaller devices need a higher amount of dopant atoms; however, above a certain limit the dopant atoms clump together, forming clusters that are not electrically active. The concentration cannot be increased further, and today's chips are close to the maximum.
2. Tunneling through the gate dielectric: The transistor gates that control the flow of electrons have become so small that they are prey to undesirable quantum effects such as tunneling through extremely small barriers (e.g., a gate oxide of 1.2 nm). High-k dielectrics are used to minimize this tunneling current.
3. Doping nonuniformity: Doping nonuniformity over a wafer surface is becoming statistically significant and appears in transistors, showing a varying degree of doping (see also point 1).
4. Interconnects: Interconnects are becoming the dominant portion of circuit speed and process complexity. By volume ICs have become all wires.
5. Power dissipation: The latest Pentium chips dissipate more heat than a stove-top cooking surface (100 W/cm^2).

A few comments and additional details on the above technical challenges follow; for yet more details, consult Volume I, Chapter 4.

Transistors with feature sizes between 0.5 and 0.25 μm are operated at low supply voltages to attain adequate gain and low leakage currents. A small supply voltage keeps the dissipated power within reasonable limits, makes for faster transistor switching, and allows the implementation of smaller transistors. The latter is an example of favorable downscaling because, as transistors get smaller, the insulation

* Dynamic random access memory. A type of memory component used to store information in a computer system. "Dynamic" means the DRAMs need a constant "refresh" (pulse of current through all of the memory cells) to keep the stored information.

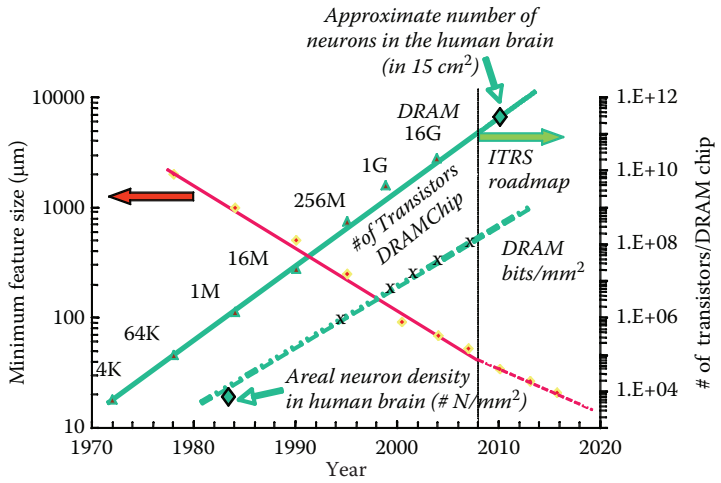


FIGURE 2.1 Semilog plot of DRAM bits and minimum feature size as a function of time. (Courtesy of Dr. Rashid Bashir, Purdue University.) See also Volume I, Chapter 4.

layer between the gate and the source/drain becomes thinner. With too high a supply voltage, electrons would "leak" through the thin insulation and compromise system stability. Eventually, the thin insulation layer will be only a couple of molecules thick, at which point transistor scaling will reach an abrupt end. In 1.8-V logic, threshold voltage is only 0.45 V. As this threshold becomes lower and lower, it will be easier for a minimal flow of electrons to be mistaken as a logical 1 instead of 0, thus causing operational problems.

Below 0.25 µm, cooling becomes necessary to maintain the same adequate transistor characteristics. Cooling of chips might be implemented using methods derived from miniaturization science, such as the use of miniature thermoelectric coolers or Joule-Thomson refrigerators (see Volume III, Chapter 8 on actuators). In the 30–50-nm range, quantum effects begin playing an important role. Quantum effects, as we saw in Volume I, Chapter 3 on quantum mechanics, can be exploited, and major efforts are underway to manufacture quantum boxes, lines, and dots for the next generation of ICs (gigascale integration or 1 billion devices per chip).

Interconnections pose another major challenge for further miniaturization of ICs. Like transistors, metal connection lines are scaled down with every new chip generation, but unlike transistors, their performance decreases as they become thinner because of increased resistance and capacitance. In this case, downscaling is disadvantageous. Connection lines were once aluminum, but copper is now replacing aluminum (e.g., in the Coppermine Pentium III) as interconnect material in building devices that not only run faster—because of copper's higher conductivity—but are also resistant to electromigration. Copper also reduces the number of steps required in manufacture. A more drastic advancement would be the use of superconducting materials for implementing interconnections, but currently no superconductor demonstrates the desired behavior under the temperature and current density conditions that exist in modern ICs. Optical interconnects are another possibility, although these have size limitations (they cannot be thinner than the wavelength of light they transmit), and the conversion of light pulses to electrical ones and vice versa could seriously compromise their performance. If optical interconnections were implemented, a switch from Si to GaAs would also be in order because GaAs offers far superior light source-making capabilities than Si. In Volume I, Chapter 5, we saw how the use of photonic crystals and metal plasmonics can circumvent size limitations of traditional optical waveguides. Photonic crystals and plasmonics offer perhaps the first real possibility for the integration of electronics and optics on the same chip.

One further comment is in order here. Ultimately, all the technological barriers listed above might be overcome, but the real question is, at what cost? Moore's second law states that capital costs increase faster than revenue and that the rate of technological

progress is controlled by financial realities. Specifically, Moore gloomily predicted that the cost of building a chip manufacturing plant doubles with every other chip generation, or roughly every 36 months. This is often referred to as Moore's "second law." The projected cost of an IC lab in 2015 is ~$100 billion.

Kurzweil: A View from an Optimist

To explore what the future of Moore's law holds, we consider Ray Kurzweil's optimistic predictions.[4] Kurzweil (Figure 2.2) and others observed that computing power had been growing exponentially long before the invention of the transistor. From Kurzweil we learn that from 1900 onward the speed of computation has doubled every three years, and that by 2020, computers will equal the human brain in terms of memory capacity and computing speed (see Figure 2.3).

Kurzweil believes that Moore's law is an illustration of a broader underlying law, *The Law of Accelerating Returns*, which is based on *The Law of Time and Chaos*. The latter, as quoted from Kurzweil, reads: "In a process, the time interval between salient events (that is, events that change the nature of the process, or significantly affect the future of the process) expands or contracts along with the amount of chaos." As a consequence, the *Law of Accelerating Returns* (a sublaw of the *Law of Time and Chaos*) states that as order exponentially increases, time exponentially speeds up, or the interval between salient events grows shorter as time passes.* It is assumed that another technology will take over when Moore's law, based on photolithography, is no longer valid. New technology may involve quantum devices, three-dimensional (3D) chips, superconductors, photonic crystals, plasmonic devices, and metamaterials. All that is needed for this evolution to take place is the growing order of the evolving technology itself and enough surrounding chaos—that is, diversity of options or alternatives—from which to draw new ideas. This is similar to natural evolution,

FIGURE 2.2 Ray Kurzweil is on the left, but does Bill believe him?

which needed the many options provided by a rich and varied environment to develop sophisticated and intelligent life forms. It follows that Moore's law, describing the speeding up of technology developments in the IC world, and Taniguchi's laws (see Chapter 6), describing the speeding up of technology developments in mechanical machining (tool development), are both examples of the *Law of Accelerating Returns*. Similarly, in data recording, the rate of progress has exponentially increased. It started slowly, perhaps with crude cave drawings, then pictorial script followed by symbolic script, then pen and paper followed by a colossal milestone, which speeded up things considerably, i.e., the Gutenberg printing press. Then magnetic recording in computers started with recording using a copper-wound coil; when this became too crude, the invention of the thin film read/write head by IBM's L. Romankiw in 1973 took over.[5] In magnetic data storage with thin film read/write, the ultimate limit at room temperature is projected to be 40 Gb/in.^{2}[2,6] A next leap forward in recording density is expected when proximal probes prove their mettle as read/write heads; 45 Gb/in.^{2} (Terris et al.[7]) and 400 Gb/in.^{2} (Kraus and Chou[8]) have been demonstrated already. Using a 32 × 32 (1024) array of AFM cantilevers, King et al.[6] thermomechanically wrote data bits 40 nm in diameter and pitch (this corresponds to a data density of 0.4 Tb/in.^{2}), and they read those bits thermally at a speed of 480 Mb/s. Figure 2.4 shows a picture of a thin film read/write head (a) and an STM-based read/write head (b) (IBM's millipede).

Yet another example of the *Law of Accelerating Returns* is exemplified in the Human Genome Project started in the mid-1980s (see Volume III,

* The other sublaw of the *Law of Time and Chaos* is the *Law of Increasing Chaos*. With chaos exponentially increasing, time exponentially slows down. The events developing after the Big Bang are an illustration of this law; most salient events occurred immediately after the Big Bang and became less frequent thereafter.

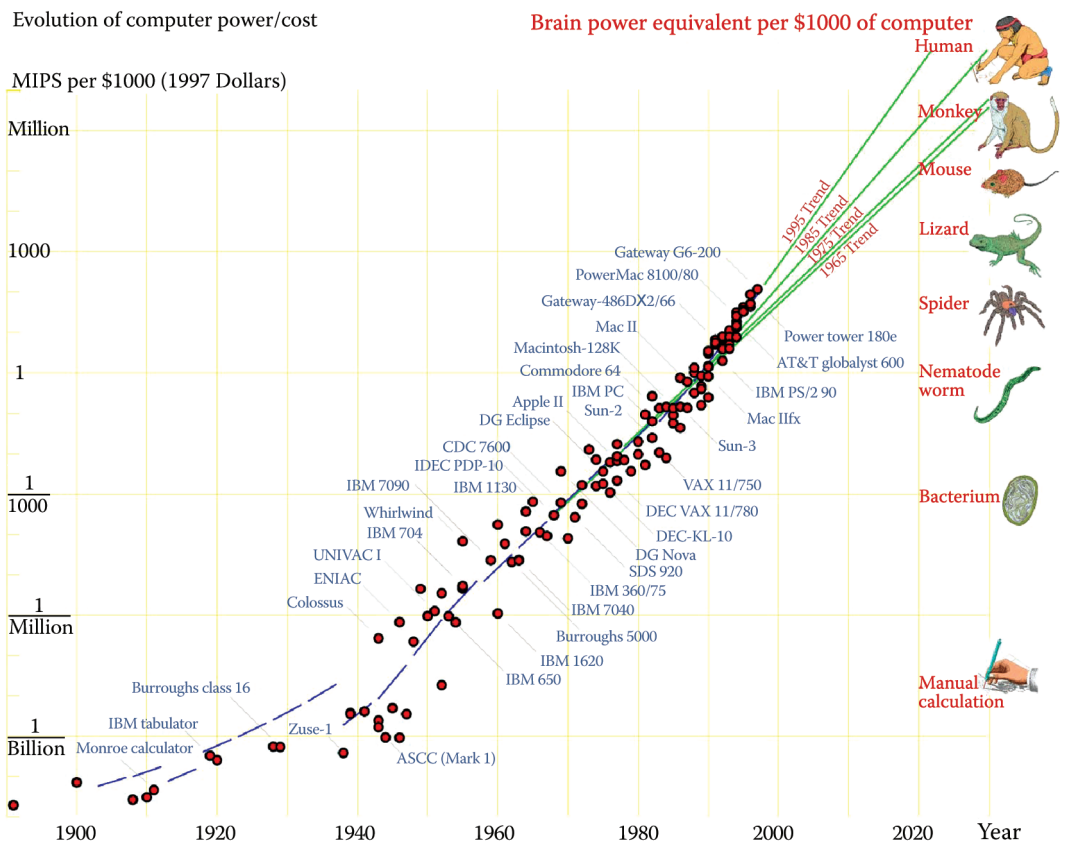


FIGURE 2.3 What $1000 of computing buys. The number of million instructions per second (MIPS) is a general measure of computing performance. (Based on Ray Kurzweil.)

Chapter 2). The speed of the process of writing down all 3.5 billion bits of the human genetic code exponentially increased since the start of the project, causing the project to finish ahead of schedule. For some additional reading from Kurzweil's work, see *The Singularity Is Near: When Humans Transcend Biology.*[9]

Next-Generation Lithographies

Introduction

Having set a rather optimistic forecast for the future of miniaturization sciences, we now introduce a set of possible next-generation lithographies

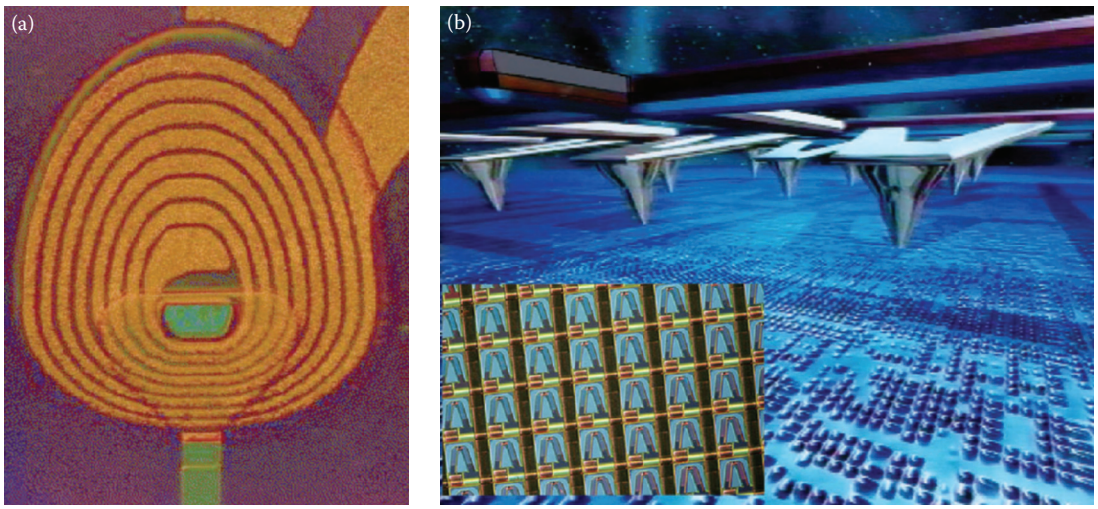


FIGURE 2.4 (a) Thin film read/write head. (b) STM-based read/write head (IBM's millipede).

(NGLs). This includes extreme ultraviolet lithography (EUVL), x-ray lithography, charged-particle-beam lithography, including scattering with angular limitation projection electron-beam lithography (SCALPEL) and ion projection lithography (IPL), as well as imprint lithography [nanoimprint lithography (NIL) and step-and-flash imprint lithography (SFIL)]. In the context of electron-beam lithography (EBL), we also look at progress made with micro-electromechanical system (MEMS)-based arrays of field emitters.

Extreme Ultraviolet Lithography

As mentioned before, the 32-nm node is viewed as beyond the scope of DUV lithography [even when applying all the resolution-enhancing technologies (RETs) or taking advantage of immersion lithography, which we learned about in Chapter 1]. EUVL is one of the most likely NGLs to achieve the 32-nm node.

Extreme ultraviolet lithography (EUVL), using wavelengths in the 10–14-nm range to carry out projection imaging, is the most natural extension of optical projection lithography as, in principle, it only differs in terms of the wavelength. This type of radiation is also referred to as *soft x-ray radiation* and *vacuum UV*. Sources for this type of radiation are laser-produced plasmas and synchrotrons.

To appreciate the merits of EUVL, we take as a starting point the expressions for resolution (R):

$$R = k_1 \frac{\lambda}{NA} \tag{1.16}$$

and depth of focus (DOF):

$$DOF = \pm\delta = \pm k_2 \frac{\lambda}{NA^2} \tag{1.17}$$

In Chapter 1 we learned that in practice, both resolution R and DOF are experimentally determined and are influenced by both resist and exposure systems. In large-scale IC manufacture with g- and i-line light, k_1 values used to be around 0.6 (see Figure 1.81) and DOFs were larger than 0.5 μm. At that time, these parameters led to the desired CD control within a tolerable process window. The push for higher resolution with 248-nm and 193-nm radiation made it necessary to achieve smaller k_1 values, and this led to processing problems; it became more difficult to control the CD, and it created an intolerably low DOF. The RETs discussed in Chapter 1 enhance resolution and the effective DOF, but, as we discussed, they are often not very manufacturer-friendly (people only half-jokingly refer to RET as "real expensive technology"). With EUVL, on the other hand, one surmounts these problems because of its shorter wavelength. The technique is capable of printing sub-100-nm features while maintaining a DOF of 0.5 μm or larger, and has k-values that make the process control less demanding. Unfortunately, things are not that simple—EUV is strongly absorbed in virtually all materials, and, consequently, imaging must be carried out in a vacuum; also, all camera optics and masks used must be reflective rather than refractive. New resists and processing techniques must be developed as well.[10–12]

In Figure 2.5 a schematic of an EUVL setup is shown. A laser is directed at a xenon gas and heats up the gas to create a plasma, radiating light at 13.6 nm—too short for the human eye to see. That EUV light is directed onto a condenser, gathering the light and directing it onto the mask. The mask is a representation of one level of the IC under construction and is made by patterning an absorber on some parts of a mirror but not on others (Figure 2.6). The pattern on the mask is reflected onto a series of four to six curved mirrors, reducing the size of the image and focusing the image onto the resist-coated silicon wafer. The entire process takes place in a vacuum because at these short wavelengths even air absorbs the light.

With respect to EUVL reflectivity, it is well known that most materials have very low reflectivity for near-normal incidences at those short wavelengths; therefore, multilayer Bragg reflectors with multiple reflecting coatings are used, as illustrated in Figure 2.6. The concave and convex mirrors in EUVL, including the mask, are coated with multiple layers (e.g., 80 of them) of molybdenum and silicon (another pair of material used is Mo-Be) with dissimilar EUVL optical constants, providing a resonant reflectivity when the period of the layers is close to λ/2. With magnetron-sputtered Mo-Si-based reflectors, peak reflectivities of up to 70% at 13.4 nm

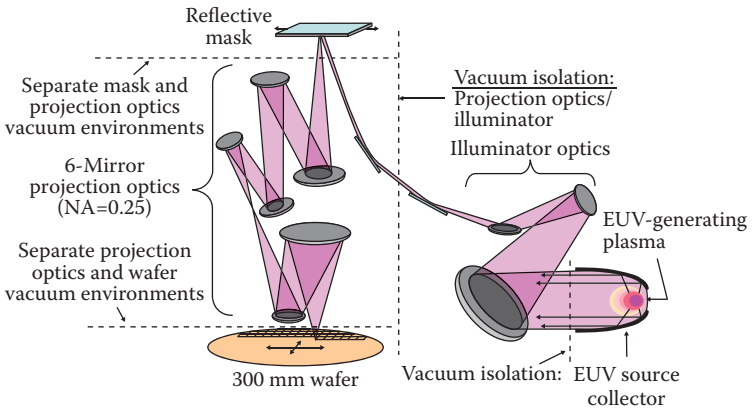


FIGURE 2.5 Schematic of an EUVL lithography setup. (From SEMATECH's Next Generation Lithography Workshop brochure.)

have been achieved (the other 30% is absorbed by the mirror). Without the coating, the light would be almost totally absorbed before reaching the wafer. The mirror surfaces have to be nearly perfect; even small defects in coatings can destroy the shape of the optics and distort the printed circuit pattern.[10]

The multilayer stacks in EUVL masks today have a demonstrated defect level of ~10^{14} defects/cm^2, still two orders of magnitude above target. Prototype EUVL cameras designed for 13.4-nm radiation have an numerical aperture (NA) = 0.1 and a four-mirror set with multilayer Mo-Si-reflecting coatings.[10] These step-and-scan 4× cameras have a resolution better than 100 nm with a ring field of 26 × 1.5 mm. In the step-and-scan system, the mask and wafer are simultaneously scanned in opposite directions, with the mask moving four times faster than the wafer. The camera wave-front error at 13.4 nm needs to be less than 1 nm so that the surface figure (that is, the basic shape of each of the mirrors) will be accurate to 0.25 nm rms or better. Achieving these accuracies will also require improved metrology techniques because current ones cannot confirm the accuracies required. The fact that EUVL masks are reflective has advantages and disadvantages. Advantages are that no thin and brittle membranes are needed as required in the case of x-ray lithography—just a solid substrate (e.g., a Si wafer) coated with the required reflective coatings and a patterned EUVL absorber; because of the 4:1 reduction, mask making is also easier than with x-rays. A major disadvantage is that the defect density in the reflective coating needs to be made so small that there are few techniques available yet to accomplish such defect-free films. Current resists absorb all the EUVL photons within less than 100 nm. Approaches for EUVL resists include silylated single-layer resist, refractory bilayer resists, and trilayer resists.

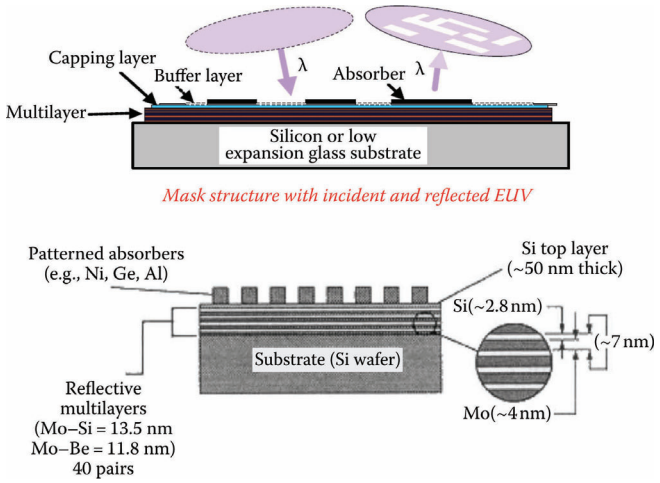


FIGURE 2.6 Two schematic representations of EUVL reflective masks. In EUVL, only reflective optics will work (see text).

The Extreme Ultraviolet LLC (EUV LLC) was created to spearhead the research involved in bringing EUVL to the marketplace. The consortium includes Intel Corporation, Motorola Corporation, Advanced Micro Devices Corporation, IBM, Infineon, Micron Technology, and the Sandia and Lawrence Livermore National Laboratories (http://www.sandia.gov/media/NewsRel/NR2001/euvlight.htm). Intel is leading this consortium, and an introduction of EUVL is planned for 2009. For the past couple of years there has been a growing realization that the resolution capabilities of the EUVL wavelength are being

countered by the effects of electrons released after absorption, and it is feared that a 13.4-nm resolution might be impossible. On the other hand, on May 30, 2007, NEDO announced, that using EUVL, it had drawn circuit patterns with just 26-nm line widths (isolated and dense lines) using EUVL (http://www.nedo.go.jp/english/archives/190820/190820.html).

X-Ray Lithography

Introduction

X-ray lithography uses a shadow printing method similar to optical proximity printing. The x-ray wavelength (4–50 Å) is much shorter than that of UV light (2,000–4,000 Å) (Table 2.2). Hence diffraction effects are reduced and higher resolution can be attained. For example, for an x-ray wavelength of 5 Å and a gap of 40 µm, R is equal to 0.2 µm. In contrast with electron lithography and ion-beam lithography, no charged particles are directly involved in x-ray exposures, which makes the need for vacuum less stringent (see Figure 2.7).[13] Like with all high-energy radiation (see EUVL), a vacuum is preferred to avoid absorption of photons by gas molecules, formation of ozone, and corrosive reactions with the exposure station's hardware. Another advantage of x-rays is that one can use flood exposure of resist-coated wafers, ensuring higher throughput than when writing with a thin electron or ion beam. The method is also referred to as deep x-ray lithography (DXRL).

Table 2.2 Light Sources for Various Types of Lithography

Wavelength (nm)	Source	Range
436	Hg arc lamp	g-line
405	Hg arc lamp	h-line
365	Hg arc lamp	i-line
248	Hg/Xe arc lamp, KrF excimer laser	Deep UV (DUV)
193	ArF excimer laser	DUV
157	F2 laser	Vacuum UV (VUV)
~10	Laser produces plasma sources	Extreme UV (EUV)
~1	x-ray tube, synchrotron	x-ray

The three main classes of sources for x-ray lithography are electron impact tubes, laser-based plasmas, and synchrotrons (Figure 2.8). In miniaturization science, we are mainly concerned with synchrotron radiation. For reasons that are explained below, in the case of ICs, less-expensive alternative x-ray sources might suffice.

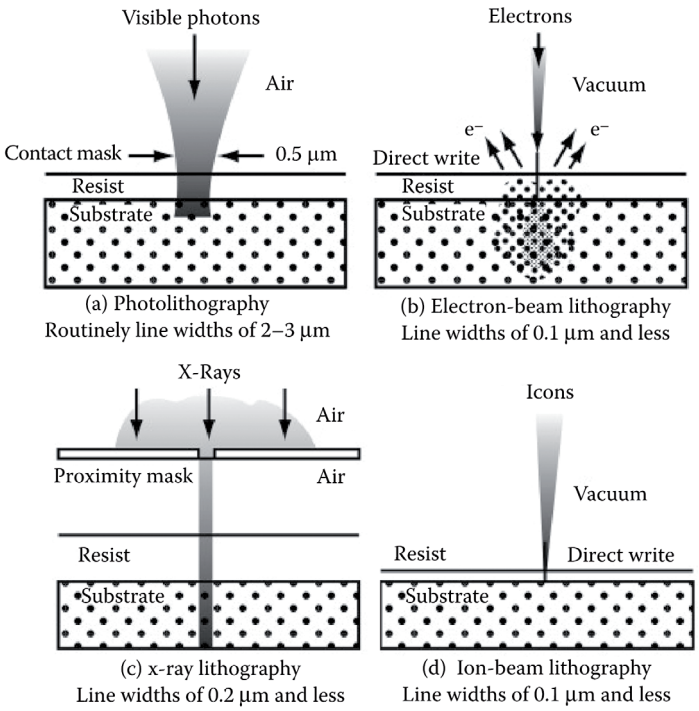


FIGURE 2.7 A comparison of (a) photolithography, (b) electron-beam, (c) x-ray lithography, and (d) ion-beam. (Based on Brodie, I., and J.J. Muray. 1982. *The Physics of Microfabrication*. New York: Plenum Press. With permission.[13])

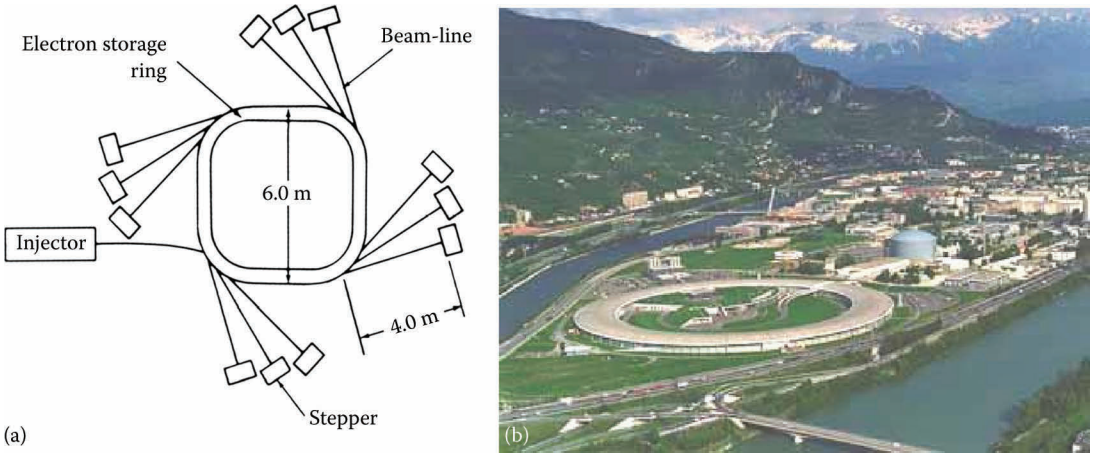


FIGURE 2.8 (a) Schematic of a synchrotron. (b) Grenoble synchrotron (http://www.esrf.eu).

Major Features of X-Ray Lithography

X-ray lithography is superior to optical lithography because of its use of shorter wavelengths and its very large DOF, as well as because exposure time and development conditions are not as stringent. Reproducibility is high because results are independent of substrate type, surface reflections, and wafer topography. Another important benefit is that x-ray lithography is immune to low atomic number (Z) particle contamination (dust). With an x-ray wavelength on the order of 10 Å or less, diffraction effects generally are negligible, and proximity masking can be used, increasing the lifetime of the mask. With a standard 50-µm proximity gap and using synchrotron x-rays, 0.25-µm patterns can be printed; by decreasing the proximity gap to 25 µm, patterns of 0.15 µm can be resolved.[14] The obtainable aspect ratio, defined as the structural height or depth to the minimum lateral dimension, reaches more than 100. With UV photolithography, under special conditions, an aspect ratio of about 25 is possible at most (see SU-8 in Chapter 1). An aspect ratio of 100 corresponds to the aspect ratio attainable by wet chemical anisotropic etching of monocrystalline Si (see Chapter 4).

In x-ray lithography, there are essentially no optics involved, and although this sounds like an advantage, it also presents one major disadvantage because one can only work with 1:1 shadow printing. No image reduction is possible, so the mask fabrication process is very complicated. In the United States, IBM remains the only major champion of x-ray lithography for NGL.[15] At the end of 1999, IBM fabricated several PowerPC 604e microprocessor batches to demonstrate the viability of the method.[16] In Japan, there are still many players involved, such as NTT, Toshiba, Mitsubishi, and NEC.[17]

LIGA

The LIGA technique (a German acronym for *Lithographie*, *Galvanoformung*, *Abformung*) was invented about 25 years ago.[18] LIGA exploits all of the advantages of x-ray lithography listed above, and the process is schematically illustrated in Figure 2.9. The LIGA process involves a thick layer of resist (from micrometers to centimeters), high-energy x-ray radiation, and resist development to make a resist mold. By applying galvanizing techniques, the mold is filled with a metal. The resist structure is removed, and metal products result. Alternatively, the metal part can serve as a mold itself for precision plastic injection molding. Several types of plastic molding processes have been tested, including reaction injection molding, thermoplastic injection molding, and hot embossing. The so-formed plastic part, just like the original resist structure, may also serve as a mold for fast and cheap mass production because one does not rely on a new x-ray exposure. LIGA enables new building materials and a wider dynamic range of dimensions and possible shapes. A showpiece structure for LIGA technology is pictured in Figure 2.10, in which an ant holds a Ni gear.

Of particular interest to miniaturization science is the possibility of creating 3D shapes with slanted

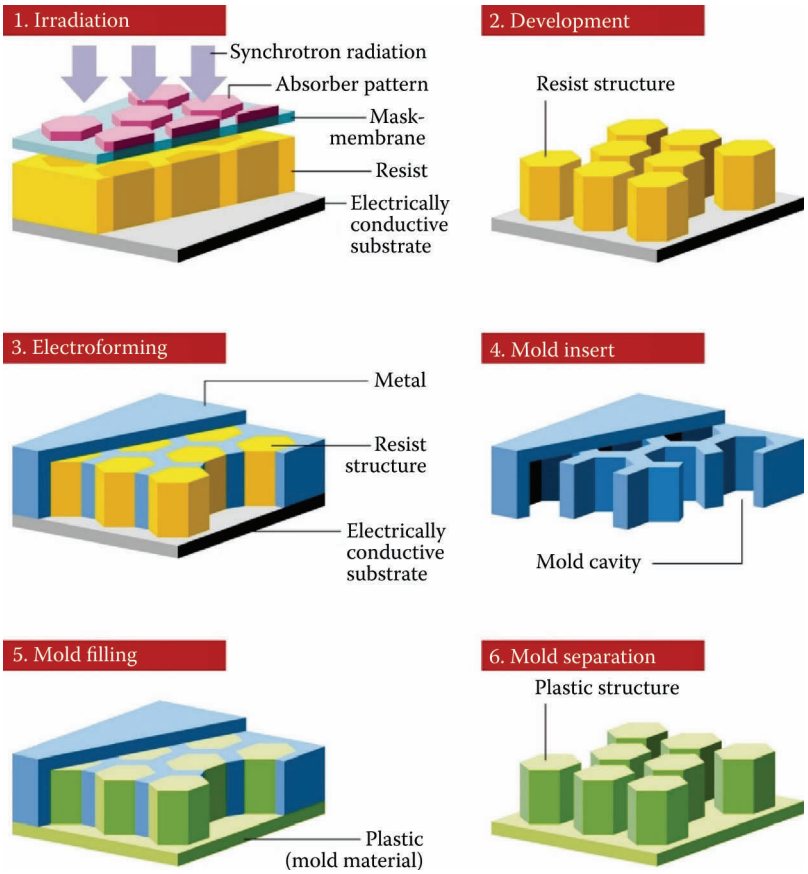


FIGURE 2.9 The LIGA process.

sidewalls (see Figure 2.11)[19] and step-like structures. Two German companies, IMM (http://www.imm-dueck.de) and Jenoptik Mikrotechnik GmbH (http://www.jo-mikrotechnik.com), have developed an x-ray scanner that allows for continuous tilt angles of the mask/substrate assembly with respect to the x-ray beam, rotation of the mask/substrate, and that contains an internal mask alignment system for different masks for multiple exposures. The unprecedented precision attainable with LIGA makes this technique stand out against other 3D lithography methods such as DUV with SU-8 (see Chapter 1). Using the IMM/Jenoptik scanner, a vertical resist wall with a precision better than 0.05 µm was demonstrated, and even resist walls inclined at a 45° angle were made with an accuracy of 1 µm over a height of up to 500 µm.[18] LIGA is further discussed in Chapter 10.

FIGURE 2.10 Ant with gear. (From Forschungszentrum Karlsruhe, Program Microsystem Technologies. With permission.)

X-Ray Resists

An x-ray resist should have high sensitivity to x-rays, high resolution and resistance to chemical, ion, and/or plasma etching, thermal stability of >140°C, and a matrix or resin absorption of less than 0.35 μm^{-1} at the wavelength of interest. No presently available resist meets all those requirements. One material predominantly used in x-ray lithography is positive poly(methylmethacrylate) (PMMA), a material better known by its trade names Plexiglas™ and Lucite™. Another name commonly used for polymers based on PMMA is *acrylics*. Clear sheets of the material are

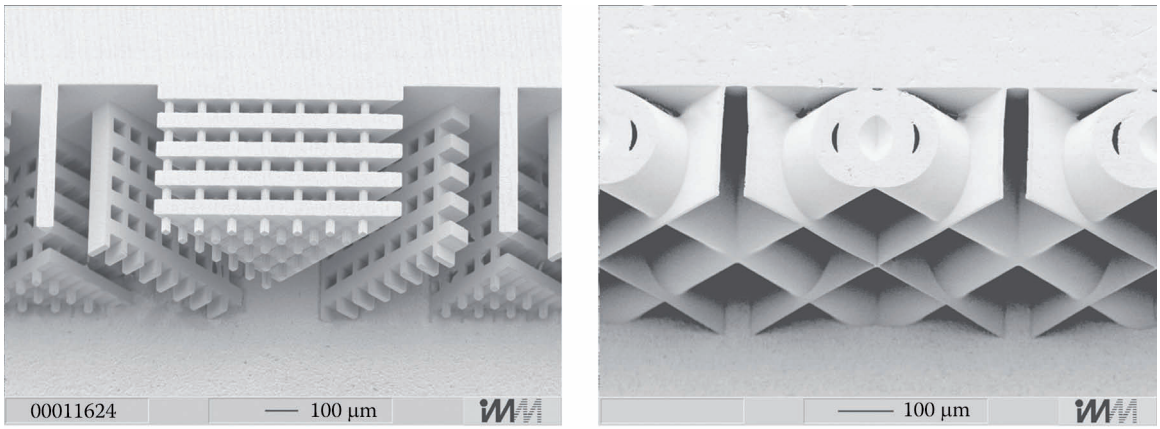


FIGURE 2.11 LIGA structures obtained at IMM using an x-ray scanner enabling continuous tilt angles of the mask/substrate assembly. (Ehrfeld, W., and A. Schmidt. 1998. Recent developments in deep x-ray lithography. *J Vac Sci Technol* B16:3526–34.[19]) (Courtesy IMM.)

used to fabricate "unbreakable" windows, inexpensive lenses, machine guards, and clear lacquers on decorative parts. At a wavelength of 8.34 Å, the lithographic sensitivity of PMMA typically hovers about 2–6.5 J/cm^2, a rather low sensitivity implying a small throughput. A possible approach to make PMMA more x-ray sensitive is the incorporation of x-ray-absorbing high atomic number atoms. Another approach, discussed in Chapter 1, involves the use of chemically amplified photoresists. Other x-ray resists explored for LIGA applications are poly(lactides)[20] and some SU-8 epoxy resin modifications.[21] These resists show a considerably enhanced sensitivity and reduced stress corrosion compared with PMMA.[19]

Negative x-ray resists exhibit inherently higher sensitivities compared with positive x-ray resists, although their resolution capability is limited by swelling. Poly(glycidyl methacrylate-co-ethyl acrylate) (PGMA), a negative e-beam resist, has been used in x-ray lithography. In general, resist materials sensitive to e-beam exposure are also sensitive to x-rays and function in the same way; that is, materials that are positive in tone for electron-beam radiation typically also are positive in tone for x-ray radiation. A strong correlation exists between the resist sensitivities observed with these two radiation sources, suggesting that the reaction mechanisms might be similar for both types of irradiation.[22]

The IC industry requires a typical resist layer not more than 1–2 µm thick. Thicker layers, between 10 and 1,000 µm, are dictated by the need for high-aspect-ratio micromachines. The technology of applying thicker layers of photoresist remains challenging. Spin coating of multiple resist layers (for relatively thin coats), resist casting with in situ polymerization of mildly cross-linked PMMA (for layers >500 µm), and plasma polymerization of PMMA with a possibility of subsequent diamond grinding of the resulting layers are some of the techniques currently in use (see also Chapters 1 and 10).

In the case of exposures with very high energy x-rays (hard x-rays), the associated wavelength is measured in angstroms, not in nanometers. At those energies, almost every type of polymer becomes a "resist," and even "resistless" lithography becomes possible as thin films can be etched, vaporized, or ion-implanted directly.

X-Ray Masks

Another challenge in x-ray lithography, besides the low sensitivity of the resists and the high cost of sufficiently bright x-ray sources, is the mask making—already complex for producing DRAMs, but even more complex for 3D structures with high aspect ratios. In Table 2.3, the procedure for making an optical mask is compared with the procedure involved in making an x-ray mask. Technologies developed for manufacturing masks for submicrometer circuitry with x-ray lithography do not directly transfer to the fabrication of x-ray

TABLE 2.3 Optical versus X-Ray Mask

Optical Mask	X-Ray Mask
Mask design: CAD	Mask design: CAD
Substrate preparation:	Substrate preparation:
Quartz	Thin membrane substrate (Si, Be, Ti)
Thin metal film deposition	Deposit plating base (50 Å Cr then 300 Å Au)
Pattern delineation:	Pattern delineation:
Coat substrate with resist	Coat with resist
Expose pattern (optical, e-beam)	Expose pattern (optical, e-beam)
Develop pattern etch Cr layer	Develop pattern
Strip resist	Absorber definition: Electroplate Au (~15 μm for hard x-rays) Strip resist
Cost: $1–3K	Cost: $4–12K
Duration: 3 days	Duration: 10 days

masks for building micromachines. An x-ray mask basically consists of a pattern of x-ray-absorbing material (a material with a high atomic number, Z, such as gold, tungsten, or tantalum silicide) on a membrane substrate transparent to x-rays (a low Z material, e.g., Ti, Si, SiC, Si_3N_4, BN, Be) (see Figure 2.12). The membrane must have windows or other features in certain areas for alignment purposes. Additionally, the material also needs to be thermomechanically stable to a few parts in 10.[23] Mechanical stress in the absorber pattern can cause in-plane distortion of the supporting thin membrane, requiring a high Young's modulus material. Also, humidity or the x-ray exposure itself might distort the membrane.

A fabrication sequence for an x-ray mask is shown in Figure 2.13. In this case, the x-ray absorber pattern is ion beam-etched in a Au-film after the pattern first has been written in a 300 Å PMMA layer with an e-beam. The absorber film typically consists of two metal layers: a thin layer of chromium for adhesion to the substrate, topped by a thicker layer of gold. The higher the required aspect ratio of the exposed resist, the thicker the gold layer of the mask absorber pattern must be to maintain a good contrast. The x-ray mask shown in Figure 2.13 only has a 400-Å-thick gold layer, which adequately covers DRAM manufacture but not high-aspect-ratio 3D structures, where 5–15 μm of gold absorber might be required for a 500-μm-thick resist and up to 50 μm for 10-cm-thick resists. A proximity x-ray mask, as shown in Figure 2.13, is placed close to the substrate, and its pattern is reproduced by exposure to x-rays. Various other schemes to make LIGA x-ray masks are explored in Chapter 10.

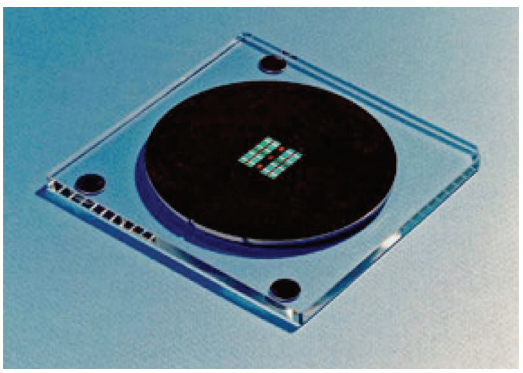


FIGURE 2.12 X-ray mask structure: Ta patterns on SiC or SiN membrane. The absorber Ta is 300 nm thick (stress ~ 0 MPa). Membrane SiC or SiN is 2 m thick (stress = 100 MPa ~ 200 MPa), 20 ~ 30 mm square. The minimum pattern width is 50 nm. Pattern position accuracy: 3σ < 30 nm (http://www.ntt-at.com/products_e/x-ray_masks/index.html).

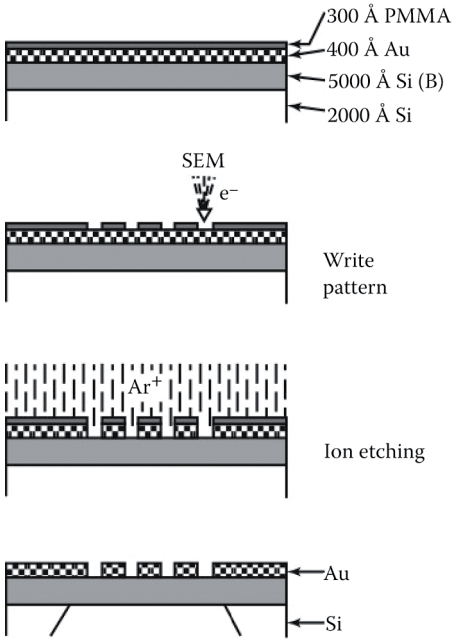


FIGURE 2.13 Fabrication of a silicon membrane-based x-ray mask with a gold absorber pattern. For use in high-aspect-ratio micromachining, the gold absorber layer must be between 5 and 15 μm. (Based on Brodie, I., and J.J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press. With permission.[13])

Why Use a Synchrotron to Generate X-Rays?

The full power of x-ray lithography for miniaturization science only materializes when using hard, collimated synchrotron radiation. To appreciate this, we will explore the procedure using a less intense, less collimated beam from an electron-beam bombardment source. A schematic diagram of an x-ray exposure system using an electron-beam bombardment-based x-ray source is shown in Figure 2.14.

The mask is typically offset above the wafer by about 10 µm after alignment. A proximity scheme, rather than a contact mask, is a useful feature, given the fact that an x-ray mask can cost up to $13,000. Because the x-ray source is finite in size and separated by a distance, *D*, the edge of the mask does not cast a sharp shadow but rather has a blurry region associated with it, known as penumbral blur, *d*. Image blurring limits the ultimate resolution power of an x-ray exposure system, as shown in Figure 2.14. As diffraction effects can be ignored, simple geometric considerations can be used for relating the image to the pattern on the mask. From Figure 2.14, we estimate that the blurring, δ, at the resist plane is given by:

$$\delta = s\left(\frac{d}{D}\right) \tag{2.1}$$

where s = mask-to-wafer gap
d = the source diameter
D = source-to-substrate distance

In a high-resolution system, δ can be controlled to within 0.1 µm. Spacing, *s*, should allow the accommodation of large-diameter masks while avoiding the high risk of contacting the resist and greatly increasing the occurrence of defects. The x-ray source must be sufficiently collimated. In practice, this translates into a small source diameter d (e.g., a few millimeters) and a large source-to-mask distance (*D*). All these factors contribute to minimizing penumbral blurring. Conventional e-beam-generated x-ray sources have sizes of a few millimeters and are about 40 cm away from the mask. Unfortunately, a large distance required for adequate collimation results in prohibitively

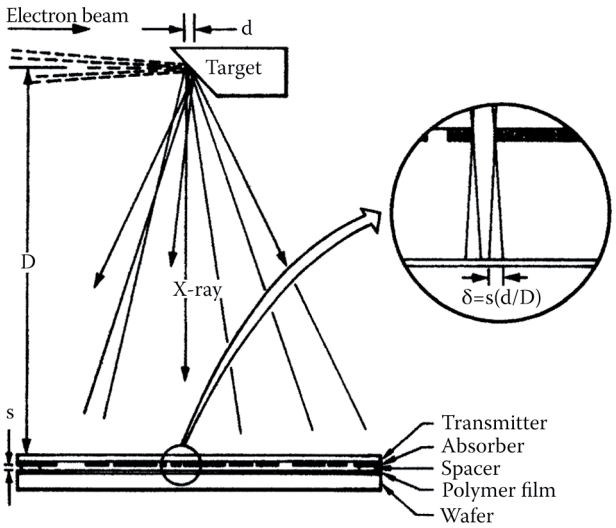


FIGURE 2.14 X-ray lithography with an electron-beam x-ray source. Inset, extent of penumbral effect calculated from geometric considerations. (Based on Brodie, I., and J.J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press. With permission.[13])

long exposure times (e.g., hours) as a result of the weak intensity of these sources. With synchrotron radiation, on the other hand, penumbral blurring does not limit the spatial resolution. Because of the high collimation of synchrotron radiation, rather large distances between the mask and the wafer can be tolerated (about 1 mm for 1-µm line width patterns). In the electron storage ring or synchrotron, a magnetic field constrains electrons to follow a circular orbit, and the radial acceleration of the electrons causes electromagnetic radiation to be emitted forward. Thus, the radiation is strongly collimated in the forward direction and can be assumed to be parallel for lithographic applications. Because of the much higher flux of usable collimated x-rays, shorter exposure times become possible. The Advanced Light Source (ALS) synchrotron in Berkeley, for example, can deliver a flux of 0.4 W/cm^2 at 30 m for 3–9-keV radiation (http://www.als.lbl.gov). Especially if one wants to generate a highly collimated photon flux in the spectral range required for precise deep-etch x-ray lithography in thick resist layers, synchrotron radiation comes close to being the ideal source because of its intensity, tunability, small source size, and small divergence.[24]

As mentioned in Chapter 1, the IC industry today favors improved photolithography over x-rays. Despite huge efforts, until recently, x-ray lithography

seemed abandoned as a candidate for NGL. The prohibitive cost of introducing a new type of industrial lithography remains a strong deterrent, pushing existing photolithography to its absolute resolution limit. For example, by using UV phase-shifting mask lithography, planar IC features approaching x-ray lithography resolution have been made. Even in the micromachining field, there are continued attempts to squeeze more out of classical photolithography. Using techniques such as DUV lithography and deep dry etching with dense plasmas, LIGA-like, high-aspect-ratio features have been produced, hemming in the potential for broader use of x-ray lithography even for building 3D miniaturized machines. But x-ray lithography might well make a comeback.[25] At recent lithography conferences, four in five NGL presentations were on x-rays. Highlights are there is no need for a synchrotron anymore as new types of light sources have been developed (e.g., from a laser); x-ray masks are cheaper than EUV masks; and 15-nm resolution has been established.[17]

Charged-Particle Beam Lithography

Introduction

In this section, we introduce lithographies based on charged-particle beams. Charged-particle versus non–charged-particle approaches were compared earlier in Figure 2.7. Both narrow-beam direct writing and flood exposure projection systems with electrons and ions will be considered. The mask fabrication process is significantly simpler in the case of narrow-beam lithography compared with flood exposure-based lithography.

In direct write systems, the computer-stored pattern is directly converted to address the writing charged particle beam, enabling the pattern to be exposed sequentially, point by point, over the whole wafer. In other words, the mask is a *software mask*. Electron-beam (e-beam) and ion-beam (i-beam) lithographies involve high current density in narrow electron or ion beams. The smaller the beam sizes, the better the resolution, but more time is spent writing the pattern. This sequential (scanning)-type system exposes one pattern element or pixel at a time. Within that area, the charged-particle beam delivers maximum current (i), which is limited primarily by the source brightness and column design. The experimental setup imposes a limit on the speed at which the writing beam can be moved and modulated, resulting in a "flash" time in seconds (t). The maximum dose (in coulombs/cm^2) deliverable by a particular beam is given by:

$$D_{max} = it/A \qquad (2.2)$$

with A the pixel area in square centimeters. It will then be necessary to work with resists that react sufficiently fast at D_{max} to produce a lithographically useful, 3D image (latent or direct image). The e-beam method displays a large DOF, as active focusing over various topographies is possible. The continued development of better charged-particle beam sources keeps widening the possibilities for nanoscale engineering through direct write lithography, etching, depositing, analyzing, and modifying a wide range of materials, well beyond the capability of classical photolithography. Table 2.4 lists some i-beam and e-beam applications.

Flood exposure of a mask in a projection system (that is, parallel exposure of all pattern elements at the same time, as we saw in DUV) is possible with ions and electrons as well. In principle, these methods can take advantage of the excellent resolution of charged-particle beams while providing the throughput levels necessary for IC manufacture. Exposure masks are fabricated from heavy metals on semitransparent organic or inorganic membranes. We will learn how SCALPEL and IPL have the potential of making e-beam and i-beam high throughput. The high cost

Table 2.4 Electron- and Ion-Beam Applications

Electron-Beam Applications	Ion-Beam Applications
Nanoscale lithography	Micromachining and ion milling
Low-voltage scanning electron microscopy	Microdeposition of metals
Critical dimension measurements	Maskless ion implantation
Electron beam-induced metal deposition	Microstructure failure analysis
Reflection high-energy electron diffraction (RHEED)	Secondary ion mass spectroscopy
Scanning auger microscopy	

of mask fabrication and the instability of the mask as a result of heating have postponed commercial acceptance of these high-energy exposure systems. Moreover, with i- and e-beams, flood exposure is limited to chip-size fields because of difficulties in obtaining broad, collimated, charged-particle beams. The most prevalent use of charged-particle beams remains the narrow-beam scanning mode.

Electron-Beam Lithography

Overview Electron beam lithography (EBL) refers to a lithographic process that uses a focused beam of electrons to form the circuit patterns needed for material deposition on (or removal from) a substrate. In contrast with optical lithography, which uses light for the same purpose, electron lithography offers higher patterning resolution than optical lithography because of the shorter wavelength associated with the 10–50-keV electrons that it uses. Electrons are easily generated and can be used to either directly write the desired structure to the resist in electron-beam direct write or in electron projection lithography (EPL) flood exposure where a much wider beam is used (such as in SCALPEL). We also analyze the status of micromachined electron emission sources and electron emission source arrays.

Direct Write e-Beam Lithography Direct write EBL is a high-resolution patterning technique in which high-energy electrons (10–100-keV) are focused into a narrow beam and used to expose electron-sensitive resists. There are two basic ways to scan an electron beam. In raster scanning, the patterns are written by an electron beam that moves through a regular pattern. The beam scans sequentially over the entire area and is blanked off where no exposure is required. On the contrary, in vector scanning, the electron beam is directed only to the requested pattern features and hops from features to features. Therefore, time is saved in a vector scan system. The direct write technique was first developed in the 1960s using existing scanning electron microscope (SEM) technology. A typical setup involves 30-keV electrons from a Hitachi HL-700F EBL. As a research solution, several groups are still "Rube Goldberging" their standard SEM to create customized e-beam writing systems. For example, Rosolen[25] modified a Hitachi S2500 with a purpose-built pattern generator and alignment system (see also Figure 6.14 in Volume III for a schematic of an SEM and a photo of a commercial SEM machine). The instrument did not require alignment marks on the sample and was able to compensate for positional errors caused by the sample stage and mask tolerances. The EBL method, like x-ray lithography, does not limit the obtainable feature resolution by diffraction because the quantum mechanical wavelengths of high-energy electrons are exceedingly small. EBL exhibits some other attractive attributes compared with photolithography. These include:

1. Precise control of the energy and dose delivered to a resist-coated wafer
2. Deflection and modulation of electron beams with speed and precision by electrostatic or magnetic fields
3. Imaging of electrons to form a small point of <100 Å, as opposed to a spot of 5000 Å for light
4. No need for a physical mask; only a *software mask* is required
5. The ability to register accurately over small areas of a wafer
6. Lower defect densities
7. Large DOF because of continuous focusing over varying topography; at 30 keV, electrons will travel on average >14 μm deep into a PMMA resist layer

Some of the disadvantages of EBL include:

1. Electrons scatter quickly in solids, limiting practical resolution to dimensions >10 nm
2. Electrons, being charged particles, need to be held in a vacuum, making the apparatus more complex than for photolithography
3. The slow exposure speed: an electron beam must be scanned across the entire wafer (for a 4-in. wafer with a high feature density this requires ~1 hour)
4. High system cost

The resolution of EBL tools is not simply the spot size of the focused beam; it also is affected

by forward scattering of the e-beam inside the resist and substrate (at relatively smaller angles) and by backscattering from the substrate exposing the resist over a greater area than the beam spot size (Figure 2.15). This phenomenon is called the "proximity effect," and hence the developed pattern is wider than the scanned pattern due to the interactions of the primary beam electrons with the resist and the substrate, as illustrated in Figure 2.16. Line width variations caused by local feature density are an immediate result. Proximity correction algorithms are used to achieve more uniform resist exposure with EBL. Such corrections are computer intensive and time consuming, however, and make a slow technique even slower.

The biggest disadvantage of direct write electron lithography is its low throughput (approximately 5 wafers/hour at less than 0.1-μm resolution). As a result, the use of EBL has been limited to mask making and direct writing on wafers for specialized applications, for example, small batches of custom ICs.

Writing with an e-beam can be additive or subtractive. In lithography the technique is subtractive; as an example of additive e-beam writing, e-beam-induced metal deposition from a metal organic gas [e.g., W deposition from $W(CO)_6$] has been used for the formation of microstructures of various geometries (see Figure 9.29). These devices are made one by one rather than in a large batch. Usually, this type of slow, expensive fabrication technique prohibits commercial acceptance. Some microstructures, especially intricate microsystems, might be worth the bigger price tag. In that case, serial microfabrication techniques may not be as prohibitive as they would be in the case of ICs.

Electron Emission Sources Three types of electron emission (field, thermionic, and photoemission) (Figure 2.17) underwrite the principle for electron emission source construction.[26] Schottky emission (SE) and cold field emission (CFE) have been in common use, especially for nanometer-sized beams for electron focusing systems. Emission of electrons from a metal under the influence of a field occurs in both SE and CFE. During SE, a blunt tungsten emitter tip coated with a low work function material (ZrO) is heated to 1800 K, and thermionic emission takes place; that is, heat thermally excites the electrons enough to bring them out of the material. In other words, the field emission is helped along by thermal excitation of the electrons (the current in this case is given by the Richardson-Dushman equation).[27] During CFE, a much smaller tungsten wire (radius of <0.1 μm) is used, and a very high field causes electrons to tunnel out of the material. In CFE sources, electrons tunnel from various energies below the Fermi level. With SE cathodes, thermally excited electrons (nontunneling electrons) escape over a field-lowered potential energy barrier (the current in this case is given by the Fowler-Nordheim equation).[27] Both SE and CFE sources display similar energy spreads, but their energy distributions are mirror images (Figure 2.18).[26]

With photo emitters, energy is transferred to electrons by incident photons, and photoelectrons generated close to the surface are able to escape.

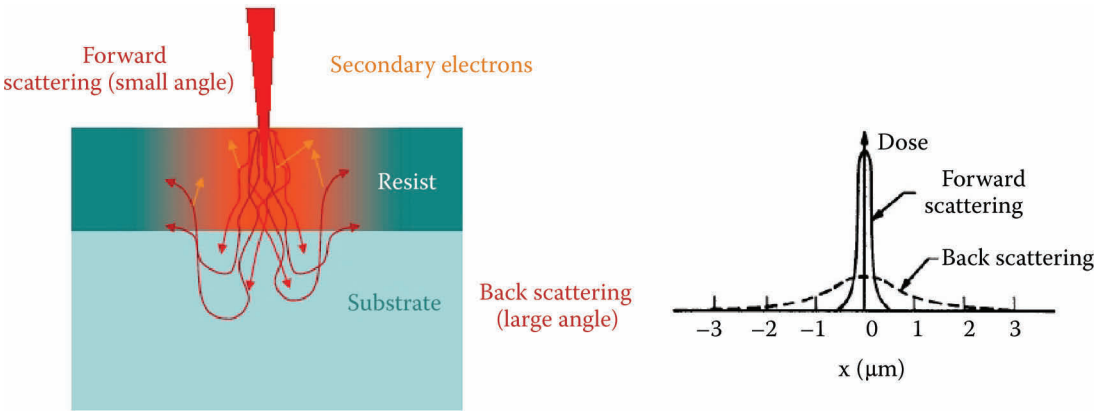


FIGURE 2.15 Scattering of the e-beam inside the resist and substrate and backscattering from the substrate exposing the resist over a greater area than the beam spot size.

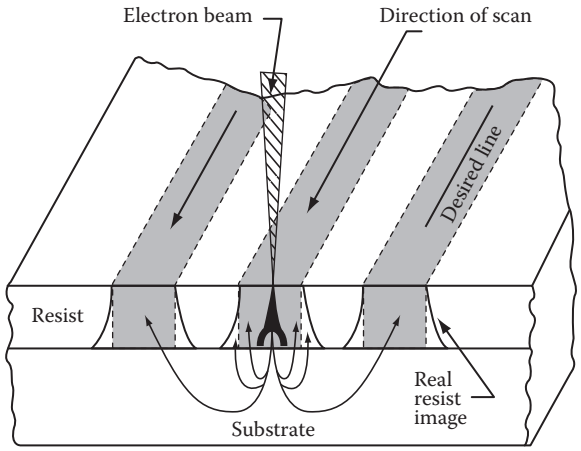


FIGURE 2.16 Proximity effect in direct write e-beam lithography.

Attaining high current levels in a submicrometer electron beam at low voltages (500 eV–1 keV) is of interest for EBL and SEM (see also Chapter 6 in Volume III). When sensitive biological samples or electron beam-sensitive resists are involved, SEM pictures must be made at voltages less than 1 keV. At these low voltages, high currents are required to attain the needed detail and to minimize edge effects. Traditional SEM cathodes using tungsten hairpin filaments are very limited at low voltages; they cannot supply enough current in submicrometer beams at the low-voltage end. The high brightness SE and CFE sources, including LaB_6 and CeB_6, can do the job (Table 2.5). SEM instruments with beam diameters of 1 nm have been made using such cathodes. For further reading about EBL, refer to Brewer.[28]

Electron emission in a water bucket

THE THREE MECHANISMS used by field emission sources all basically involve emitting electrons and ions from a metal surface under the influence of a strong electric field.

Understanding these mechanisms is where the water bucket comes in.

In this analogy, the water level in a bucket represents the Fermi level—the highest occupied energy level in a cathode material. The work function is the energy required to get the water droplets (electrons) from the top of the liquid out of the bucket. This is the distance equivalent to the potential energy barrier.

In photoemission, photon energy excites electrons at the Fermi level of the cathode material and can impart enough kinetic energy to allow the electrons to escape from the bucket.

In thermionic emission, heat thermally excites the electrons, providing enough energy to boil the electrons off and out of the bucket.

In field emission a high electric field can thin the side of the bucket enough so that the electrons can tunnel through it.

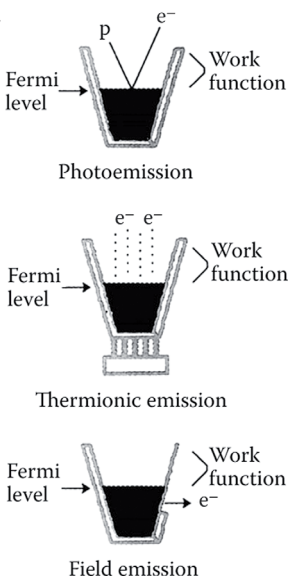


FIGURE 2.17 Photoemission, thermionic emission, and field emission.

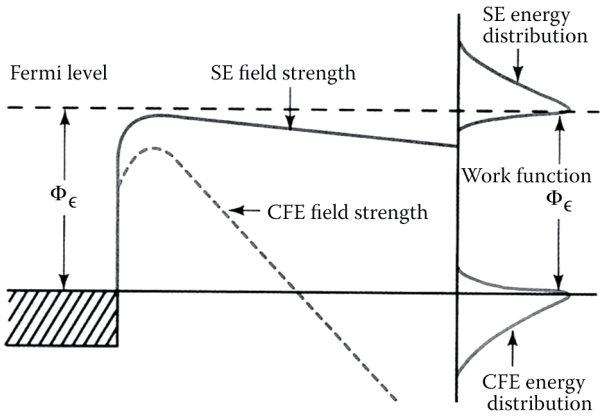


FIGURE 2.18 SE and CFE energy distributions.

Electron-Beam Resists Numerous commercial e-beam resists are produced for mask making and direct write applications. Bombardment of polymers by electrons causes bond breakage; thus, in principle, any polymer material can function as a resist. However, the important considerations include sensitivity, tone, resolution, and etching resistance. PMMA exemplifies an inexpensive positive e-beam resist with a high-resolution capability and a moderate glass transition temperature T_g (114°C). Microposit SAL601 (Shipley) is an often-used negative e-beam resist. SAL601, being novolak based, has much better dry etch resistance than PMMA resists. The same materials act as x-ray resists as well. This is not coincidental, as there is a strong relation between x-ray and e-beam sensitivity. A copolymer of glycidyl methacrylate and ethyl acrylate (COP) is another frequently used negative resist in mask manufacture. This material, although exhibiting good thermal stability, has (as is typical for acrylates) poor plasma-etching resistance. The measured G value (see Chapter 1) of representative polymers in the COP family is about 10. Whelan et al.[29] developed a low-energy electron beam top surface imaging chemically amplified (AXT) resist. This AXT positive-mode resist incorporates a poly(hydroxystyrene) base resin, achieves sub-100-nm resolution with 2-keV electrons, and features a sensitivity less than 1 μC/cm^2.

TABLE 2.5 Comparisons of Electron Sources

Working Principle	Gun Material	Brightness (B) (A/cm^2 Sr)	Energy Width (eV)	Filament Temperature	Gun Vacuum (torr)
Electron Source					
Thermionic Emission					
Electron emission at high temperature	W	~10^5	2–3	~3000 K	10^{-5}–10^{-6}
	LaB_6	~10^6	2–3	2000–3000 K	10^{-7}–10^{-8}
Field Emission					
Electron tunneling in high field	W	10^5–10^{10}	0.2–0.5	Room	<10^{-9}

Streblenchenko et al.[30] patterned FeF_2 and CoF_2 with 100-keV electrons from a scanning transmission electron microscope using a 0.5-nm diameter electron probe. The fluoride resist films are prepared by thermal evaporation on thin carbon films. During electron bombardment, fluorine escapes and the transition metal coalesces. The resolution in a 20-nm-thick, very small-grained CoF_2 film is about 5 nm. An electron dose of 1000 C/cm^2 at 100 keV removes 90% of all the fluorine in the 20-nm-thick CoF_2 film. Arbitrarily shaped nanometer-scale magnetic structures have been written in such CoF_2 films.

For further reading, see *Handbook of Microlithography Micromachining and Microfabrication*, edited by P. Rai-Choudhury (published by SPIE Press), and Joe Nabity's web site (http://www.jcnabity.com) describing the Nanometer Pattern Generation System. This system is the top-selling SEM lithography system at research institutions in North America, and its use is becoming widespread worldwide.

Electron Projection Lithography: Scattering with Angular Limitation Projection Electron Beam Lithography Electron projection lithography (EPL) is electron beam lithography (EBL) with a wide electron beam, using a mask for high throughput. Scattering with angular limitation projection lithography or SCALPEL, a projection electron beam technique, using a 4× reduction and a step-and-scan writing strategy, is the most prominent EPL technique. The aspect of SCALPEL (1995, Bell Laboratories and Lucent Technologies) that differentiates it from previous attempts at projection EBL is the use of a mask as illustrated in Figure 2.19. As in the case of x-ray lithography, the mask consists of a low atomic number membrane covered with a layer of a high atomic number material: the pattern is delineated in the latter. Although the mask is almost completely electron-transparent at the energies used (100 keV), using the difference in electron-scattering characteristics between the membrane and the patterned material generates contrast. The membrane scatters electrons weakly and to small angles, whereas the patterned layer scatters them strongly and to high angles. An aperture in the back-focal (pupil) plane of the projection optics blocks the strongly scattered electrons, forming a high-contrast aerial image at the wafer plane.

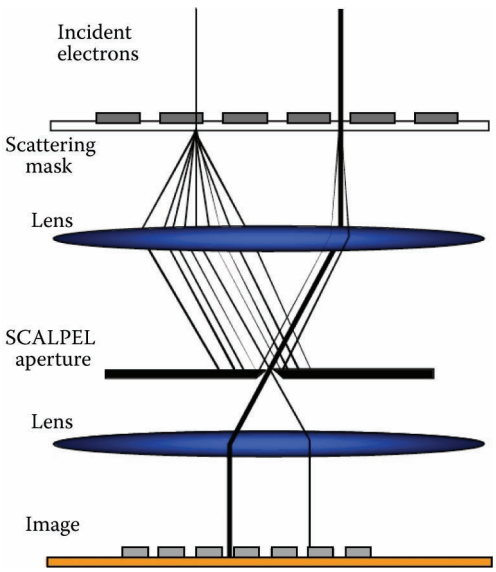


FIGURE 2.19 Scattering with angular limitation projection lithography or SCALPEL, a projection electron beam technique.

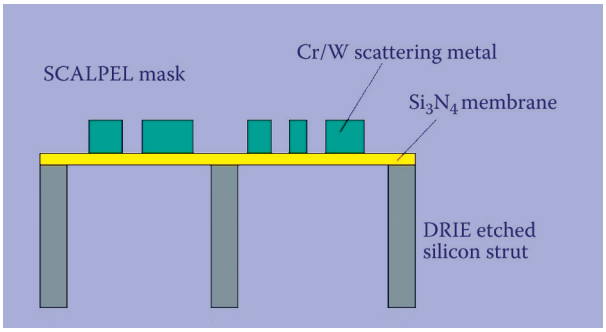


FIGURE 2.20 SCALPEL mask.

The mask is a continuous thin membrane (~100 nm) made of a low atomic number material like silicon nitride, patterned with a 25–50-nm-thick layer of a high atomic material such as tungsten (Figure 2.20). To minimize image distortion, thick silicon struts support the thin silicon nitride membrane. A broad beam (2–3 mm in diameter) of high-energy electrons (~100 keV) passes through both membrane and W pattern with minimal energy loss, leading to minimal heating effects. In SCALPEL, the aperture, rather than the mask, absorbs unwanted energy. Thus, the functions of contrast generation and energy absorption are divided between the mask and the aperture. This means that very little of the incident energy is absorbed by the mask, minimizing thermal instabilities in the mask. It should be noted that, although the membrane scatters electrons weakly compared with the scatterer, a significant fraction of the electrons passing through the membrane are scattered sufficiently to be stopped by the SCALPEL aperture. Image field is still limited today to 1 × 1 mm at the mask level. The small images at the wafer are assembled (also stitched together) from the individual portions on the mask through a combination of interferometrically monitored mechanical motions of the mask and wafer stages and electronic deflections of the beam (see Figure 2.21). The relatively low cost masks in SCALPEL [$27,000 vs. $40,000 for a phase shift mask (PMS), $40,000–59,000 for an EUV mask, and $5,000–10,000 for standard binary masks] may make this technique a cost-effective solution for lithography at 130 nm and less.[31,32]

In 1999, IBM introduced Projection Reduction Exposure with Variable Axis Immersion Lenses (PREVAIL).[33] In this system, the optical axis of the electron system is shifted, so aberrations are reduced, enabling larger scan fields. The lens system shifts the electron optical axis along a preset curvature and precisely deflects the electron beam to follow the variable axis curvature. The SCALPEL

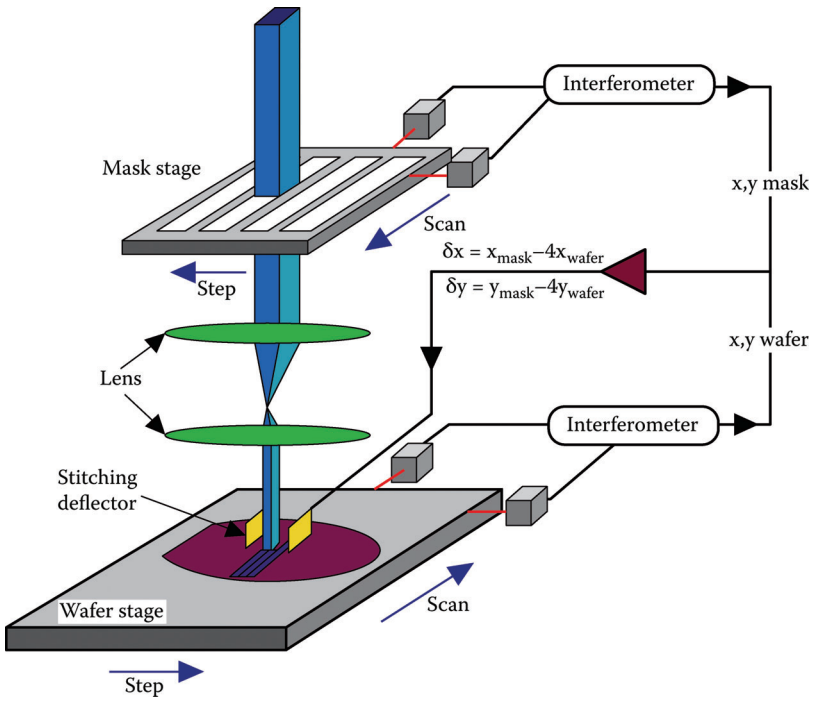


FIGURE 2.21 SCALPEL writing strategy involves step-and-scan writing.

proof-of-concept system was the first system to implement sequential illumination of the mask in an e-beam reduction projection system by mechanical scanning of reticle and wafer at a 4:1 speed ratio underneath a stationary beam. PREVAIL carries this concept further by combining electronic beam scanning with continuous stage motions.[34] This approach provides a significantly larger effective field size needed to achieve commercially viable throughput levels.

Micromachined Electron Emission Sources There have been many attempts at decreasing the accelerating voltage in direct write EBL because lower energy electrons have a more confined lateral backscattering range. Moreover, it was found that low-energy EBL leads to higher sensitivity—that is, for a given resist, a lower exposure dose is required than with higher energy electrons.[35] Low-energy e-beams require thinner resist layers, as the penetration depth falls off very quickly (e.g., in a typical resist, 2 keV will penetrate more than 120 nm). Micromachining is ideally suited to deliver these low-energy electron writing tools. We only present some examples here of micromachined e-beam tools. A first example involves the work by Zlatkin et al.,[36] who developed an elegant array of focused electron writing beams operating at 300 eV or less. The emitters used are cold field emission (CFE) sharpened tungsten tips, although thermionic or Schottky emitters would be feasible as well (for a description of different types of emission sources see above). The emitters are positioned several millimeters above the micromachined extraction holes of a lens array fashioned in a single crystal Si substrate. Because of the relatively large distance between the emitter and the extraction anode (a 1-µm aperture), no precise alignment is necessary. The setup and a detail of the lens system are shown in Figure 2.22.

Each lens has a total area of 1 × 1 cm on a 500-µm-thick silicon wafer and is generally built up from two 0.5-µm thin Si layers. The two thin Si layers are separated from each other and the Si substrate by insulating silicon oxide. An anisotropic etch into the single crystal Si side opposite the thin layered sandwich structure exposes the oxide/Si/oxide/Si/oxide thin layered structure. A short isotropic etch of the exposed oxide is followed by a 1-µm dry-etch bore in the first thin Si layer (first aperture, 1-µm diameter). The small aperture is used to extract electrons from the emitter and to spatially confine the initial electron beam. Another isotropic etch through the second oxide is followed by the larger dry etched aperture (50 µm) in the second thin Si layer. This

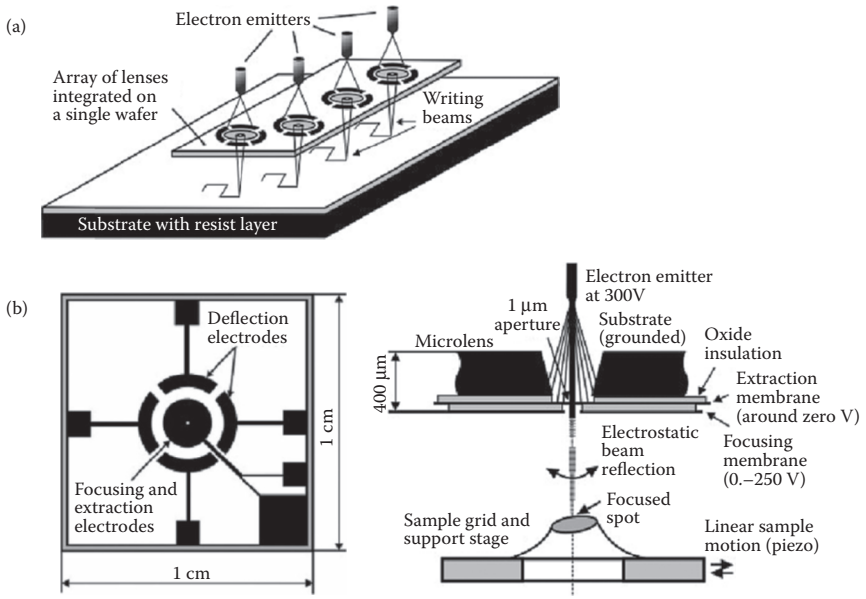


FIGURE 2.22 Array of focused electron writing beams operating at 300 eV or less. Setup (a) and detail of the lens system (b). (Redrawn from Zlatkin, A., and N. Garcia. 1999. Functional scanning electron microscope of low beam energy with integrated electron optical system for nanolithography. *Microelectron Eng* 46:213–17.[36])

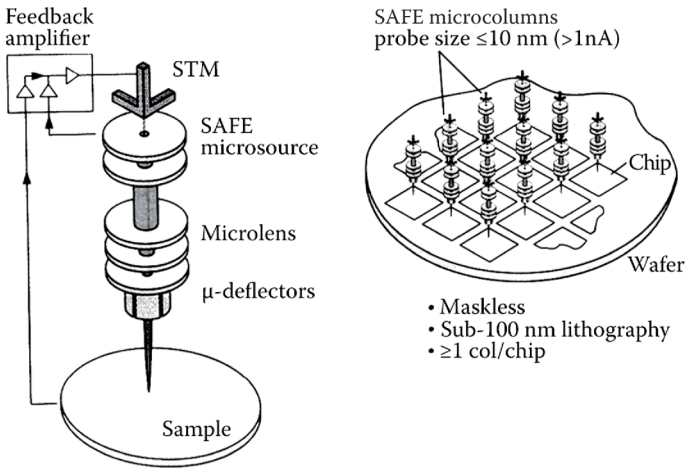


FIGURE 2.23 A microcolumn based on STM aligned field emission (SAFE) and arrayed microcolumn lithography. (From Editorial. 1993. Novel electron-beam lithography system being explored at Cornell's NNF. *Solid State Technol* 36:25–26. With permission.[37]) Acronyms are explained in the text.

second aperture is used to focus the electron trajectories and is well aligned with the first aperture. Changing the voltage on the extraction and focusing electrodes on the thin Si membranes can change the focal length. To deflect the electron beam for scanning, a set of four electrodes symmetrically placed around the extraction and focus apertures is energized. The four electrodes are placed on top of another insulator layer on the focusing membrane. Applying potentials to opposite pairs of deflection quadrants deflects the electron beam laterally, and the focused spot can be swept over an area either to create an SEM image or to pattern a design in EBL mode. The aim is to eventually be able to image and write in the sub-100-nm range. Imaging in the 30-nm range was demonstrated already, but writing is still inadequate (limited to >200 nm).

Micromachining presents other potential solutions to make "nanolithography" a cost-effective and adequately fast proposition in the future. The next example addresses the issue of speed in writing with small e-beam systems by introducing massive parallelism. At Cornell's National Nanofabrication Facility, a research group has been working on arrays of microfabricated, miniaturized electron lithography systems based on STMs. In this STM aligned field emission (SAFE) system, the physical dimension of the electron beam column (length and diameter) are on the orders of millimeters. A field emission tip is mounted onto an STM; the STM feedback principle is used for precision x, y, and z piezoelectric alignment of the tip to a miniaturized electron lens to form a focused probe of electrons as shown in Figure 2.23.[37] Because many electron-optic aberrations scale with size, microfabrication techniques enable lenses with negligible aberrations, resulting in exceptionally high brightness and resolution. STM controls also allow for stability of the emission by automatically adjusting the z-position via a piezo element.

An array of these microcolumns, each with a field emission tip as the source with individual STM sensors and controls, can generate patterns in parallel, one or more columns per chip (see Figure 2.23). The low voltage of operation of these tips obviates the need for proximity-effect corrections, as low-voltage operations have proven to eliminate proximity effects.[37,38] For example, low-energy electrons (15–50 eV) from a SAFE system have been used to write patterns with 23-nm feature sizes, more than four times smaller than can be written on the same substrate and in the same resist with a tightly focused 50-kV e-beam.[38]

Nanolithography with ultrasharp field emitters as discussed above needs to be distinguished from scanning probe lithography (SPL), discussed further below. SAFE is based on an emission technology, whereas tunneling is the mechanism in SPL. The key differences are the distance between tip and substrate (1 nm with SPL vs. 100 nm or more with sharp field emitter tips) and the electron generation mechanism (tunneling vs. emission).

Besides the application described above, micromachined emission sources are also used or are intended to be used for flat panel displays (see Spindt cathodes in Figure 2.42), high-temperature and high-radiation amplifiers, vacuum gauges, and radio frequency oscillators and amplifiers.[39]

Ion-Beam Lithography

Introduction In ion-beam lithography, resists are exposed to energetic ion bombardment in a vacuum. As in the case of e-beam systems, ion-beam lithography offers direct write and flood exposure fabrication opportunities. Direct write ion-beam lithography consists of point-by-point exposures with a narrow beam scanning source (e.g., 20 nm in diameter) of liquid gallium metal. In ion projection lithography (IPL), flood exposure, with a broad ion beam of H^+, He^{2+}, or Ar^+ (e.g., 1–2 cm^2 in size), of a stencil mask is used to pattern a substrate. Ion-beam lithography uses ions of a kinetic energy from a few keV up to several MeV. Ion lithography achieves higher resolution than optical, x-ray, or electron beam techniques because ions come with a smaller wavelength and therefore undergo almost no diffraction and scatter much less than electrons because the secondary electrons produced by an ion beam are of lower energy and have a short diffusion range. The total spread including forward and backward scattering of the "stiffer" ion beams is typically less than 10 nm, and they only require about 1–10% of the electron dose to expose a resist. The resist can be an ordinary PMMA resist that absorbs most of the ions during exposure. Therefore, radiation damage to sensitive, underlying structures is minimized and smaller compared with EBL and x-ray lithography. There is also the possibility of a resistless wafer process. However, the most important application of focused ion beams remains the repair of masks for optical or x-ray lithography, a task for which commercial systems are available.

Focused Ion-Beam and Deep Ion-Beam Lithography For ion-beam construction, liquid metal ion (LMI) sources are becoming the choice for producing high-current-density submicrometer ion beams. With an LMI source, liquid metal (typically gallium) migrates along a needle substrate. By applying an electrical field, a jet-like protrusion of liquid metal forms at the source tip. The gallium-gallium bonds are broken under the influence of the extraction field and are uniformly ionized without droplet or cluster formation. LMI sources hold extremely high brightness levels (10^6 A/cm^2 sr) and a very small energy spread, making them ideal for producing high-current-density submicrometer ion beams. Beam diameters of less than 50 nm and current densities up to 8 A/cm^2 are the norm. In addition to Ga, other pure element sources are available, such as indium and gold. By adopting alloy sources, the list expands to dopant materials such as boron, arsenic, phosphorus, silicon, and beryllium.

Compared with photons (x-rays and DUV light) or electrons, ions chemically react with the substrate, allowing a greater variety of surface modifications such as patterned doping. The ion-beam spot size is the smallest possible—smaller than UV, x-ray, or electron-beam spots. The smallest focused ion-beam (FIB) spots reached are about 4–8 nm, accomplished by using a two-lens microprobe system and a single-isotope gallium ion source.[40, 41] Ion-beam lithography experiences the same drawbacks as an electron-beam system in that it requires a serially scanned beam and a vacuum.

FIBs can be used to perform maskless implantation and metal patterning with submicrometer dimensions. FIB also has been applied to milling in IC repair, maskless implantation, circuit fault isolation, and failure analysis (see Table 2.4). In Figure 2.24, deposition, milling, and imaging are illustrated. Micromachining applications of ion-beam technology are reviewed in more detail in Chapter 6 on mechanical energy-based removing (see Figure 6.32). As a machining tool, FIB is very slow. Except for research, it may take a long time to become an accepted "micromachining tool." For additional reading on ion-beam lithography in general refer to Selinger et al.[42]; for more specific reading on FIB-induced deposition, see Brodie.[13]

Using high-energy (2 MeV) protons, deep ion-beam lithography (DIBL) in PMMA produces submicrometer (300 nm) walls with an aspect ratio approaching 100. Three-dimensional complex microstructures with smooth walls and corners have

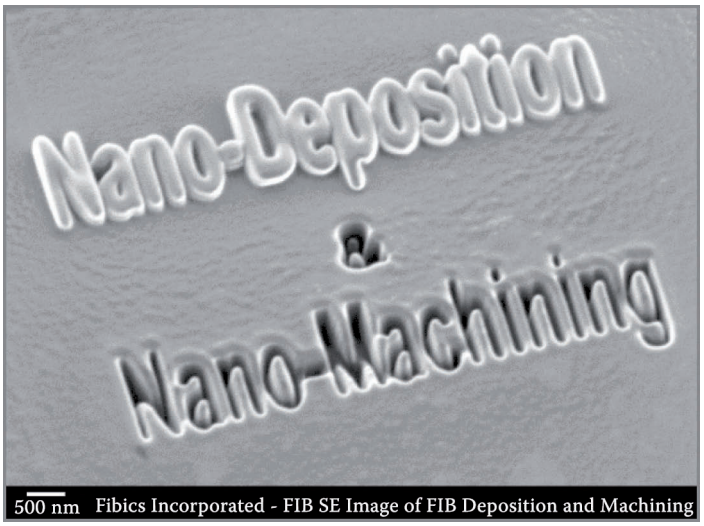


FIGURE 2.24 FIB deposition can produce features 200 nm or less in thickness. FIB milling can produce even finer results. Approximately 30 minutes of FIB time was required to produce this structure (http://fibics.com/Micromachining.html).

been produced this way. The range of 2-MeV protons in PMMA is 63 µm.[43] Multiple exposures at different ion energies (e.g., 0.6 MeV and 2 MeV) allow production of multilayer structures in single-layer resists such as SU-8.[44]

FIB systems have been produced commercially for approximately 20 years, primarily for large semiconductor manufacturers. FIB systems operate in a similar fashion to an SEM. When operated at low-beam currents, the system can be used for imaging, and when operated at high-beam currents, it can be used for site-specific sputtering or milling (see Figure 2.24).

Ion-Projection Lithography Ion projection lithography (IPL), illustrated in Figure 2.25, is another of the candidates for high-throughput lithography dedicated to future 32-nm and sub-32-nm IC generations. In ion projection lithography, a stencil mask, consisting of an Si membrane with a fine pattern of holes, is illuminated with a uniform beam of hydrogen or helium ions at about 10 kV. The image of this mask is projected through an ion optical column and demagnified by 4× on the wafer. Ions are very efficient at exposing resists so that the exposure time per die is well under a second even for comparatively insensitive (1 µC/cm^2) resists. Protons or helium ions are generated by a radio frequency-driven filament (ion source). An extraction system consists of

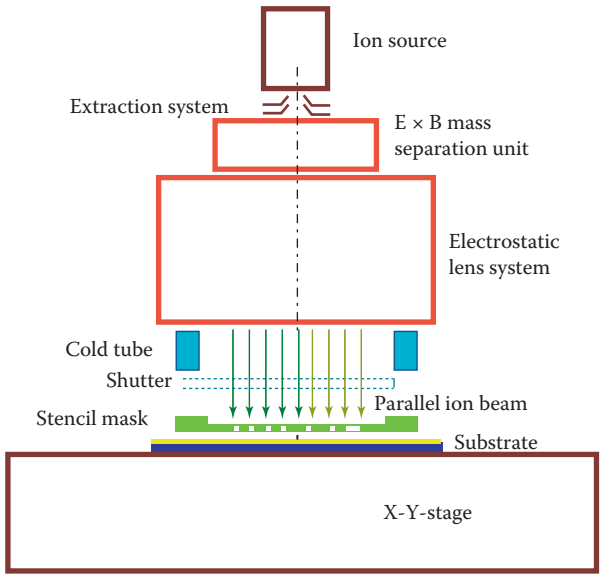


FIGURE 2.25 Ion projection lithography (IPL).

an **E** × **B** mass separation unit and an electrostatic lens system.

The ion flood lithography mask typically consists of a silicon membrane a few micrometers thin (3 µm) with pattern openings that allow protons to pass through.[45] The mask fabrication process involves silicon-on-insulator (SOI) and dry etching.[46] The holes are defined by EBL and reactive ion etching, whereby an oxide is used as a mask. Current IPL approaches, like SCALPEL, are 4× technologies. DOF is large and may reach up to 500 µm. The optics in IPL are all electromagnetic, and the

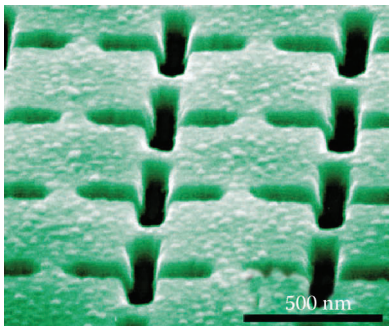


FIGURE 2.26 IPL mask 80-nm feature sizes etched into a Si membrane mask. The patterns were defined using a Leica VB6 EBL. (From http://www.cnf.cornell.edu/image.)

potential for a 50 × 50-mm exposure field exists. The membranes for this work typically consist of low-stress silicon nitride or Si. An example of a Si stencil mask created at the Cornell Nanoscale Science and Technology Facility for masked ion-beam lithography is shown in the electron micrograph in Figure 2.26.

Ion-beam lithography has at least two advantages over EBL: 1) it has almost two orders of magnitude higher resist sensitivity, and 2) it has negligible ion scattering in the resist and very low backscattering from the substrate. A major problem is the potential for damage to sensitive electronic functions from the high energy ions.

Efforts in IPL are especially strong in Europe (e.g., at IMS, http://www.ims.co.at).

Imprint Lithography

Nanoimprint lithography (NIL) and step-and-flash imprint lithography (SFIL) are techniques that use hard molds instead of the soft molds used in soft lithography (see below). Stephen Chou at Princeton University invented NIL in 1994, with the aim of overcoming the diffraction-limited minimal feature sizes obtained in semiconductor manufacturing based on DUV lithography (http://www.princeton.edu/~chouweb). It is possible to break through the diffraction-controlled size limit of photolithography, for example, by using electron beam lithography (EBL). But the infrastructure of this approach, we know by now, costs billions of dollars. NIL patterns a resist by deforming the resist shape through embossing (with a mold/stamp/template) rather than by altering resist chemical structures through radiation (with particle beams). After imprinting the resist, a dry anisotropic etch is used to remove the residual resist layer in the compressed area to expose the substrate underneath. In NIL, a template (the mold/stamp/template) is made of a hard material (usually Ni or Si) and is pressed against a layer of polymer. High temperature and pressure conditions mold and harden the polymer layer. As illustrated in Figure 2.27, the NIL technique is based on the craft of hot embossing, with an adaptation to modern semiconductor needs. In this example the technique is used to pattern a metal film via liftoff. The method relies on the excellent replication fidelity obtained with polymers and combines thermoplastic molding with common pattern transfer methods. Once a solid stamp with a nanorelief on the surface is fabricated, it can be used for the replication of many identical surface patterns. The resolution of the NIL process is a direct function of the resolution of the original template/stamp fabrication process.

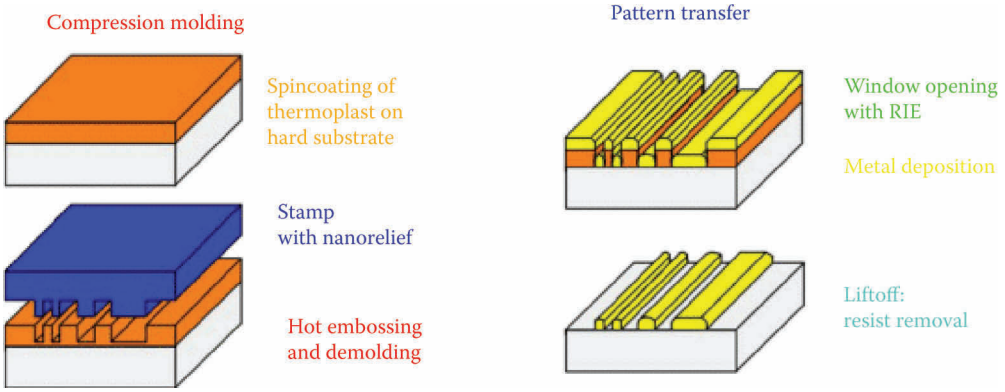


FIGURE 2.27 Schematic illustration of the nano imprint lithography (NIL). The technique is used here to pattern a metal film via liftoff.

Electron beam writers that provide high resolution but lack the throughput required for mass production are used to make them.

The University of Texas–Austin developed its own version of nanoimprint lithography, i.e., step-and-flash imprint lithography (SFIL), in 1998.[47] In 2001, the SFIL concept was licensed to Molecular Imprints, Inc., a company that develops SFIL-capable tools (S-FIL® systems) and related processes (http://www.molecularimprints.com). The SFIL method is distinct from the original NIL in its use of UV-assisted nanoimprinting that molds photocurable liquids in a step-and-repeat, die-by-die fashion rather than by heat-assisted molding of full, polymer-coated wafers. As shown in Figure 2.28, in SFIL, a hard but transparent template/stamp/mold (fused silica) is used to mold a polymer photoresist layer. The fused silica surface of the stamp, coated with a release layer, is gently pressed into the thin layer of low viscosity photoresist. The photoresist is subsequently exposed to UV light through the transparent template to harden it. On separation of the fused silica template, one layer of the circuit pattern is left on the wafer surface. Molecular Imprints uses a family of photocurable, low viscosity materials called MonoMat™ as its imprint resists. MonoMat™ is composed of an organic monomer that polymerizes in seconds using low-cost, broadband UV light sources. Whereas other imprint tools use spin-on imprint resist, S-FIL® systems dispense low viscosity imprint resist in very small droplets, one field at a time, immediately before the imprint step. Consequently, both the monomer dispenses and imprint functions happen in a step-and-repeat fashion. A residual layer of polymer between features is eliminated by a dry etch process, and a perfect

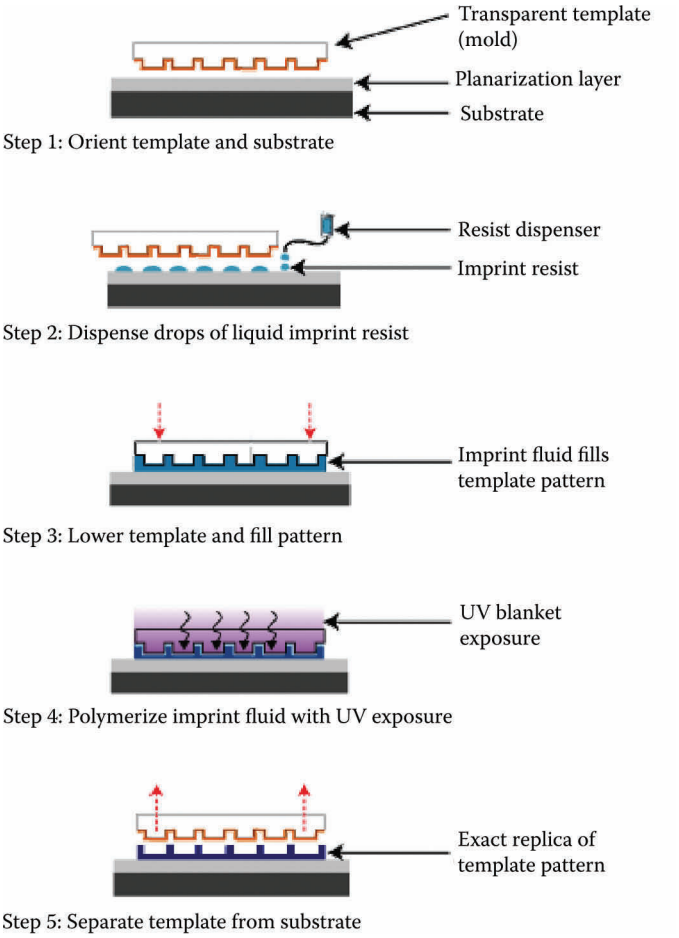


FIGURE 2.28 Schematic of step-and-flash imprint lithography (SFIL). (Colburn, M., S. Johnson, M. Stewart, S. Damie, T.C. Bailey, B. Choi, M. Wedlake, T. Michaelson, S.V. Sreenivasan, J. Ekerdt, and C.G. Wilson. 1999. Step and flash imprint lithography: a new approach to high-resolution patterning. *Proc SPIE* 3676:379–89.[47])

replica of the pattern is ready to be used in semiconductor processing for etch or deposition. Only the template fabrication process, typically accomplished with an e-beam writer, limits the resolution of the features. It has been demonstrated that the mold templates do not deteriorate even after 1500 imprints with sub-100-nm feature sizes. Although more progress is needed for widespread adoption of NIL and SFIL, sub-100-nm feature size devices can be reliably fabricated at a reasonable throughput. Implementing SFIL for high-volume manufacturing of the gate or contact level of transistors is not feasible in the near term, but it is one of the few methods currently available for low-volume prototyping at the 32-nm node. Sub-20-nm features have been made already that exceed the present requirements of the ITRS. For example, using Molecular Imprints' Imprio 250 system, Toshiba was able to print 18-nm isolated features and 24-nm dense features with <1-nm critical dimension uniformity and <2-nm line edge roughness.[48] The Toshiba demonstrations also showed lower defectivity and improved results for imprint lithography. Defectivity levels of as low as <0.3 defects/cm^2 are approaching those of immersion lithography, and device overlay results were also within Toshiba's required specifications. However, the approach faces alignment challenges, and critics say the cost of making the reticles will be high. Unlike conventional lithography, with mask patterns that are four times larger than the printed image, nanoimprint is a 1:1 template technology.

Others argue that imprinting processes are economical because of the simplicity of the equipment involved and the potential for high throughput. As the rate of improvements in optical lithography decelerates and as the costs of manufacturing continue to escalate, there is an increased interest in printing and molding as alternative processes for microfabrication. As imprint lithography depends on another lithography technique (perhaps an NGL) to make the master, it does not really qualify as an NGL itself. Whereas it may initially have targeted semiconductor applications, the SFIL process is now being explored for manufacturing of several emerging technologies, such as photonic crystals, micro/nano-optical components, and nanopatterned magnetic media for future hard disk drives. Progress in these new areas has been such that SFIL is likely to find its first commercial manufacturing application in one of these emerging technologies well before it would be required for high-volume, sub-50-nm semiconductor lithography. A notable feature of nanoimprint technologies is their relatively low cost, which allows researchers to explore applications of nanopatterning that would never be economically feasible given the extraordinary cost associated with EUV lithography or even current-generation 193-nm steppers. This is very similar to the current applications of soft lithography (see below) with this difference that using a hard stamp renders NIL and SFIL more industrial manufacturing capable.

Thermal NIL using thermoplastic polymer films is focused on applications such as biochips, life sciences, storage media, and optical devices, whereas SFIL is focused on LSI and magnetic storage applications, as summarized in Figure 2.29. Hitachi, with NEDO support, has also developed a sheet or roll-to-roll prototype NIL machine and demonstrated the capability to process 15-m sheets.[49] The goal is to create a continuous imprinting process by using belt molding (nickel-plated molds) and polystyrene sheets for large geometry applications, such as membranes for fuel cells, batteries, and possibly for displays. Currently the prototype tool does not

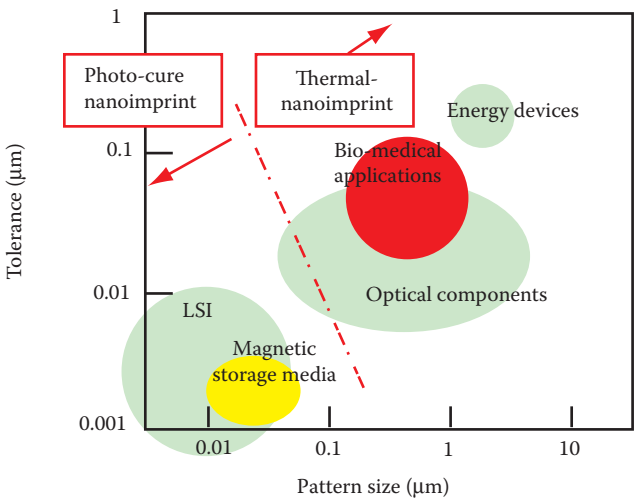


FIGURE 2.29 Usage domains of photocure (SFIL) and thermal nanoimprinting (NIL). (Based on a figure by Hitachi.)

offer the desired throughput, but the concept has been demonstrated, with future efforts focused on throughput, reliability, and repeatability.

Lithography Techniques in the Research and Development Stage

Introduction

In this section we cover the more futuristic lithography methods, some of which could cause a paradigm shift in the improvement of CD printing capability by perhaps a factor of 10. These might eventually enable devices that are a few nanometers in size and have switching speeds in the terahertz range. We review lithography with ultrathin resist layers and block copolymers, zone plate array lithography (ZPAL), proximal probe-based lithographies such as atomic force microscopy (AFM), scanning tunneling microscopy (STM), dip-pen lithography (DPL), quantum lithography (two-photon 3D lithography), near-field scanning optical microscopy (NSOM), apertureless near-field scanning optical microscopy (ANSOM), holographic lithography, plasmonic lithography, and lithography with superlenses. In terms of an important lithography method that is capable of patterning resist on nonplanar substrates (e.g., patterning a resist on the surface of a cylinder), we detail soft lithography, in which a soft mask can be bent around nonplanar substrates.

Very Thin Resist Layers

Introduction

Radiation scattering during exposure of a resist limits its resolution to no better than the resist thickness. Using ever thinner resist layers further improves the resolution and renders a low DOF—associated with high NA lenses and short wavelength lithography—less critical. It is in this context that physisorbed Langmuir-Blodgett (LB) resists, self-assembled chemisorbed monolayer (SAM) resists, and ultrathin film (UTF) resists are explored.

Langmuir-Blodgett Resists

LB resist films were considered for some time for the finest lithography applications such as ultraresolution nanolithography. In this method, LB resist films are prepared by transferring organic monolayers floating on a water surface onto solid substrates. The procedure is explained in Chapter 8 on chemical, photochemical, and electrochemical-based forming (see Figure 8.13). Multilayer films of PMMA deposited by the LB technique, for example, have been investigated as a potential e-beam resist for nanolithography. The range of backscattered electrons limits the resolution of e-beam resists through the proximity effect, and nanometer-thick layers minimize this effect. Despite extensive work in the early 1980s to automate LB resist deposition systems, the technology is not envisioned as a viable manufacturing option because of its low mechanical and chemical stability.[50,51] Other major problems concerning LB films need to be addressed before the method can become viable, and research in this area focuses on speeding up the coating process, simplifying surface cleaning procedures, and improving etching resistance. Because the bonding of an LB film with the substrate is based on physisorption, the adhesion is weak, and chemisorbed SAMs have become much more popular (see next section).

Self-Assembled Monolayers

Molecular self-assembly is a chemical process in which molecules spontaneously organize to form larger ordered structures. Self-assembly is one aspect of supramolecular chemistry, i.e., chemistry beyond the molecule or the chemistry of the intermolecular bond.[52] Self-assembly is studied extensively with an eye toward mimicking nature's spectacular use of this bottom-up, automated manufacturing approach (see Volume III, Chapter 4 on packaging, assembly, and self-assembly). In one special type of self-assembly, self-assembling precursor molecules, from a solution or the vapor phase, react at interfaces to produce layers of monomolecular thickness that are chemically bonded to solid surfaces. Such layers belong to a class of materials known as *self-assembled monolayers*. The thermodynamically favorable bond formation involves chemisorption, which results in monolayers that are more stable than those possible with physisorbed LB films (see above). Therefore, SAMs potentially make for better resist candidates.

Homogeneous and densely packed molecular layers incorporate reactive groups that form bonds with the substrate on one side, whereas an organic group (e.g., an alkyl group, R) on the other side imparts the desired chemical functionality to the surface modified with the thin film. Changing even one atom at the end group of self-assembling molecules is sufficient to dramatically alter macroscopic properties such as wettability, biocompatibility, and adhesion. For example, fluorinated SAM precursors form monolayers ~1-nm thick with the same low wettability and resistance to adhesion typical of thick samples of Teflon™.[53] Many SAM films can be molecularly engineered to be patterned by various types of energetic radiation, including DUV, soft x-rays, ion beam, and low-energy electrons.[54] Scratching and microcontact printing (see below) can also be used to pattern SAMs. Structures less than 20 nm have been obtained by exposure with STMs and conventional EBL systems.[55] SAMs have successfully been deposited on metals like gold, aluminum, titanium, zirconium, silver, copper, and platinum, as well as on SiO_2, GaAs, and other surfaces.[53] The quality of the films depends strongly on surface pretreatment. Pinhole density on a Au surface of less than 5/mm^2 has been reported (see Müller and references therein[55]). Alkane-thiols and dialkyldisulfides are typical precursor materials for SAMs on gold, the most extensively studied substrate. The formation of long chain ω-substituted dialkyldisulfides on gold was first demonstrated in 1983.[56] Films of better quality are formed by the adsorption of alkylthiols.[57] Alkanethiols and dialkyldisulfides are lipid-like organic molecules having the general formula HS-$(CH_2)_n$-X and X-(CH_2)-SS-$(CH_2)_m$-Y, respectively, where n and m indicate alkyl chain length and X,Y the end groups (e.g., -CH_3, -azobenzene, -OH).[58] Organosilanes are often used to form SAMs on Si. These molecules form a Si-O-substrate siloxane bond on the Si surface, and the alkyl group R is responsible for the ordered nature of the film. For $n > 6$ (n is number of carbon groups in the R chain), SAMs have substantial crystalline order at the air/monolayer interface.[53, 58]

Microcontact printing, discussed below, also takes advantage of SAMs, and so does DNA and protein patterning, discussed in Chapter 8. Microcontact printing can be considered a merging of top-down manufacturing techniques (i.e., traditional lithography used to make the elastomeric stamp) with bottom-up manufacturing (i.e., proteins used to ink the stamp as building blocks). Protein and DNA patterning has evolved into an important research application for lithography. It is easy to imagine arrays with several different enzymes, antibodies, or DNA probes immobilized precisely onto a small transducer surface as a diagnostic panel for clinical applications. Different approaches for patterning organic materials are reviewed in Chapter 8.

Ultrathin Film Resist Layers

Polarity changes are at the heart of the newest ultrathin film (UTF) resist strategies. UTF resists represent the next evolutionary step in surface imaging technologies (see Chapter 1). In UTF, only the top monolayer of the resist changes, opening up the potential for further resolution improvement. Calvert et al.[59] use a few monolayers (<10 Å) of organosilanes, which are chemisorbed onto a substrate. These chemisorbed films are easier to prepare and more robust than the physisorbed LB films (see above). DUV irradiation of these films cleaves organofunctional groups from the film and produces an extremely hydrophilic surface (water contact angle < 10°) to which colloidal Pd/Sn catalyst does not adhere. Subsequent electroless deposition occurs only on those surface regions that were not irradiated. Thus, patterns of electroless Ni, Co, and Cu are produced on a variety of substrates, including silicon, quartz, alumina, metals, and polymers. These metal patterns serve as an efficient plasma-etching mask for pattern transfer. Fabrication of 0.3-μm line widths has been demonstrated in this way.[50, 59]

Block Copolymer Lithography

Block Copolymers Background

Self-assembly of block copolymers is widely investigated as a simple, inexpensive means for patterning high-density arrays of nanoscale features.[60] Block copolymers comprise two or more different monomer units, strung together in long sequences, which

can self-assemble into highly ordered lattices with unit cell dimensions of 10–100 nm. This length scale reaches well below the limits of conventional optical lithography. The simplest block copolymer architecture is the diblock. In diblock copolymers, two different types of polymer chains are connected at one end with a covalent bond (in a triblock copolymer, three different polymer chains are involved, etc.). The immiscibility between the unlike segments drives them to self-assemble into periodic domains, a process termed "microphase separation." This self-assembly process is driven by an unfavorable mixing enthalpy and a small mixing entropy while the covalent bond between the two blocks prevents macrophase separation. The size and shape of the periodic domains are dependent on the molecular weight and composition of the copolymer and typically assume morphologies of spheres, cylinders, and lamellae. These morphologies that represent different phases are dictated by the Flory-Huggins interaction parameter χ, the degree of polymerization N (i.e., number of monomers per chain, proportional to the polymer molecular weight), and the volume fraction f of the blocks.[61] The Flory-Huggins interaction parameter χ is a dimensionless parameter, which expresses, in units of kT, the energy change when one takes a molecule of pure A from an environment of pure A and puts it in an environment of pure B; it is inversely proportional to temperature and reflects the interaction energy between the different segments of the polymer chains. As χ increases, phase separation becomes more likely, and as N increases, phase separation becomes less so. From Figure 2.30, we glean that if the volume fractions of the blocks are close to equal, a layered morphology is observed. When moving toward less equal block ratios, the observed morphologies go through a bicontinuous gyroid structure, hexagonally packed cylinders, and finally body-centered cubic-packed spherical domains. Thus, microphase separation leads to a wonderful variety of exotic ordered phases, increasing rapidly with the number

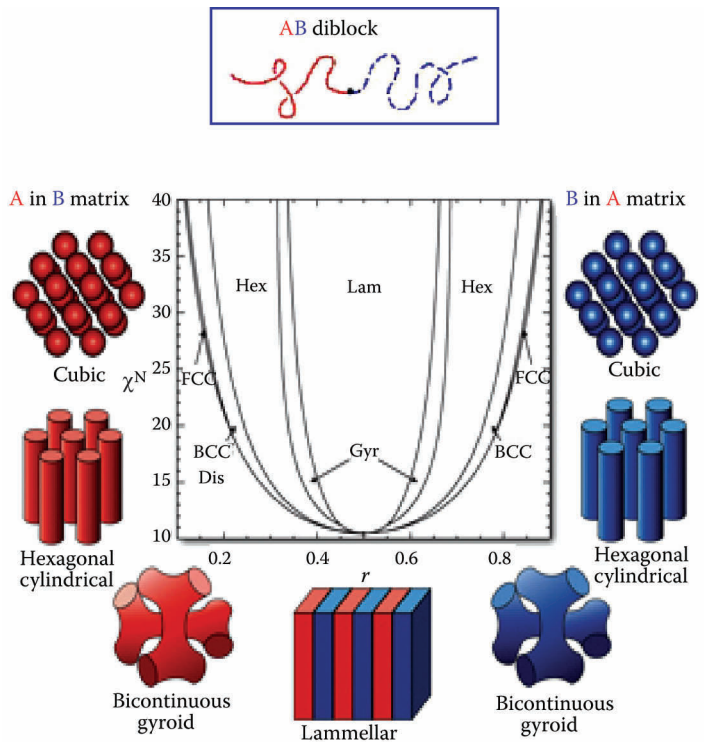


FIGURE 2.30 Schematic diblock copolymer phase diagram: f = volume fraction of A in B, χ = Flory-Huggins interaction parameter, N = diblock degree of polymerization. Known equilibrium mesophases are spheres, cylinders, gyroid, and lamellae, as well as the disordered (homogeneous) state at small interblock segregation strength (χN). (Diagram adapted from Matsen, M.W., and F.S. Bates. 1996. Unifying weak- and strong-segregation block copolymer theories. *Macromolecules* 29:1091–98.[62])

of blocks (with a triblock copolymer, a veritable "zoo" of structures results).

Block Copolymers Lithography

The previous section applied to bulk block copolymers. What about applying thin layers of block copolymers to surfaces? On surfaces, the microphase separation also results in nanoscale structures (domains) with sub-100-nm length scales, unless the surface has too strong an interaction with the polymer blocks that it prevents them from assembling. Block copolymer lithography refers to the use of block copolymers in the form of thin films, in which the domain structure provides a template for additive or subtractive pattern transfer operations, as illustrated in Figure 2.31. The idea of using block copolymer thin films as lithographic masks, as shown here, was first proposed in 1995 by Mansky et al.[63] In Figure 2.31, we compare traditional lithography with block copolymer lithography. The material chosen for the block copolymer lithography demo is from IBM work, where a diblock copolymer, in which polystyrene and PMMA are tied together in one chain (referred to as PS-b-PMMA, with total molecular weight in the range of 60,000 g/mol), is used.[64] When spun onto the surface of a silicon wafer, the two polymers separate, and although the molecules stretch out, the chemical bonds keep them attached. Subsequent heat treatment above the polymer glass transition temperature completes the microphase separation. Typical annealing conditions are 200°C (in a vacuum or N_2) for times ranging between 30 minutes and 24 hours, depending on the application. In the end, PMMA ends up concentrated in small cylinders surrounded on all sides by polystyrene. Thus, the diblock copolymer forms on its own into a nearly

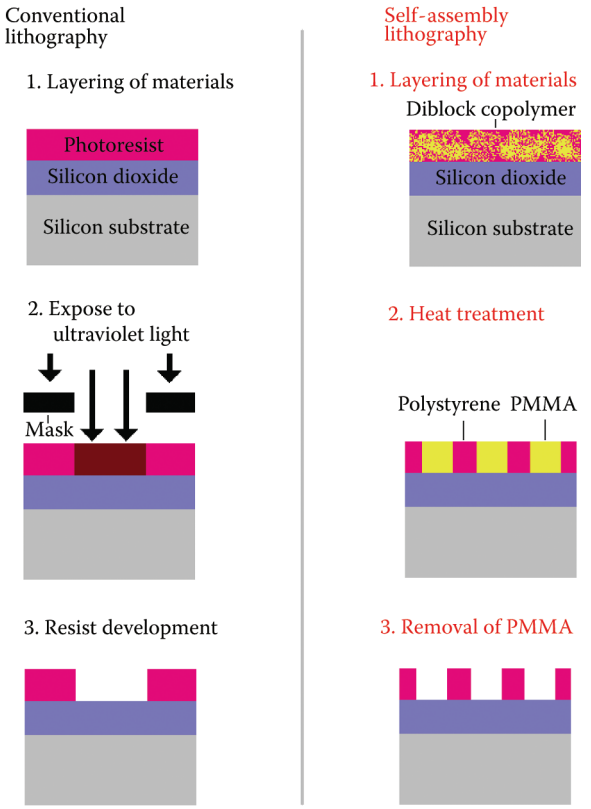


FIGURE 2.31 Traditional lithography compared with block copolymer lithography. Conventional lithography exposes a photoresist to ultraviolet light. A solvent etchant then removes the exposed part of the photoresist (for a positive type photoresist). Self-assembly patterning occurs when a diblock copolymer is heated, thereby separating the two polymers in the material into PMMA and polystyrene domains. The PMMA can then be selectively etched away, and the template of cylindrical holes can be transferred into the silicon dioxide. (Adapted from Stix, G. 2004. *Nano patterning: IBM brings closer to reality chips that put themselves together.* http://www.fractal.org/Fractal-Research-and-Products/Nano-patterning.htm.[65])

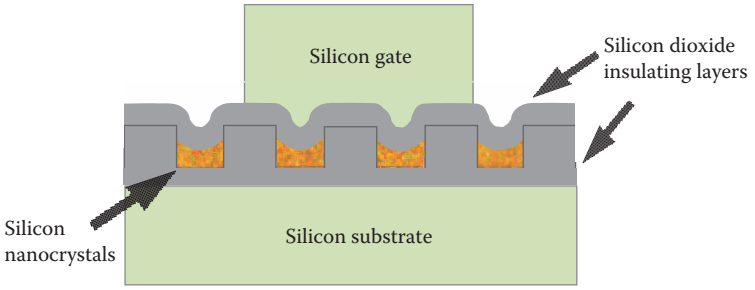


FIGURE 2.32 Flash memory: a layer of self-assembled silicon nanocrystals is inserted into an otherwise standard device as part of a novel IBM manufacturing process. (Based on Stix, G. 2004. *Nano patterning: IBM brings closer to reality chips that put themselves together*. http://www.fractal.org/Fractal-Research-and-Products/Nano-patterning.htm; and Black, C.T., R. Ruiz, G. Breyta, J.Y. Cheng, M.E. Colburn, K.W. Guarini, H.-C. Kim, and Y. Zhang. 2007. Polymer self assembly in semiconductor microelectronics. *IBM J Res Dev* 51: 605–33.[65,64])

perfect honeycomb-like template without the need of UV exposure or a mask.

Finally, the IBM team created an array of 20-nm-wide pores in the polystyrene matrix by removing the PMMA using an organic etching solvent (acetic acid). The same team used the thus-obtained perforated polystyrene structure to create the flash memory cells illustrated in Figure 2.32.[64] After the PMMA etch, a subsequent etching step was used to transfer the same honeycomb pattern into an underlying layer of silicon dioxide. Next the polystyrene template is removed, and subsequently, nanocrystal formation is accomplished by conformal a-Si (amorphous or nanocrystalline Si)* deposition over the whole wafer, and a dry anisotropic a-Si etch is used to remove all the deposited silicon except for that deposited in the holes in the oxide matrix. All that is left in the end are nanocrystalline cylinders surrounded by silicon dioxide. The final steps place an insulating layer of silicon dioxide (7-nm-thick control or program oxide) over the top of the Si nanocrystal islands and a polysilicon control gate is defined on top of the control oxide using conventional lithography. These flash devices are programmed by injecting charge into the nanocrystals (through the program oxide) and erased by expelling charge from the nanocrystals.

Creating closely spaced holes for a flash memory as demonstrated here is exceedingly difficult to accomplish with ordinary lithographic and deposition methods. Forming nanocrystalline silicon using these conventional techniques creates elements of different sizes. In contrast, the self-assembled nanocrystals are evenly spaced and of uniform size, improving their durability and their capacity to retain a charge while allowing the cylinders to shrink to dimensions even smaller than 20 nm.

For most lithography applications, we must also be able to control the orientation of the self-assembled pattern with respect to the underlying substrate surface. Whereas the self-assembly process itself is driven by microphase separation of the constituent copolymer blocks, the ultimate pattern orientation is determined by the relative strength of the surface affinity for each block. Control of block copolymer film thickness plays a critical role in determining vertical and lateral ordering and alignment of the nanostructures. A variety of approaches for controlled domain orientations have been attempted, including topographic prepatterns and/or chemical prepatterns, an external electric field, thermal field, and several other methods as reviewed by Segalman,[66] Hawker and Rusell,[67] Black et al.,[64] Stoykovich and Nealy,[68] and Hamley.[60]

In Figure 2.33, the patterned wettability (polar vs. nonpolar stripes) of the surface is used to align the nanodomains. Substrates are often preferentially wetted by one of the blocks of the copolymer, and the domains, e.g., lamellae or cylinders, tend to orient parallel to the substrate (see Figure 2.33a and b, where all the lamellae align parallel to

* Nanocrystalline silicon (nc-Si) is an allotropic form of silicon that is similar to amorphous silicon (a-Si), in that it has an amorphous phase. However, nc-Si has small grains of crystalline silicon embedded within the amorphous phase. This is in contrast to polycrystalline silicon (poly-Si), which consists solely of crystalline silicon grains, separated by grain boundaries. The term nanocrystalline silicon refers to a range of materials around the transition region from the amorphous to the microcrystalline phase in a silicon thin film.

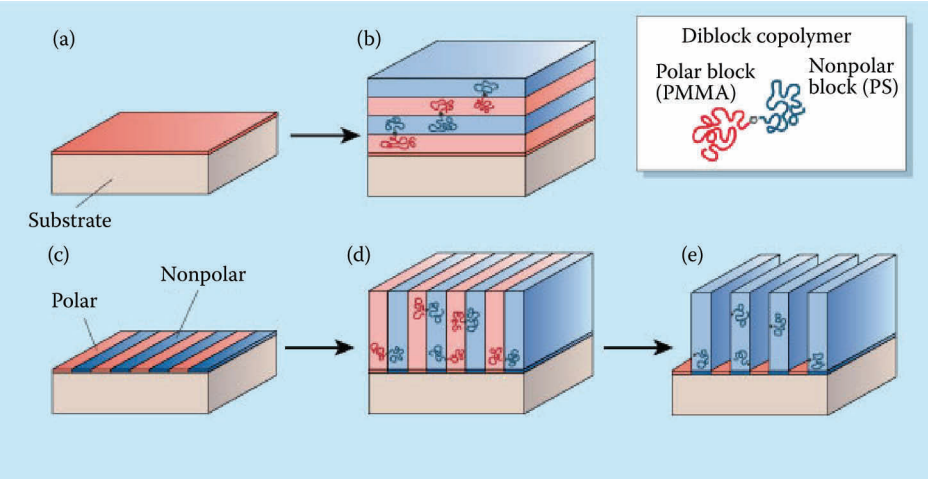


FIGURE 2.33 Patterned wettability of the surface causes the alignment of the nanodomains of a block copolymer. PMMA, polymethylmetacrylate; PS, polystyrene.

the substrate). Chemical modification can be used to neutralize or overcome surface and interfacial forces that tend to drive the domains to form parallel to the plane of the film. Controlling the interactions at the interfaces and scaling of the prepatterns can achieve parallel and perpendicular alignments of the lamellae achieved as shown in Figure 2.33c and d: the polar PMMA block aligns over the polar stripes and the nonpolar polystyrene blocks over the nonpolar stripes. The block copolymers on the surface arranged themselves into the underlying pattern without imperfections. In Figure 2.33e we show a finished structure after etching away the PMMA.

In a significant accomplishment, Kramer's group[69] (University of California, Santa Barbara) has patterned single crystals of 10-nm block copolymer spheres across an entire 2-in.-diameter silicon wafer. The silicon wafer was first patterned with photolithography to produce 12-µm-wide hexagonal wells. A single layer of block copolymer spheres was then self-assembled inside those wells. Assessing the quality of the self-assembly of the block copolymer in perfect crystal patterns is very challenging because of the small feature size. Whereas high-resolution microscopy can resolve individual block copolymer domains, this technique is limited to extremely small sampling areas (~10^{-6} mm^2). The Santa Barbara team used grazing-incidence small-angle x-ray scattering (GISAXS) to check for the crystal quality over macroscopic areas (~2 mm^2), and they established it as nearly perfect, demonstrating that single crystals resulted in each well. These results clarify why copolymers constitute a viable avenue toward building photonic crystals.

Summary/Applications

The obvious interest in using block copolymers for patterning is derived from the fact that they can self-assemble to form dense arrays of nanostructures with dimensions and spacings that are difficult or even impossible to create by other means or are prohibitively expensive to fabricate using conventional lithographic materials and processes. Polymer self-assembly can define only a limited set of pattern geometries. However, within these constraints the materials provide a straightforward means for achieving feature sizes (<20 nm), pitches (<40 nm), and densities (~$10^{11}/cm^2$) that are currently not achievable by optical lithography.

The use of block copolymer films to pattern dense arrays of features is more than a decade old, and other reviews have covered this topic in more detail (for example, see Segalman,[66] Hawker and Rusell,[67] Stoykovich and Nealy,[68] and Hamley[60]). These review articles reveal that patterns of dense periodic arrays fabricated with block copolymer templates have been demonstrated, and in some cases are being commercialized for the fabrication of quantum dots, magnetic storage media, flash memory devices (see Figure 2.32), semiconductor capacitors, nanowires, photonic crystals, and

nanopores. At IBM, a polymer self-assembly process was successfully integrated into a 200-mm semiconductor fabrication facility at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York.[64]

Zone Plate Array Lithography

Fresnel Zone Plate

Before getting into the details of zone plate array lithography (ZPAL), we take a look at the building blocks of a ZPAL, i.e., Fresnel zone plates. A Fresnel zone plate, invented by the French physicist Augustin-Jean Fresnel, is a single plate consisting of alternating opaque and transparent circular zones as illustrated in Figure 2.34. Zone plates use constructive interference of light rays from adjacent zones (Fresnel zones) to form a focus. The zones are spaced so that diffracted light constructively interferes at the desired focus. The Fresnel zone plate is a relative of the pinhole camera in that it does not use mirrors or lenses for its imaging properties. Because a diffraction lens can be manufactured using planar technologies, it is easier to make than a traditional refractive lens. The zone plate is especially useful in the UV and x-ray regions of the spectrum, for which other imaging devices are hard to find. Self-supporting gold zone plates have been manufactured for these spectral regions. Moreover, it comes with a higher refractive index, allowing for better contrast at the image plane. Unfortunately, the zone plate has low efficiency and suffers from veiling glare because most of the light incident on the zone plate passes through it undiffracted and falls onto the image plane.

The focal length f of a zone plate with many zones is a function of the radius of the outermost zone, r_n, and its width, Δr_n, where n is the number of zones and depends also on the wavelength λ (see Figure 2.34), or:

$$f = \frac{2r_n \Delta r_n}{\lambda} \tag{2.3}$$

To get constructive interference at the focus, the zone boundaries in a zone plate should switch from opaque to transparent at radii where:

$$r_n^2 = n\lambda f + \frac{n^2\lambda^2}{4} \tag{2.4}$$

$$r_n = \sqrt{n\lambda f + \frac{n^2\lambda^2}{4}} \tag{2.5}$$

and simplifying when the zone plate is small compared to the focal length:

$$r_n = \sqrt{n\lambda f} \tag{2.6}$$

where n is a positive integer, r_n is the radius of the nth zone, f is the primary focal length, and λ is the wavelength of the illumination source. A Fresnel zone plate functions very much like an optical lens. Its optical properties do not depend on the material used, though, but are determined at the design time. For example, the diffraction limited resolution for

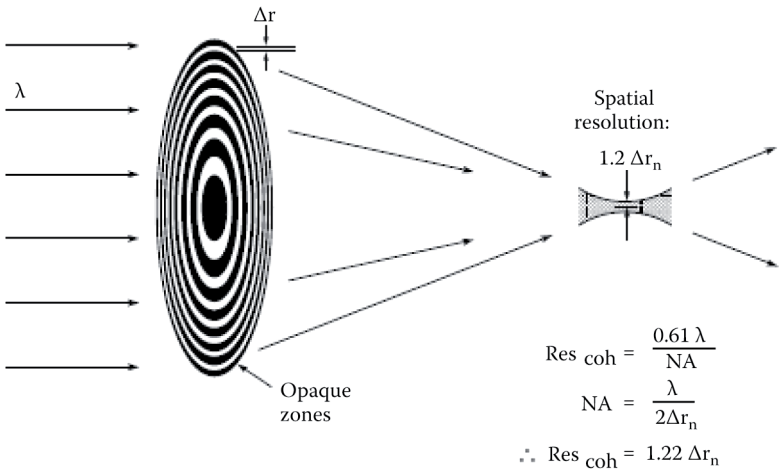


FIGURE 2.34 A Fresnel lens. (From http://xradia.com.)

coherent light Res_{coh} of a zone plate is governed by the Rayleigh principle as:

$$\text{Res}_{coh} = 0.61\frac{\lambda}{NA} \text{ and with}$$

$$NA = \frac{\lambda}{2\Delta r_n} \text{ we obtain:} \quad (2.7)$$

$$\text{Res}_{coh} = 1.22\Delta r_n$$

The resolution depends only on Δr_n, the outermost zone width. NIL and EBL can be used for the fabrication of large arrays of high NA diffractive Fresnel lens arrays.

In optical lithography, further increasing the NA of a lens is difficult and costly (see Figure 1.80 where a maximum NA of around 0.9 is projected). In diffractive optics, making a high NA lens is not significantly different than making a low NA lens. At MIT, Fresnel lenses with an NA of 0.7–0.9 have been achieved.[70]

Zone Plate Array Lithography Detailed

Zone plate array lithography (ZPAL) requires no masks but rather uses arrays of individually targetable optical beams as shown in Figure 2.35. Beamlets of photons, which can be rapidly turned on and off,

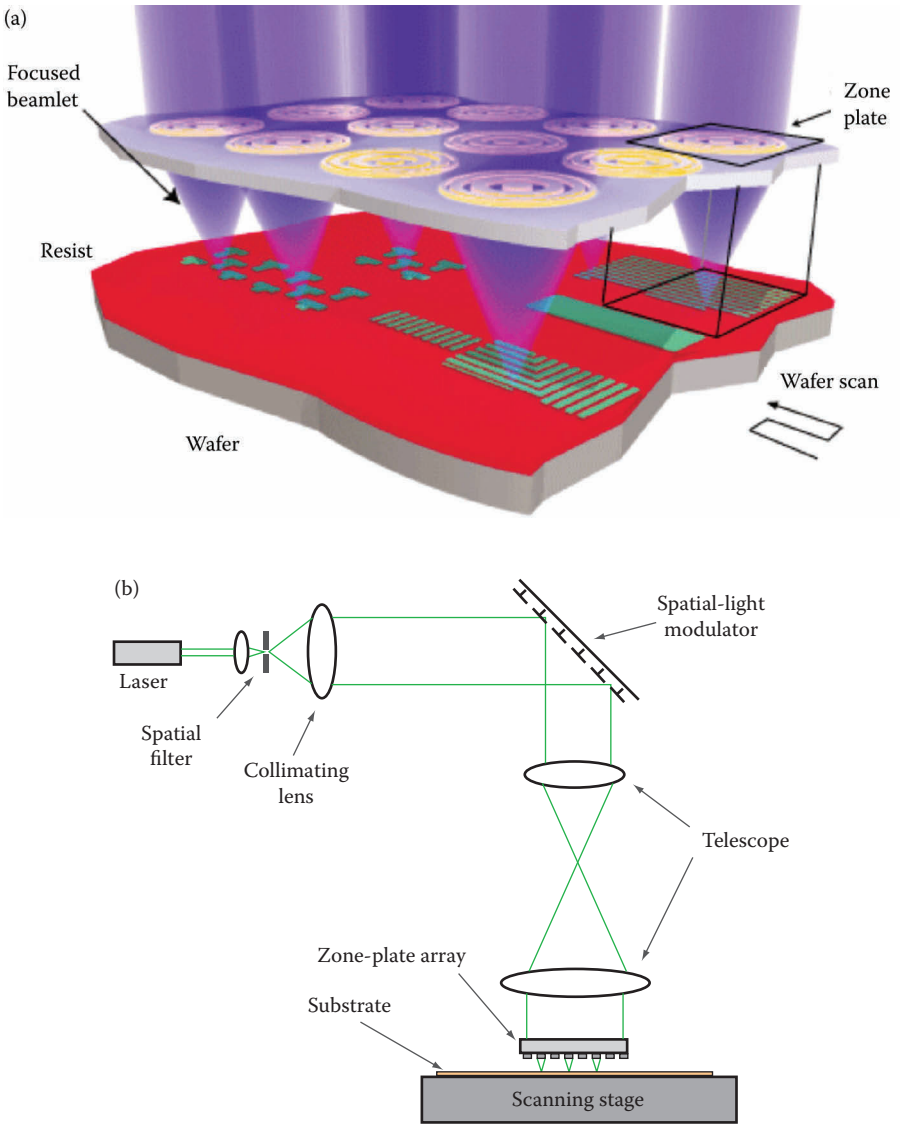


FIGURE 2.35 (a) Lumarray (http://www.lumarray.com) has demonstrated fully multiplexed ZPAL writing. (b) A laser beam is passed through a spatial filter and a collimating lens onto a digitally controlled programmable spatial light modulator (SLM), which replaces the photomask.[70,71]

are projected through diffractive Fresnel zone plate lenses (see above), allowing myriad complex shapes to be fabricated. The zone plates are made using lithography techniques, and the shutters for each beamlet are micromechanical movable mirrors. ZPAL was originally developed at MIT's nanostructures laboratory in the mid to late 1990s.[70] As shown in Figure 2.35a, each zone plate is responsible for one unit cell of photoresist exposure corresponding to the diameter of one individual zone plate. The writing with the shutters on and off results in dot-matrix-type lines, and moving the wafer in a serpentine fashion under the focused beamlets results in a full pattern.

ZPAL is an example of maskless optical projection lithography (MOPL, see Chapter 1); a laser beam is passed through a spatial filter and a collimating lens onto a digitally controlled programmable spatial light modulator (SLM), which replaces the photomask (see Figure 2.35b). In one of the MOPL systems described in Chapter 1, the SLM used is the Texas Instrument micromirror array (see DMD™, Example 7.10). In operation, a desired pattern is fed into the SLM, and the movable mirrors create the desired image. This image is demagnified and focused by a single reduction lens on the wafer to expose the photoresist. The use of refractive optics prevents the extension of this type of MOPL to wavelengths much shorter than 157 nm.

In the ZPAL system, the MIT team uses the Grating Light Valve™ (GLV)* from Silicon Light Machines (http://www.siliconlight.com) as the light modulator. In contrast to MOPL, not a single refractive reduction lens is used. Instead, an array of (high NA) diffractive Fresnel lenses ("the zone plates") is used to focus the incident light onto an array of spots on the photoresist-coated substrate. The light source in the ZPAL system preferably is a continuous wave laser (a pulsed laser source, as used in MOPL, severely limits system throughput). Light from the laser is passed through a collimating lens to create a uniform beam incident on the SLM, and the SLM breaks that beam up into individually controllable beamlets. A telescope is placed such that each beamlet is normally incident on one zone plate in the array. A Fourier filter within the telescope ensures that there is sufficient contrast between the on and off states. Each zone plate behaves like a microscopic lens focusing the light into a tight on-axis spot in its focal plane. The diffractive Fresnel lens optics in ZPAL allow the technique to scale to much shorter wavelengths than what is possible with MOPL (even x-rays are feasible).

In Chapter 1 we used a dimensionless constant k_1 to define the practical limiting resolution R in projection printing as:

$$R = k_1 \frac{\lambda}{NA} \tag{1.16}$$

where k_1 depends on resist parameters, process conditions, mask aligner optics, and so on. The MIT team and a startup company called Lumarray (http://www.lumarray.com) have demonstrated fully multiplexed ZPAL writing, multilevel alignment, and resolution corresponding to $k_1 < 0.3$ [this is without any resolution-enhancing technology (RET) applied to it].[71] In theory, ZPAL could use 193-nm optical lithography to inexpensively produce 45-nm chips. Moreover, this technology should theoretically work with 157-nm lithography and EUVL to create 20-nm ICs. ZPAL's main competition at this point is regular 193-nm lithography enhanced with immersion lithography, a combination that could etch 45-nm features (see Chapter 1). ZPAL, however, might be a less-expensive and an easier solution. As in optical projection lithography, to increase NA, immersion lithography can also be applied to ZPAL; in experiments at MIT, a resolution improvement of 1.36 was achieved this way.[71]

Quantum Lithography

Background

In optical projection lithography, a reduced-size image of complicated patterns is reproduced onto a photoresist. If light always followed the law of geometrical optics, there would be a point-to-point

* The Grating Light Valve™ (GLV) technology comprises a series of microscopic ribbons on the surface of a silicon chip; the GLV device is a diffractive MEMS that acts a dynamic, tunable grating that can switch, attenuate, and modulate laser light.

relationship between the object (mask) and the image planes (resist). But, as we saw in Volume I, Chapter 5, light is a wave, and resolution is limited to a finite spot size, defined by the "point-spread function". This point-spread function determines the spatial resolution of the imaging setup and limits the ability to produce demagnified images (see Rayleigh limit of 0.6 λ/NA, Volume I, Chapter 5). It has been demonstrated that quantum lithography with entangled *N*-photon states beats this Rayleigh diffraction limit by a factor of *N*; thus, in two-photon lithography the resolution is improved by a factor of two (as if one used a classical source with wavelength λ/2).[72,73] Einstein, Poldosky, and Rosen (EPR) described the entangled two-particle state according to the principle of quantum superposition in 1935, and they pointed out a surprising consequence: the momentum ΔP_x (position Δx) for neither photon is known (see Heisenberg uncertainty principle, Volume I, Chapter 3); however, if one particle is measured to have a certain momentum (position), the momentum (position) of its "twin" is known with certainty, despite the distance between them (this is known as the EPR paradox).[74] The entangled photon pairs come out from a point of the object plane, undergo two-photon diffraction, and result in twice-narrower point-spread function on the image plane as shown in Figure 2.36.

Two-Photon 3D Lithography

Two-photon lithography is an intrinsic 3D lithography technique that enables the fabrication of structures difficult to produce by conventional single-photon processes and with far greater spatial resolution than most other 3D microfabrication techniques. As we discuss in Chapter 8 in the section on stereolithography, the use of laser rapid prototyping by polymerizing a resin point by point enables the computer-aided manufacture (CAM) of 3D polymer MEMS structures. Linear absorption of laser light by a monomer for cross-linking typically limits the resolution to the wavelength of the laser that is used. Moreover, deep penetration of the photons into the resin bulk is impossible because of the strong absorption of the light, so that the process is carried out necessarily thin layer by thin layer (see Figure 8.30). Using two-photon excitation changes all this as it enhances the resolution and can penetrate deep into the bulk of the resin. When high-intensity light shines on a material, the probability for two-photon absorption is proportional to the square of the field intensity (see Volume III, Chapter 6 for an illustration of this effect), and thus is greatest at the center of a Gaussian laser spot. Using a high NA lens, the laser spot can be made less than 1 μm in diameter, and significantly narrower structures can be obtained because polymerization starts as soon as a certain threshold of intensity has been reached.

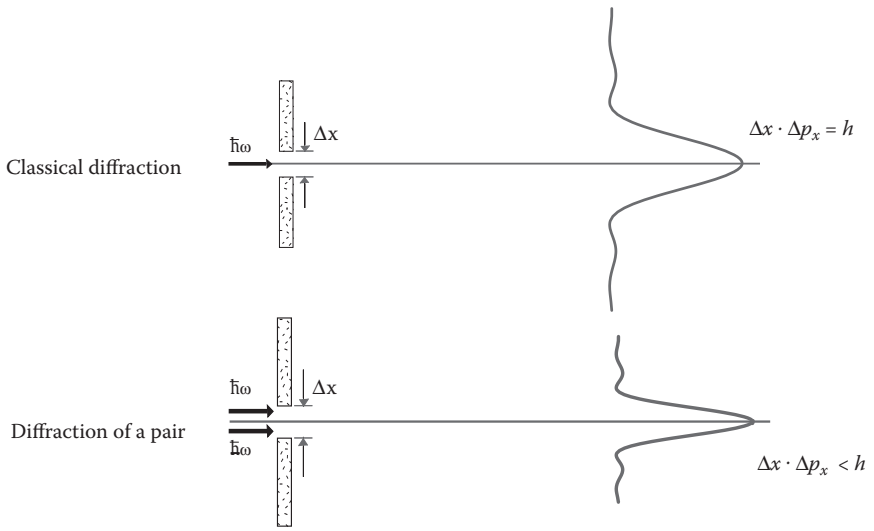


FIGURE 2.36 The entangled photon pair comes out from a point of the object plane, undergoes two-photon diffraction, and results in twice narrower point spread function on the image plane.

Therefore, by tightly focusing a femtosecond laser beam into a resin, subsequent photo-induced reactions such as polymerization occur only in the close vicinity of the focal point, allowing the fabrication of subwavelength lithography and 3D polymer structures. Importantly, because the energy required to initiate polymerization is twice that of the incoming radiation, it is possible to excite deep into the monomer as the material is transparent to one-photon excitation, whereas at the focus, where the intensity is high, two-photon excitation occurs.

Parthenopoulos and Rentzepis[75] first demonstrated the 3D microfabrication capability possible through two-photon polymerization in optical memory in 1989, and Wu et al.[76] pioneered the technique's utility for lithography in 1992. In two-photon lithography one often works with epoxy resins and urethane acrylates. Several two-photon photoinitiators with very large two-photon absorption cross-sections to be added to base polymers have been developed.[77] In these photoinitiator molecules the photon energy that is absorbed is emitted as photons of a higher energy (up-converted fluorescence). The up-converted fluorescence is then used to induce various photochemical reactions such as photopolymerization and photodissociation of the base resin. Two-photon photolithography experiments using this idea recently have achieved 120-nm spatial resolution using femtosecond laser pulses with a 780-nm central wavelength.[78] Below we will see that using two-photon lithography combined with apertureless near-field scanning optical microscopy (ANSOM), a resolution nearly a factor of 2 better than the resolution achieved in far-field two-photon lithography experiments can be achieved. We will also learn that combining a metallic ANSOM tip and femtosecond laser pulses, lithographic patterning with resolution of 70 nm in a SU-8 photoresist can be obtained.

As we clarified above, energy transferred to the photosensitive resin material by two photons of wavelength λ corresponds to that transferred by a single photon of wavelength $\lambda/2$. At the same time one takes advantage of long-wavelength exciting light with larger penetration depth, which enables 3D patterning of bulk material. Scanning the focus within the material allows one to produce microscale and submicroscale structures. Subwavelength structures can be produced because the absorption profile for the two-photon process is narrower than the beam profile. Some examples of 3D two-photon lithography are shown in Figure 2.37. The two-photon 3D lithography has also been used to make photonic crystals. In Chapter 8 on photochemical forming, we will come back to this two-photon 3D lithography process and its use in stereolithography or microphotoforming.

Quantum lithography is in its infancy, and many experimental challenges remain. The main problem is how to generate *N*-state entangled photons

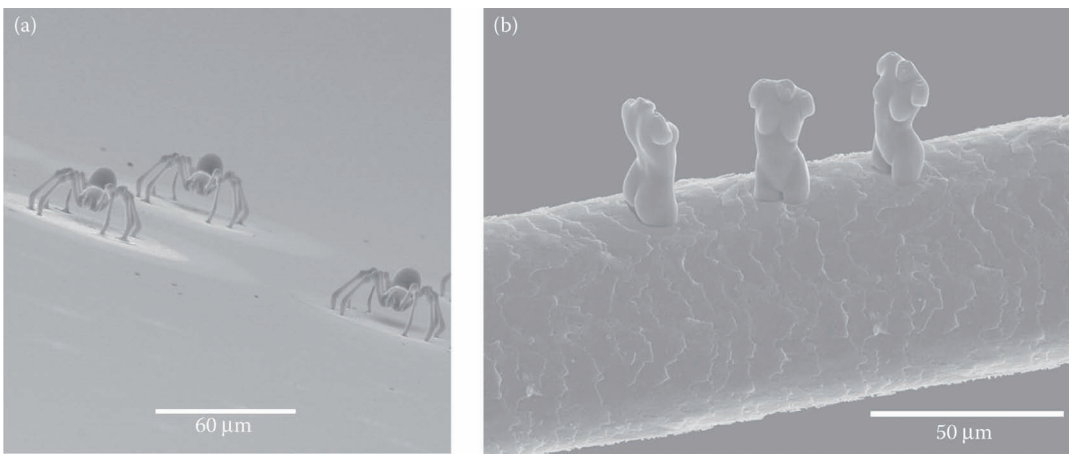


FIGURE 2.37 Two-photon 3D lithography. (a) Microspider array, the body of the spider is above the substrate; it is supported by eight 1-µm thick legs (http://www.laser-zentrum-hannover.de). (b) Venus micromodels on a human hair (http://www.laser-zentrum-hannover.de).

in a reasonable way. The next problem is how to develop an *N*-photon-sensitive photoresist. Third, the exposed field is very small. However, if we could overcome all three problems, we can beat the Rayleigh limit and approach the Heisenberg limit instead, thus approaching a resolution on the order of the atomic size.

Plasmon Lithography

Atwater and coworkers from Caltech introduced plasmon lithography in 2002. Plasmon lithography is based on plasmon resonance occurring in nano-sized metallic structures, allowing for the replication of patterns with a resolution limit considerably below the diffraction limit.[77] As we discovered in Volume I, Chapter 5 on photonics, when metallic nanoparticles are excited in an electromagnetic field, they can exhibit collective electron oscillations known as surface plasmon oscillations. With the diameter of the particle much smaller than the wavelength of the incoming light, the oscillating charges produce an oscillating dipole field around the metal particles. At resonance, i.e., when the excitation frequency coincides with the natural relaxation frequency of the nanoparticles (ω_p), this results in a strongly enhanced local electrical field. One can choose the nanoparticles such that their resonance frequency falls within the sensitivity range of an underlying photoresist. The field enhancement at the interface of the particle and the resist is then used to locally expose the resist, as schematically illustrated in Figure 2.38. A transparent mask with a configuration of metal nanoparticles deposited on it is brought into intimate contact with a thin layer of photoresist. Kik et al.[79] used spray deposition to obtain ~41-nm-diameter silver nanoparticles. To ensure that the local field intensity is enhanced directly below the metal particles, the illumination beam should be at a glancing angle to the resist (p-polarized light). Although the whole resist film is illuminated, development of the resist only affects the areas that received locally enhanced exposure. In this way an image of the mask pattern is transferred onto the resist layer. The choice of nanoparticle material and size depends on the required wavelength; as stated earlier,[79] the resonance wavelength should fall within the resist sensitivity window. For example, to be able to use standard g-line photoresist, the resonance wavelength should lie between 300 nm and 460 nm. Additionally, to obtain high-field enhancement, long electron scattering times (τ) are desirable. Silver is a good candidate, with long reported electron scattering times up to 10 fs and a resonance occurring at a wavelength of ~360 nm in air. The particles should also be small enough (<40 m) to attain confined dipolar enhanced fields at the resist interface. Particles that are too large develop multipolar oscillations as we detailed in Volume I, Chapter 5.

From Volume I, Chapter 5 we know that the resonance frequency of the metal nanoparticles is not a fixed quantity. For a spherical particle, for example, the surface plasmon resonance occurs at the wavelength λ for which ε' (dielectric constant of the particle) = $-2\varepsilon_d$ (dielectric constant of the medium). Consequently, the wavelength at which resonance occurs can also be tuned by changing the refractive index of the surrounding medium. Using this method Kik et al.[79] have shown that the local field enhancement occurring around metal nanoparticles when they are excited at the surface plasmon resonance frequency can be used to print nanoscale features in thin resist layers. Feature sizes less than $\lambda/10$ were generated in a parallel fashion using visible illumination and standard g-line photoresist. Further below we analyze ANSOM

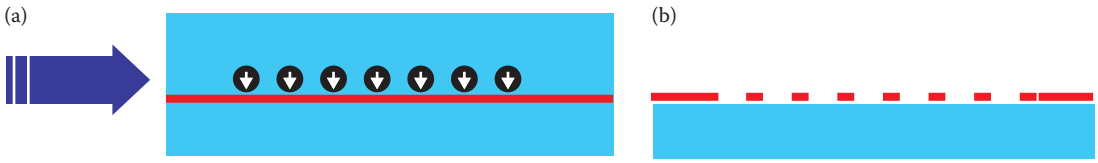


FIGURE 2.38 Schematic representation of plasmon printing showing glancing angle illumination using polarized visible light, producing enhanced resist exposure directly below the metal nanostructures in the mask layer (a), and the resulting pattern in the resist layer after development (b).

lithography, where a scanning metal tip is used to generate plasmons for exposing the underlying resist.

In Volume I, Chapter 5, we discussed how Pendry had the amazing insight that a superlens could be made from a flat slab of a negative refractive index material, which not only would bring incident rays to a focus but also has the capacity to amplify the near-field radiation so that it could contribute to image formation, thus removing the wavelength limitation completely. Thus, a superlens would prevent image degradation and beat the diffraction limit established by Abbe! We also discussed the development of a poor man's near-field superlens ($\varepsilon < 0$ and $\mu = 1$) in 2003 by Zhang's group at the University of California, Berkeley, who showed that optical evanescent waves can be enhanced as they pass through a silver superlens.[80] This team imaged objects as small as 40 nm across with their superlens, which itself is just 35-nm thick.[81] In contrast, current optical microscopes can only resolve objects down to around 400 nm. With the superlens, using 365-nm illumination, features of a few tens of 10 nm were imaged, clearly breaking Abbe's diffraction limit.

Scanning Proximal Probe Lithography

Introduction

Proximal probe techniques rely on the use of nanoscale probes, positioned and scanned in the immediate vicinity of a material surface. Their development is often viewed as a first step toward nanotechnology because they demonstrate the feasibility of building purposeful structures one atom or one (macro) molecule at a time. Proximal probes used in lithography might involve electrical methods where a scanning tunneling microscope (STM) tip generates a local field/current that modifies the region directly under the tip (e.g., SiH $\rightarrow$ Si). A second approach involves mechanical methods where a scanning force microscope (SFM/AFM) tip scrapes, thermally deforms, or transfers material at the surface; the latter material transfer method corresponds to dip-pen lithography (DPL). Third, it may involve a near-field scanning optical microscope (NSOM) tip or apertureless near-field scanning optical microscopy (ANSOM) that exposes photoresist under the tip only.

The use of single proximal probe tips for lithography poses a serious drawback in terms of processing speed. To use these techniques in the manufacture of ICs and data storage devices, it is necessary to devise a scheme for parallel processing by making arrays of these proximal probes.[82]

STM/AFM Background

The scanning tunneling microscope (STM) images the surface of conducting materials with atomic-scale detail (see Volume III, Chapter Six). STM was invented by Gerd Binnig and Heinrich Rohrer of IBM's Zurich Laboratory in 1981 (they received the 1986 Nobel Prize for their invention).[83] In general, STM works by bringing a small conducting probe tip up to a conducting surface. When the probe is very close to the surface (<10 Å) and operating voltages in the ±10-V range are applied, very small currents are produced because the electrons in the probe and the surface have wave functions extending beyond the physical surface boundaries. To the extent that these spillover wave functions overlap, a measurable current results. The interesting part about this current is that it depends exponentially on the spacing between the two conductors (as well as the voltage) (see Volume I, Chapter 3 in the section on tunneling). A piezoelectric transducer accomplishes the *z*-axis distance variation between tip and sample. By changing the distance over 1 Å, the current changes by a factor of 10. In practice, the current is kept constant through a feedback mechanism, and the probe moves up and down over the surface following the atomic contours it "sees." The images produced by the STM come from the electronic structure and from the geometry of the sample. Up to 100 times more powerful than SEMs, STMs measure objects in the angstrom range.

Over time, many proximal probe were developed, and STM belongs now to a large new family of very local, proximal probes, such as atomic force microscopes (AFMs), scanning electrochemical microscopes (SECMS), scanning thermal microscopes, scanning capacitance microscopes, magnetic force microscopes, and scanning pH probes, enabling microscopy of almost any type of material and

property (for more details see Volume III, Chapter 6 on metrology). The common feature of these instruments is that their resolution is not determined by visible light used for the interaction with the probed object, as in conventional microscopy.[84]

Besides their use for imaging and characterization of surfaces, proximal probes can also be used to modify solid surfaces. Below we introduce some examples of this type of proximal lithography.

Scanning Probe Lithography

Introduction In lithography, STM can produce nanostructures at higher resolution but at lower speed, whereas AFM can operate at higher scanning speeds but offers lower resolution. AFM can be used to image and pattern any kind of material, making AFM the more widely used technique. AFM/STM lithography can be classified as force assisted (including thermomechanical), bias assisted, and material transfer assisted.

Scratch Lithography—Mechanical Effect The simplest (and perhaps the crudest method) to pattern a surface with a proximal probe is by scratching in scratch lithography (see inset in the heading of this chapter: copy of Pablo Picasso's "Don Quixote" in polycarbonate patterned with scratch lithography). In this mechanical process, simply drawing an AFM tip across a surface with enough force may cut a trench, or the piezomechanism driving the AFM cantilever may be modulated in the *z*-direction to tap the tip on the surface as it is drawn over it. Cuts as narrow as 20 nm and 2 nm deep have been made in III–V semiconductor surfaces.[85] One difficulty with the scratching approach has been that probe tips often break following direct collision with a solid surface. As a possible solution, carbon nanotubes have been mounted on the tip of silicon cantilevers, enabling lines as narrow as 10 nm (Figure 2.39). Carbon nanotubes constitute the strongest material known along the axial direction and yet are highly elastic and flexible along the radial direction (see also Volume III, Chapter 3).[86]

The Naval Research Laboratory (NRL) worked with proximal probes to pattern thin films of chemically amplified negative e-beam resist (SAL601 from Shipley). Resist films of 30–70-nm thick were

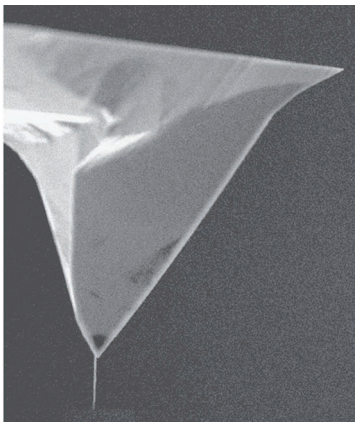


FIGURE 2.39 Nanotube mounted on the micromachined tip of a Si cantilever is used as a nanopencil for lithography of 20–10-nm lines (courtesy of Dr. M. Meyyappan, NASA Ames). SEM picture of nanotube mounted on tip of Si cantilever (http://cnst.rice.edu/pics.html).

patterned with typical tip-sample voltages from –15 to –35 V, resulting in minimum feature sizes of 23 nm.[54] Using SAMs as resists (see above), it was shown that the lower the exposure threshold energy of the film, the better the lithographic resolution. There are indications that, with these low voltages (~4 V), SAM resists will eventually yield sub-10-nm CDs.

In a variation on scratch lithography, a resistively heated AFM probe writes a data bit by scanning over a polymer surface. The combined heat and mechanical force of the tip causes the polymer to soften and flow, thus facilitating the writing of data bits in a storage medium (see the millipede chip developed by IBM for thermal writing in a polymer for memory; Figure 2.4b).

Electrical Field-Induced Chemical Modification—Electrochemical Effect In an alternative approach, surface oxides were induced on Si and GaAs by slightly increased tip voltages on samples held in a wet nitrogen atmosphere.[38] In this method, the water meniscus formed at the tip-sample gap is dissociated by the negative tip bias, and the O^- and OH^- oxidative ions react with the substrate to form localized oxide nanostructures. Because the molecular volume of the oxides is usually larger than that of the substrate materials, raised nanopatterns are formed after the oxidation reaction. These thin oxides, although only a few monolayers thick, are sufficiently robust to act as a mask for subsequent reactive ion etching of

the substrate.[54] The oxidation process is fairly general and may also be applied to Ti and Cr. At NRL, oxide features with lateral dimensions as small as 10 nm have been achieved.[54] Multilayer resist films for nanopatterning with SPL have also been developed. Sugimura et al.[87] work with a three-layer resist to pattern insulating substrates such as thermally grown SiO_2 with current injection from a scanning probe tip. The process sequence is sketched in Figure 2.40.

The bottom layer of the three-layer resist consists of 20 nm of amorphous Si (a-Si) and is prepared by ion-beam sputtering; a second layer consists of an intermediate 2-nm-thick Si oxide and is prepared by photo-oxidation of the top layer of the a-Si. The top resist layer is 2-nm thick and consists of an octadecylsilyl self-assembled monolayer (ODS-SAM). To pattern the resist on top of the insulating SiO_2, a bias voltage is applied between the AFM probe and the conductive a-Si layer, which is made positive. When a scanning probe is operated in the presence of atmospheric water vapor, the probe and sample are automatically connected via a tiny water column created by capillary forces of the adsorbed water. This assembly serves as a minute electrochemical cell. As a result of the electrochemical anodization reaction, the monolayer resist becomes degraded in the region where the probe has passed, and the underlying photo oxide grows thicker. A subsequent etch in 0.5 wt% HF removes the exposed oxide, and that pattern is then transferred to the underlying a-Si by an etch of the a-Si in an aqueous solution of 25 wt% tetramethylammonium hydroxide (TMAH). The latter solution etches Si but not SiO_2. The a-Si is etched, and the etch stops when the underlying insulator is reached. The underlying thermal oxide can now be etched in HF and the resist stripped. Minimum feature sizes reached so far are about 50 nm.

Electron Scanning Probe Lithography-Bias Effect

Lithography using electrons from a scanning microprobe offers several potential advantages over writing with a traditional e-beam source.[84,88] One important benefit is that the low energy of electrons in SPL (<50 eV) compared with those in EBL (10–100 keV) avoids the detrimental effects of electron backscattering, thereby virtually eliminating proximity effects and thus producing enhanced resolution and superior pattern fidelity.[54] Moreover, SPL using an STM can be operated in air, whereas EBL relies on a vacuum. Because of the small tip-to-sample distance,

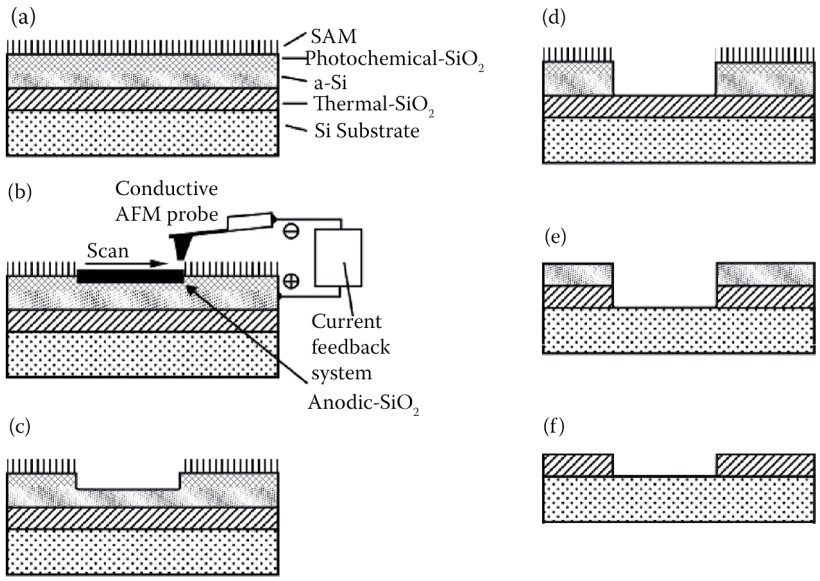


FIGURE 2.40 Schematic of AFM lithography on SiO_2 using a multilayered resist system. (a) Cross-section. (b) Exposure by drawing patterns into the ODS-SAM layer by current injection from an AFM probe. (c) First development step by HF etching to remove the SiO_2 just formed. (d) Second development step by TMAH etching to remove the a-Si in the exposed area. (e) Pattern transfer with HF to remove the thermal oxide in the exposed area. (f) Resist removal, i.e., removal of all the remaining a-Si in TMAH etch. (Based on Sugimura, H., O. Takai, and N. Nakagiri. 1999. Multilayer resist films applicable to nanopatterning of insulating substrates based on current-injecting scanning probe lithography. *J Vac Sci Technol* B17:1605–08.[87])

extremely small spot sizes are achievable so that the exposure dose may be confined to a beam diameter of less than 10 nm. The method enables wider exposure latitude than EBL, a fact demonstrated in Figure 2.41.[89] In this figure, line width is plotted versus dose for SAL601 resist using both EBL and SPL lithography systems. From the lower slope of the SPL curve, one deduces that SPL has higher dose latitude than EBL—in other words, SPL is less sensitive to dose variations. On the other hand, SPL is less sensitive and requires a higher dose to write the same feature size. The mechanism of electron bombardment in SPL is very different from that in EBL. In EBL, the mean free path of the bombarding electrons is long compared with the resist thickness. In contrast, low-energy electrons in SPL have a mean free path less than 2 nm and travel through the resist under the influence of an electrical field, undergoing a number of scattering events before reaching the resist/substrate interface.[89] The latter makes it clear why a thinner resist will result in a better resolution. The influence of secondary electrons (i.e., all electrons emanating from the resist/substrate) in proximal probe lithography was investigated, for example, by Völkel et al.[90]

Writing speeds of an SPL tool of 0.1–100 µm/s are typical, and several thousands of micrometers per second have been demonstrated.[91] However, very fast scanning will remain limited because the close tip-to-substrate distance (typically 1 nm with STM) is prone to cause tip crashes.

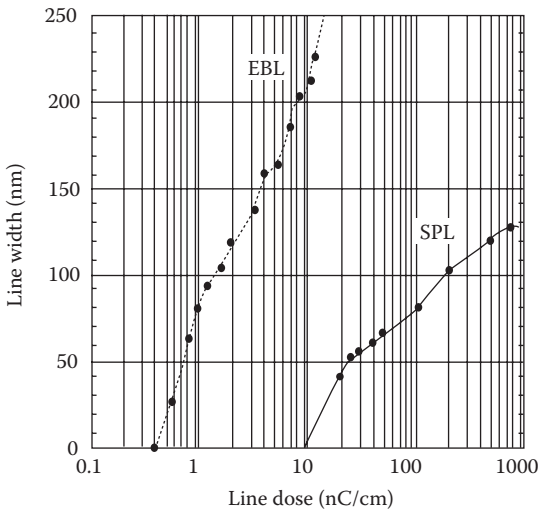


FIGURE 2.41 Line width versus dose for SAL601 resist using both EBL and SPL lithography systems. (Wilder, K., C.F. Quate, B. Singh, and D.F. Kyser. 1998. Electron beam and scanning probe lithography: a comparison. *J Vac Sci Technol* B16:3864–73.[89])

With an ultrasharp tip (radius < 10 nm) in the emission mode rather than the tunneling mode, one may stay as far as 100 nm away from the surface. This approach not only enables faster writing (limited by the response of the piezoelectric material only) without fear of crashing the tip as frequently, but the high beam current also contributes to the possibility for exposure of large area patterns. Even at a distance of ~100 nm, the beam diameter remains small at ~30 nm.[55] Structures of lateral dimension of ~20 nm have been created in SAMs using ultrasharp field emitters.[92] The use of arrays of micromachined electron emission sources to improve throughput was covered earlier (see Figures 2.22 and 2.23) as a type of lower energy EBL with less lateral backscattering. In this context a useful array of field emitters for EBL is the so-called Spindt cathode array[93] shown in Figure 2.42. The use of such an array in a field emission display is shown in Chapter 7 (Figure 7.90a). The cathode tips in Figure 2.42 are molybdenum, and electrons are extracted from the tip at gate voltages as low as 20 V (70–100 V is more typical). In the newest embodiments of Spindt arrays, the molybdenum tip is replaced with carbon nanotubes. The use of nanotubes as field emitters was first proposed by Walt de Heer, André Chatelain, and Daniel Urgate in 1995.[94] Carbon nanotubes are excellent field emitters because of their chemical, structural, and electronic properties, which afford important aspects of robustness that have been lacking in the conventional metal and silicon field emitter arrays.

Material Transfer Methods: Atom Lithography or Mechanosynthesis and Dip Pen Nanolithography

Atom Lithography or Mechanosynthesis As we just learned, proximal probe equipment locally modifies surfaces[95] and is used to expose resists and to oxidize and mechanically scribe surfaces. Also, a variety of direct atomic manipulations have been demonstrated with proximal probes.[95] The electric field strength in the vicinity of a probe tip is very strong and inhomogeneous (e.g., a field of 2 V Å^{-1} concentrated around the probe tip). This field can manipulate atoms, including sliding of atoms over

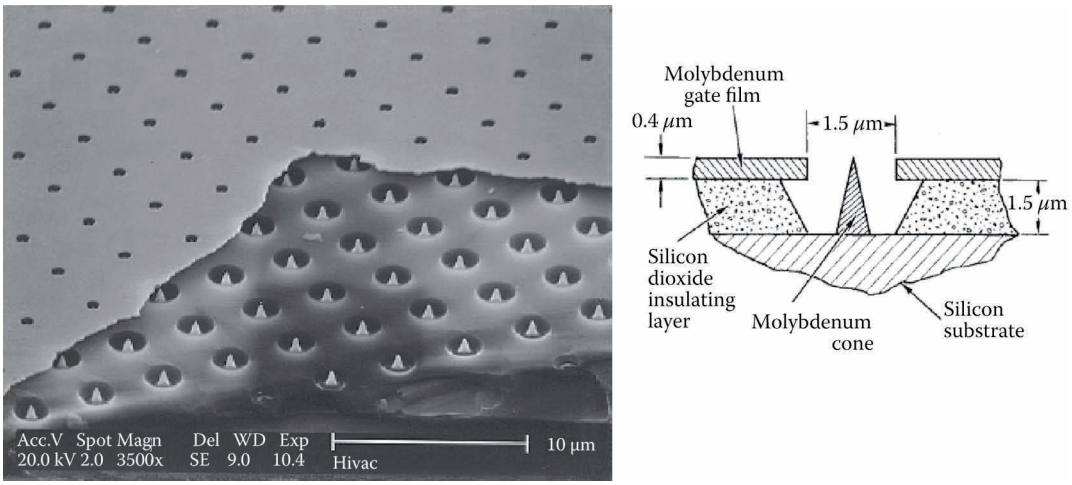


FIGURE 2.42 Spindt field emitter array. For its fabrication see Chapter 7, and for its application in a field emission display see Volume III, Chapter 10.

surfaces and transferring atoms by pick (erase) and place (write). Drexel calls this type of machining mechanosynthesis.[96] Although both positive and negative voltages have been used for the deposition, a tip with a negative polarity is usually more stable with higher transfer probability. These atomic manipulation processes may be classified in parallel processes and perpendicular processes. In parallel processes, an adsorbed atom or molecule is induced to move along the surface (sliding); in perpendicular processes, the atom or molecule is transferred from the surface to the tip of a proximal probe or vice versa. Patterning surfaces at the atomic level is demonstrated in Figure 2.43a where a 60 × 48-Å image of four Pt adatoms is shown. The four atoms were herded into a linear array.[95] Figure 2.43b shows seven Pt atoms compacted together. The theoretical resolution of a lithography technique based on these atomic probes is a single atom. Even with "only" a 100-Å resolution, a single memory bit could be stored in an area that measures 100 Å on a side. This will enable bit storage of 10^{12} bits/cm^2 as compared with 10^9 bits/cm^2 with conventional technology.

In Figure 2.44 we show linear Au chains 2–20 atoms long, written by Ho et al.[97] from the University of California at Irvine on a NiAl (110) surface via the manipulation of single atoms with an STM. Differential conductance (dI/dV) images of these gold chains reveal 1D electronic density oscillations at energies 1.0–2.5 eV above the Fermi energy. Thus, Ho et al. were able to make a direct correlation between the geometric structure and the electronic properties of the Au chains, where the former is precisely known and the latter is directly measured. The same team, again using STM, also connected Au chains to a copper phthalocyanine (CuPc) molecule.[98] A typical artificial molecule assembly is CuPc@2Au$_3$, consisting of a CuPc molecule and two Au$_3$ (three Au atoms in each chain) chains. When the separation between the Au chains is six lattice constants [of the underlying NiAl (100) substrate]

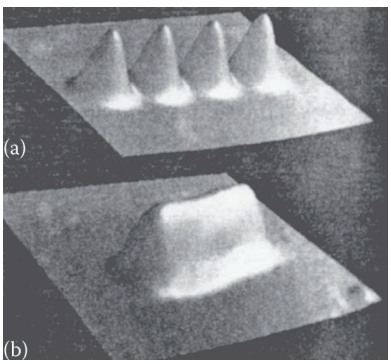


FIGURE 2.43 Atom manipulation for nanomatching. (a) A 60 × 48 Å STM image of four Pt adatoms assembled into a linear array on a Pt (111) surface. Pt atoms were herded four unit cells apart along a close-packed direction of the Pt (111) surface. (b) A 40 × 40 Å STM image of a compact array of seven Pt adatoms. (From Stroscio, J.A., and D.M. Eigler. 1991. Atomic and molecular manipulation with the scanning tunneling microscope. *Science* 254:1319–26. With permission.[95]) For more pictures of STM atom manipulation, visit http://www. almaden.ibm. com:80/vis/stm/gallery.html.

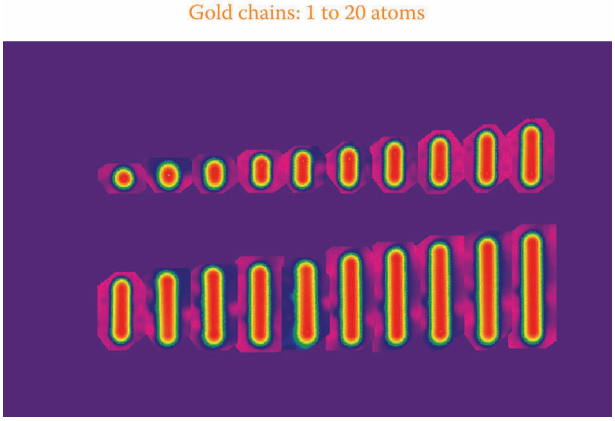


FIGURE 2.44 STM topographic images from a single Au atom to a linear chain of 20 atoms, arranged in single-atom increments from left to right. The images are cut from 20 separate scans, each taken with a sample bias voltage between 2.0 and 2.5 V and a tunneling current between 1.0 and 1.5 nA. Each chain has an apparent height between 2.4 and 2.7 Å. The chains between 12 and 20 atoms long were constructed and imaged with a tip different than the tip used to construct and image the other chains. (Wallis, T.M., N. Nilius, and W. Ho. Electronic density oscillations in gold atomic chains assembled atom by atom. *Phys Rev Lett* 89:236802-1–2-4.[97])

or more, a Pc molecule is unable to bridge the gap between the chains. With five lattice constants, the gap is small enough for a Pc molecule to fill it, and the molecule adsorbs symmetrically with two opposite benzene rings attached to the Au chains (see the schematic in Figure 2.45j).

This work by Ho's team is contributing to the detailed understanding of the electron transport through a molecule that bridges the gap between two metal electrodes. Electron transport through single-molecule junctions is determined by the local electronic structure of the nanoscale region that includes the molecule and a number of metal atoms in proximity to the molecule-metal contacts. As we will detail in Volume III, Chapter 3, a detailed understanding of the contact between metal-contacting electrodes and molecules is essential for the understanding of molecular electronics.

One major drawback to bear in mind with atom placing or removing techniques for micromachining is the time involved in generating even the simplest of features. If one wanted to deposit a metal line 10 µm long, 1 µm wide, and 0.5 µm high, 10^{16} atoms would need to be manipulated. Even at a deposition rate of 10^9 atoms/s, this would take more than 100 days. In the late 1980s, it took an IBM team 1 week to spell out the IBM logo with an STM, putting Xe atoms in place on a Ni surface. Only use of massive parallelism can help this cause.

Dip Pen Lithography AFMs suffer from an annoying problem: water from the air tends to attach itself to the tip, and the greater the relative humidity, the more water that collects. This water meniscus interferes with high-resolution AFM imaging. In 1999, Chad Mirkin, a Northwestern University researcher, realized that this meniscus could be turned into an asset. Because water is always moving from the tip to the substrate surface or vice versa, he decided to use it to float ink molecules onto the substrate. The principle of dip pen nanolithography (DPN) is shown in Figure 2.46.[99] DPN is named after the 4000-year-old dip pen, with an AFM tip for a sharp point.

A reservoir of "ink" is stored on the cantilever holding the scanning probe tip, which is manipulated across the surface, leaving lines and patterns behind. Lines as thin as 15 nm have been drawn. The attainable resolution depends strongly on the substrate roughness, the writing speed, and the relative humidity. Smooth surfaces, fast scan rates (and thus shorter pen/paper contact), and lower relative humidity (narrower water meniscus) all make for higher-resolution DPN. Thus far, most of the work has been carried out with a sulfur-containing compound called 1-octadecanethiol on gold surfaces. However, many inks and many different types of surfaces can be written on, making this an interesting nanolithography option. A review of different ink/substrate combinations in DPN was listed by Ginger et al.[100] An additional attractive feature of DPN lies in the fact that the same AFM can be used to write and image the results. The downside is that DPN is slow, but to offset this negative aspect, arrays of dip pens, similar to the millipede chip developed by IBM (see Figure 2.4b), can parallelize and speed up the technique: the newest array from Chad Mirkin's group has 55,000 pens.[101] Much needed now is the development of an automated ink-coating system for inking individual tips with its own ink; perhaps such a system

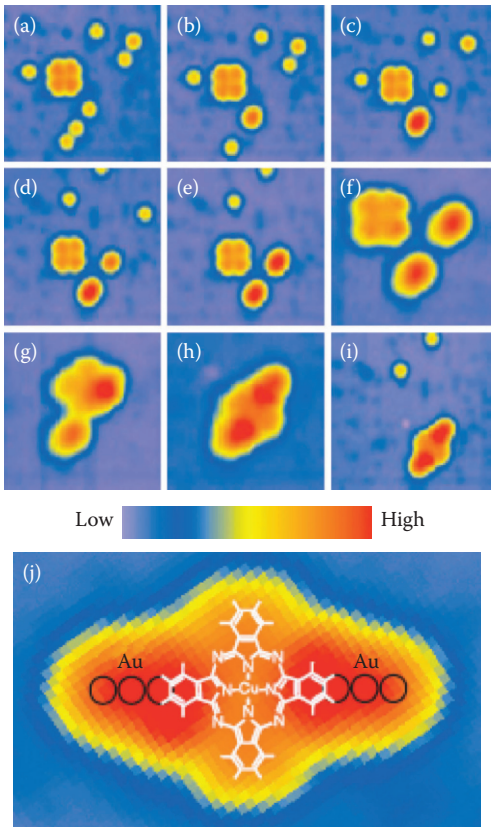


FIGURE 2.45 Assembly sequence and schematic for a structure consisting of a CuPc molecule and two Au_3 chains ($2Au_3$). All images were taken with V_{bias} = 1 V, I = 1 nA. (a) CuPc molecule (appears as a protruding four-lobe structure) on the NiAl surface together with seven Au atoms, which appear as round protrusions. Four of the atoms have been aligned along the same Ni trough. Image size is 74 Å by 74 Å. (b) One of the atoms has been manipulated along the Ni trough to create a Au_2 chain. (c) First Au_3 is created; another atom is positioned five Ni-Ni lattice sites (5 × 2.89 Å) away from Au_3, marking the starting position for another Au_3 chain. (d) A second Au_2 chain is created. (e) Assembly of the second Au_3 chain and the $2Au_3$ junction is completed. (f) Zoom-in image of the $2Au_3$ and CuPc molecule; the area of image is 47 Å × 47 Å. The image was taken with a tip modified by a CuPc molecule adsorbed on it. (g) The CuPc molecule has been moved into the junction between the two Au_3 chains. As is clear from the image, the molecule is adsorbed only on one of the chains. (h) The position of the molecule was adjusted by manipulating the leftmost lobe of the molecule in (g) toward the lower Au_3 chain. The molecule is adsorbed symmetrically between the two chains, forming CuPc@$2Au_3$. (The tip still has a CuPc molecule adsorbed on the apex.) (i) Image of same structure taken after removing the molecule adsorbed on the tip. The area corresponds to that of (e). (j) Schematic showing the relation between the internal molecular structure, adsorption geometry, and STM image. Black circles represent the Au atoms comprising the two Au_3 chains. (From Nazin, G.V., X.H. Qiu, and W. Ho. 2003. Visualization and spectroscopy of a metal-molecule-metal bridge. *Science* 302:77–81.[98])

can be achieved through the integration of microfluidic technology. Besides its use in nanolithography one can envision the use of DPN for placing of molecules on a surface precisely where one wants them as in protein and DNA arrays and to build 3D nanostructures. Mirkin et al.[102] have shown that through the use of DPN, protein nanoarrays can be constructed with high fidelity of specific protein binding to proper sites on a Si substrate. NanoInk (http://www.nanoink.net) is a start-up company commercializing the DPN technology.

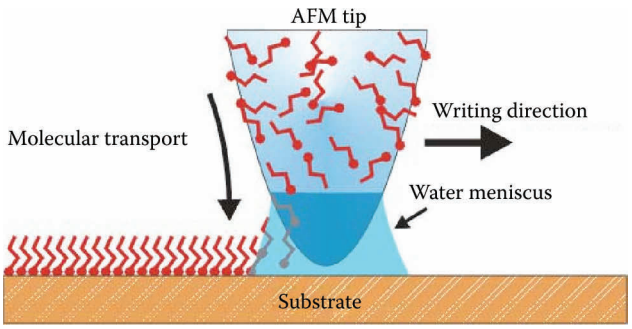


FIGURE 2.46 Dip pen nanolithography: transport of molecules to the surface via water meniscus (http://chem-groups.northwestern.edu/mirkingroup/dpn.htm).

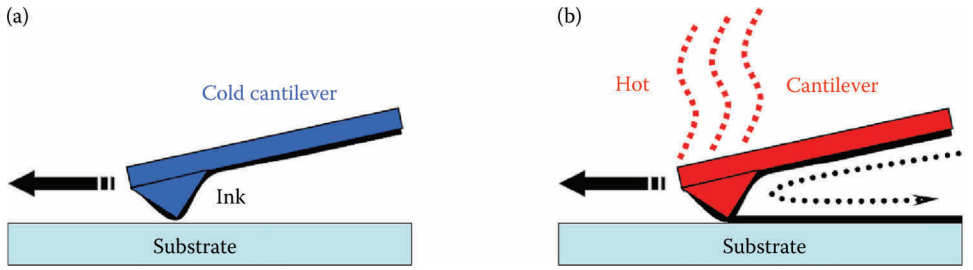


FIGURE 2.47 Diagram illustrating thermal dip pen nanolithography. When the cantilever is cold (a), no ink is deposited. When the cantilever is heated (b), the ink melts and is deposited onto the surface. (Image courtesy of Naval Research Laboratory, http://gtresearchnews.gatech.edu/newsrelease/tdpn.htm.)

In Figure 2.47, we illustrate an extension of DPN, a writing method dubbed "thermal dip pen nanolithography." In conventional DPN, the probe tip is coated with a liquid ink, which then flows onto the surface to make patterns wherever the tip makes contact. This approach has been hampered by an inability to turn the ink flow on and off. A thermal-DPN (tDPN) method developed by Georgia Tech's William King and NRL's Lloyd Whitman solves this specific problem by using easily melted solid inks and special AFM probes with built-in heaters that allow writing to be turned on and off at will.[103,104] Moreover, conventional DPN cannot be used in a vacuum because liquid inks evaporate, whereas the solid materials used in the thermal process bond to surfaces, allowing them to be used in vacuum environments that are part of conventional semiconductor manufacturing. The thermal materials also provide sharper features because they do not spread out like liquid inks.

In Figure 2.49 a topographic image of a surface scanned with a heated AFM cantilever tip for 256 sec in each of four 500-nm squares is shown. The cantilever temperature is shown for each of the four scans. No writing is observed from the two low-temperature scans. The scan at 98°C resulted in light deposition. Robust deposition occurred during the final scan when the cantilever temperature was increased to 122°C.

Combining thousands of individually controlled AFM pens into arrays could allow writing of faster and more complex semiconductor patterns. One of the first inks used in thermal writing was octadecylphosphonic acid, which melts at about 100°C. But the tDPN technique has been used also to apply other materials, including solders and polymers.

Near-Field Optical Scanning Microscope and Apertureless Near-Field Scanning Optical Microscopy

Introduction: Breaking the Diffraction Limit To obtain a resolution better than Abbe's optical microscopy limit ($\Delta x \sim \frac{\lambda}{2NA} \sim 200$ nm, based on Volume I, Equation 5.1), complicated and costly electron microscopy (with resolution of a few nanometers) or STMs (with atomic resolution) are required. Unfortunately, these techniques sacrifice many of the advantages associated with traditional optical microscopes (including nondestructiveness, low cost, high speed, reliability, versatility, accessibility, ease of use, informative contrast, spectroscopy, and

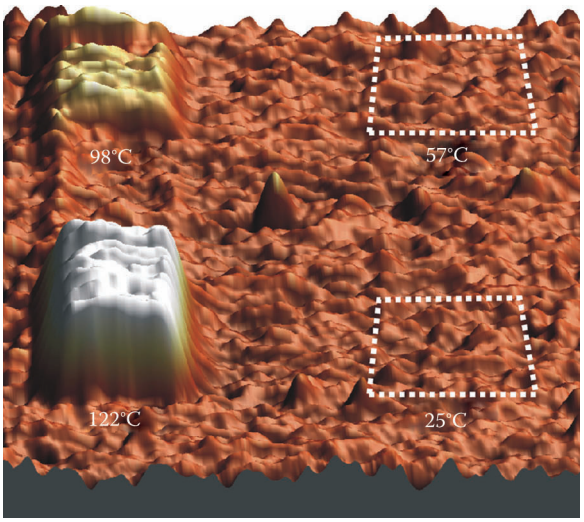


FIGURE 2.48 A topographic image of a surface scanned with a heated AFM cantilever tip for 256 s in each of four 500-nm squares. The cantilever temperature is shown for each of the four scans. No deposited material is observed from the two low-temperature scans. The scan at 98°C resulted in light deposition. Robust deposition occurred during the final scan when the cantilever temperature was 122°C. (Courtesy of Naval Research Laboratory.)

real time). As seen in Volume I, Chapter 5 on photonics, combining a scanning proximal probe technique with optical microscopy in so-called scanning near-field optical microscopy, NSOM provides an attractive solution to this dilemma. In NSOM, the sample is illuminated by a nanoscopic light source located close to the surface (10 nm), and the resolution is dictated by the source diameter *a*, or $\Delta x \sim a \leq 100$ nm. This is achieved by using nanoscale apertures in NSOM or by using apertureless techniques (e.g., a scattering metal tip) in ANSOM or scattering scanning near-field optical microscopy. A typical NSOM probe is a commercial single-mode optical fiber, tapered and coated by a thin CrAl film with an aperture of ~100 nm. In ANSOM, light interaction over nanoscale dimensions is enhanced with the use of scattering nanoscale tips, nanospheres, and so on. The conceptual idea of near-field optics and near-field microscopy came from the seminal proposal of E. H. Synge. It was Synge (in 1928) who proposed that if source or detector is scanned close to a surface, resolution is dependent on probe size and probe-sample distance instead of the Abbe limit.[105] Synge corresponded about his idea with Einstein, who accepted the concept but was doubtful that it could be reduced to practice. But Einstein was wrong; today NSOM is an important tool for nanotechnologists. NSOM constitutes a form of lensless optics with subwavelength resolution, independent of the wavelength of the light being used! The light coming from an imaged object in the near field contains a considerable amount of nonpropagating, evanescent waves containing high lateral spatial resolution, information that decays exponentially within a distance comparable to a wavelength. It is this evanescent wave information, the information-rich light, we capture in NSOM and ANSOM, thus enhancing the resolution. Several NSOM implementations using apertures and SNSOM were illustrated in Volume I, Chapter 5.

Aperture-based probes are the best developed and come with a typical resolution of 50–100 nm. Tip fabrication involves heating and pulling and chemical etching of glass optical fibers. Apertureless probes enhance local scattering and feature a 1–20-nm resolution. NSOM and SNSOM microscopy are reviewed in Volume III, Chapter 6.

NSOM and ANSOM Lithography Besides their utility for imaging applications, NSOM and ANSOM may also be used for direct write optical nanolithography.[106] From the above, near-field scanning optical microscopy refers to the interaction of light with a sample close to a metal aperture (NSOM) or a solid metal tip (ANSOM). The light source is held in place in a manner similar to those used for other scanning proximal probe microscopes. To write patterns into a photoresist, one uses aperture NSOM in illumination mode or ANSOM in plasmon mode. The first NSOM lithography results obtained in several labs dealt with nanometric tips on standard positive photoresist and on self-assembled monolayers of alkanethiols and using laser sources visible up to the UV range. The NSOM approach to improved spatial resolution in lithography unfortunately is limited because of the very small apertures, through which only a very small fraction of the light can be transmitted. ANSOM overcomes this problem by using sharp metallic tips instead of fiber apertures to achieve nanometer-scale resolution. In ANSOM, light is polarized along the sharp axis of the metallic tip, and this induces a high concentration of surface plasmons that results in strong enhancement of the electromagnetic field in the local vicinity of the tip—typically more than a few tens of nanometers (see plasmonic printing with metal nanoparticles discussed above)! Yin et al.[107] demonstrated near-field two-photon optical lithography by using 120-fs laser pulses at 790 nm in an apertureless near-field optical microscope and produced lithographic features, even without optimization, of 70 nm in a commercial SU-8 resist. Thus, combining ANSOM with two-photon absorption can provide a spatial resolution as high as $l/10\ \lambda$, nearly a factor of two better than the resolution achieved in previous far-field two-photon lithography experiments.

Limitations of NSOM and ANSOM lithography include very low working distances, very shallow depth of field, and very long scan times for patterning large surfaces with a high-resolution image.

Holographic Lithography

Holography Background

As this section concerns holographic lithography, a short reminder about what holography is, is in

order. Holography comes from the Greek words for *whole* (holo) and *graphy* (or graphic, image), standing for "the entire image." The Hungarian physicist Dennis Gabor invented holography in 1948 and received the 1971 Nobel Prize for Physics for it. However, the field of holography only took off after 1962 with the invention of the laser, providing the coherent light needed to make a hologram. In holography two laser beams are used, a reference and an object beam, to snap pictures of an object. If these two beams are coherent, optical interference between the reference beam and the object beam, as a result of the superposition of the light waves, produces a series of fringes on standard photographic film. These fringes form a type of diffraction grating on the film, which is called a hologram. To reconstruct the image in the imaging medium one can go about it two different ways. First, on using the same reference light that was used to record the hologram, one obtains a virtual image of the object. Second, if one uses the phase conjugate light of the reference light to shine through the recording medium, the image of the object is reproduced.

Holographic Lithography or Interference Lithography

As mentioned above, a hologram is a complex interference pattern that can recreate a 3D image. In Figure 2.49 we show how a generic hologram is recorded. The "partially silvered" mirror in this figure reflects and transmits fractions of the incident light. This is useful because the reflected and transmitted beams are coherent (i.e., have a definite common phase) even when the incident light has a time-varying random phase. From Volume I, Chapter 5 we indeed recall that reflection from a surface produces a "phase-flip" of 180° in the reflected light. This is true for a 180° reflection. However, when light is reflected at an angle other than 180°, the phase change in the reflected light is the same as the reflection angle. Thus, light reflected at 90° has a 90° change in phase. In holographic lithography the specific object is the mask.

In holographic or interference lithography (Figure 2.50), a holographically constructed photomask replaces the standard photomask.[108,109] In the holographic recording or construction phase of the hologram (Figure 2.49), one uses the interference of two mutually coherent beams. A well-collimated flood laser (object beam) passes through the photomask and is diffracted by the mask features. This signal beam contains the amplitude and phase information of the photomask. When this diffracted object beam passes through the holographic recording layer, it interacts with the reference beam to create the interference pattern. The reference beam converts the amplitude and phase information into intensity information, which is stored in the photosensitive holographic medium (Figure 2.50a). The light-sensitive recording layer stores the holographic image data as variations in the refractive index of the photopolymer. Pattern printing or image reconstruction is accomplished by scanning a collimated laser illumination beam—the phase conjugate of the reference beam—to create the hologram. By interaction with the recorded hologram, the latter beam generates an image of the original photomask at precisely its original position in space (Figure 2.50b). When a

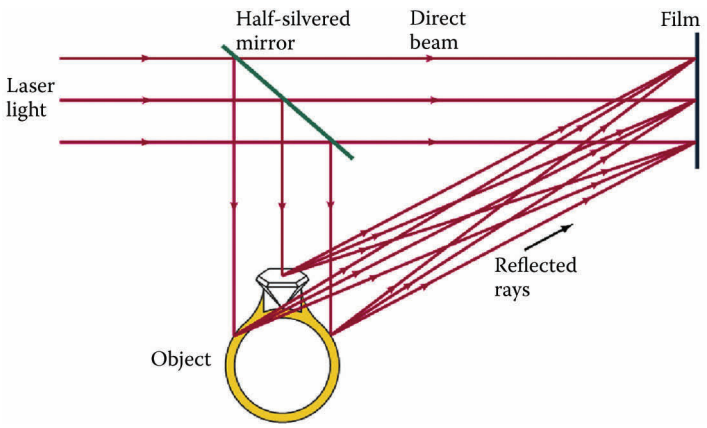


FIGURE 2.49 Recording a hologram. (From John G. Cramer.)

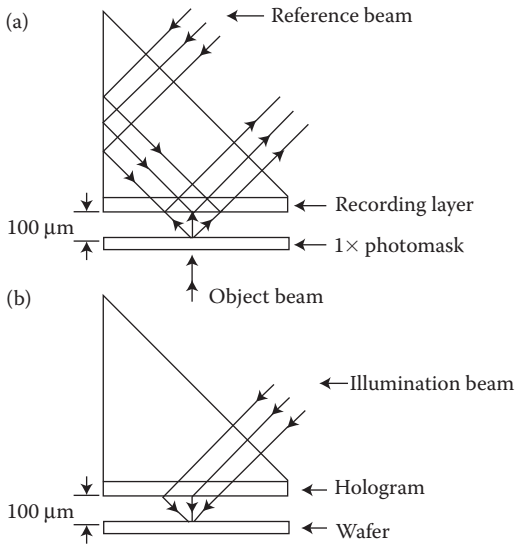


FIGURE 2.50 Basic arrangement for holographic lithography. (a) The photomask pattern forms a hologram in the polymer recording layer. (b) Using the illumination beam, the high-resolution holographic mask image is reconstructed into a printable masking layer in the resist-coated substrate. (From Brook, J., and R. Dandliker. 1989. Submicrometer holographic photolithography. *Solid State Technol* 32:91–94; and Omar, B., S. Clube, F. Hamidi, M.D. Struchen, and S. Gray. 1991. Advances in holographic lithography. *Solid State Technol* September:89–94.[108,109])

photoresist-coated substrate resides in this plane, a copy of the original mask can be printed. The image is diffraction limited, but because the holographic mask and the wafer can lie very close, a high NA and thus a very good resolution are possible (0.3 µm has been reported). Only the size of the photomask itself restricts the image field size. The holographic image of the original mask exactly overlays the wafer surface with a full field (large area recording).

Holographic lithography may also use the interference pattern of multiple coherent lasers. Simple grating patterns, 2D dot arrays, and 3D lattice patterns can be fabricated using two-, three-, and four-beam interference, respectively.

Holographic lithography is readily extendible to electrons. Spacings of a few nanometers to even less than 1 nm have been reported using electron holograms.[110] This is because the wavelength of an electron is always shorter than for a photon of the same energy. Atom lithography is also possible: the momentum of an atom is even larger than for electrons or photons, allowing even smaller wavelengths, per the de Broglie relation.

A benefit of using interference lithography is the quick generation of dense features over a wide area without loss of focus. A major drawback of interference lithography is that it is, in general, limited to patterning arrayed features. Recently, holographic lithography has emerged as a very promising technique for the inexpensive fabrication of high-quality 3D photonic crystals. Together with direct laser writing approaches, holographic lithography offers the attractive possibility of controlled incorporation of functional elements using direct laser writing in a second step.

Soft Lithography

Introduction

Soft lithography refers to a series of methods that use a patterned elastomer as a stamp, mold, or mask to generate micropatterns and microstructures instead of using a rigid photomask. Soft lithography has become an important tool in the nanofabrication arsenal.[111] Today the name is a collective term for a set of new techniques: microcontact printing (µ-CP), microtransfer molding (µ-TM), micromolding in capillaries (MIMIC), and microreplica molding.[112] The method is a very exciting research tool and may offer advantages over conventional methods for patterning of nonplanar substrates, unusual materials, and large areas.

In soft lithography, a master mold is first made by lithographic techniques, and an elastomeric stamp [e.g., polydimethylsiloxane (PDMS); see Figure 2.51] is cast from this master mold. A simple example procedure for making a PDMS (Sylgard 184, Dow Corning, http://www.dowcorning.com) stamp from a photolithographically patterned resist layer as master mold is outlined in Figure 2.52. A thin layer of SU-8 photoresist [SU-8 (50) from MicroChem, Newton,

- Low cost
- Easy fabrication
- Elastomeric/soft
- Moldable
- Surface chemistry
- Bonding
- Biocompatible

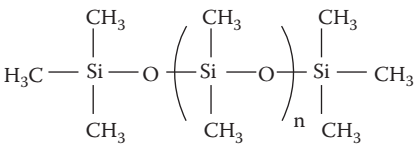


FIGURE 2.51 Polydimethylsiloxane (PDMS). (Sylgard 184, Dow Corning, http://www.dowcorning.com.)

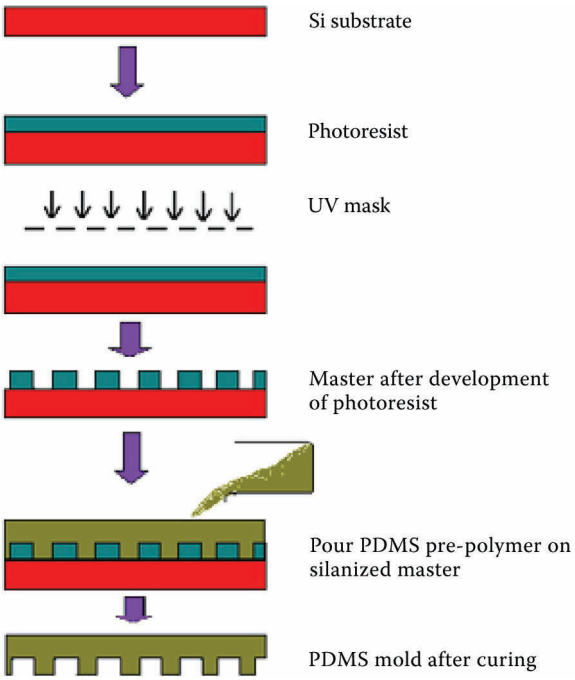


FIGURE 2.52 Making a PDMS mold. (Based on Xia, Y., and G.M. Whitesides. 1998. Soft lithography. *Angew Chem Int Ed* 37:550–75; and Xia, Y., and G.M. Whitesides. 1998. Soft lithography. *Ann Rev Mater Sci* 28:153–84.[113,114])

MA] is coated on a Si wafer. The resist is patterned by UV lithography. To make things really fast (<24 h) and inexpensive (<$20), the contact mask for the UV lithography may be a transparency on which—using drawing software such as Freehand—a design is printed with a high-resolution printer (>3300 dpi). After development, the photoresist is treated with (tridecafluoro-1,1,2,2-tetrahydrooctyl)-1-trichlorosilane (Hüls Chemicals, http://www.degussa.com) vapor to facilitate PDMS removal once cured. A 10:1 ratio of a PDMS mix is cast on the photoresist film and cured for 1 hour at 60°C in an oven.[111]

PDMS has many properties that make it suitable as a stamp material. It 1) provides a surface that has a low interfacial free energy (~21.6 dyn/cm), 2) is chemically inert, 3) is nonhygroscopic (does not swell with humidity), 4) passes gas easily (don't start laughing now), 5) has good thermal stability (~186°C in air), 6) is optically transparent down to ~300 nm, 7) is isotropic and homogeneous, 8) is durable (stamps may be used > 50 times over several months without degradation), and 9) has interfacial properties that are easy to modify.[114]

Many PDMS stamps can be generated, and each one may be used many times. The rubber stamp can be used in a number of different ways. Some different soft lithography processes are outlined below. For further reading on soft lithography, consult Xia et al.[113,114] and visit the soft lithography homepage at University of Washington/Washington Technology Center (http://www.ee.washington.edu/research/microtech/cam/CAMsoftlithhome.html). Another excellent site to follow the experimental steps involved in soft lithography is on JoVE (http://www.jove.com), the "YouTube" for experimental procedures.

Microcontact Printing

In microcontact printing (μ-CP), the PDMS rubber stamp is coated with an ink of the molecules (e.g., alkylthiols or a protein) that one wants to print in selected patterns on a solid substrate. During stamping, only the raised parts of the stamp collect the "ink." The inking of the substrate consists of self-assembled monolayers (SAMs) on the solid surface formed by covalent chemical reaction. The inked areas are self-passivating and exhibit very low interfacial tension that repels additional molecular layers; therefore, SAM forms only in areas of conformal contact between polymer and substrate. STM studies of the transferred SAMs reveal an achievable monolayer order indistinguishable from that found for SAMs prepared from solution.[53] The SAM pattern acts as a highly localized and efficient barrier to some wet etches. This lithographic technique—once the master is made—is not subject to diffraction or DOF limitations. The deformability of the elastomeric stamp allows it to accommodate rough surfaces. The method even works on rounded substrates (such as optical fibers or lenses) with radii of curvature of less than 10 μm. PDMS-based elastomers do not adhere to novolak or PMMA-based polymers, allowing convenient replication of masters formed by EBL.[53] The technique has been used, for example, to build an antibody grating on a Si wafer by inking the rubber stamp with an antibody solution.[115] The antibody grating alone produces insignificant diffraction, but on immunocapture of cells the optical phase change produces diffraction.

In Figure 2.53 we illustrate a generalized scheme for microcontact printing of proteins.

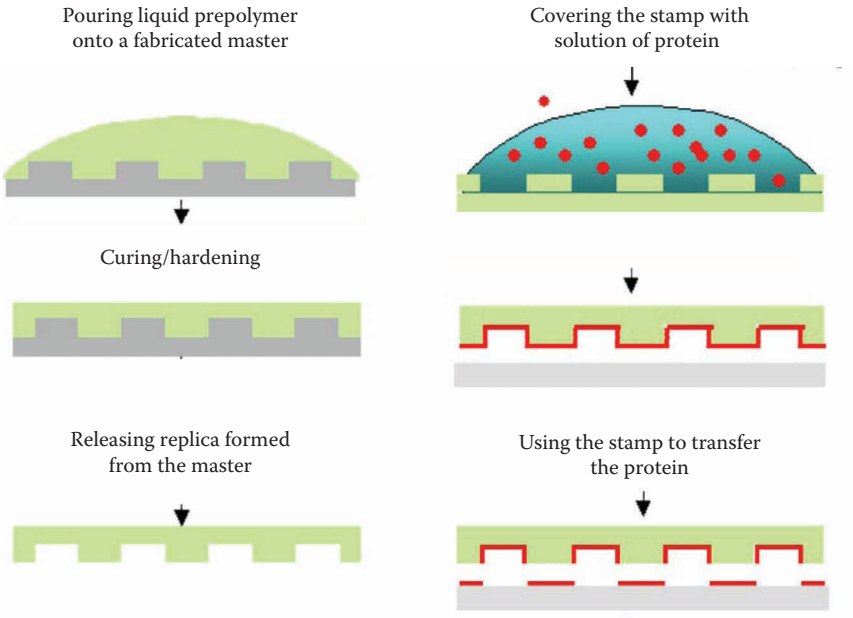


FIGURE 2.53 Microcontact printing of protein.

Microtransfer Molding

In microtransfer molding (μ-TM), illustrated in Figure 2.54, the rubber mold is filled with a polymer precursor, and the rubber stamp is pushed up against a substrate (e.g., a Si wafer or another flat sheet of PDMS).[116] The polymer in the rubber stamp relief is cured and transferred to the substrate, and the stamp is peeled off. Schueller et al.[111] used this technique to make curved glassy carbon microgrids. In this case, a photoresist was used as polymer precursor and cured while sandwiched between a patterned PDMS rubber stamp and a planar PDMS sheet. To obtain the curved carbon grid, the PDMS sandwich was deformed against a curved cylinder during curing. After removing the two thin strips of PDMS, the free-standing and curved photoresist grid was pyrolized (carbonized, see Volume III, Chapter 5) at high temperature in an argon atmosphere (Figure 2.55).

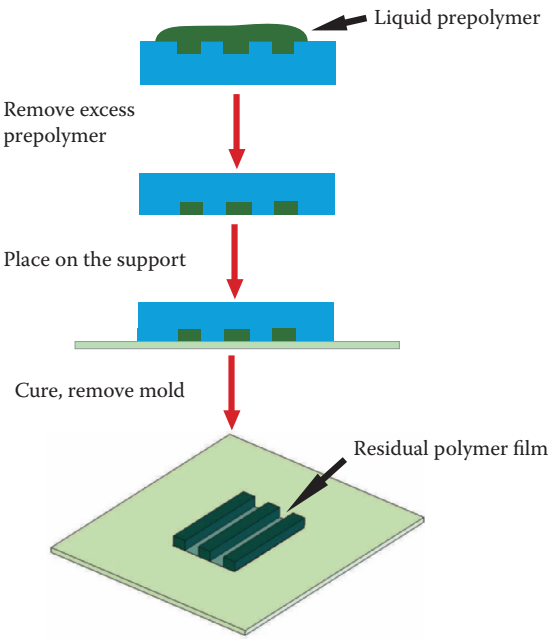


FIGURE 2.54 Schematic description of microtransfer molding (μ-TM). (Based on Xia, Y., and G.M. Whitesides. 1998. Soft lithography. *Angew Chem Int Ed* 37:550–75; and Xia, Y., and G.M. Whitesides. 1998. Soft lithography. *Ann Rev Mater Sci* 28:153–84.[113,114]) PU is prepolymer (polyurethane).

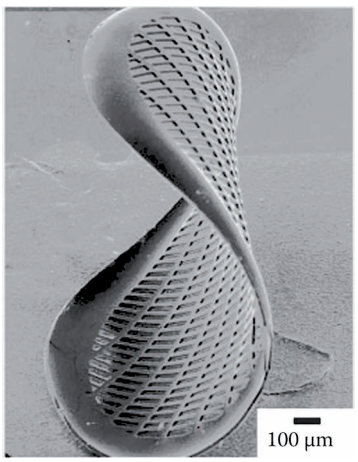


FIGURE 2.55 Curved glassy carbon structure. (Schueller, O.J.A., S.T. Brittain, and G.M. Whitesides. 1999. Fabrication of glassy carbon microstructures by soft lithography. *Sensors Actuators A* A72:125–39.[111] Courtesy of Dr. G.M. Whitesides, Harvard University.)

Micromolding in Capillaries

In micromolding in capillaries (MIMIC), illustrated in Figure 2.56, the rubber stamp is pushed up against a substrate, and liquid is applied to access holes in the mold. The liquid wicks into the cavities formed by the rubber mold against the substrate. Sometimes this process is pressure, electro-osmosis, or vacuum assisted.[117] The polymer is cured, and the stamp is removed. MIMIC has been used to fabricate polymeric transistors.[118]

Microreplica Molding

In microreplica molding, the master mold is replicated in PDMS by casting and curing the PDMS prepolymer (polyurethane, PU); this negative replica is then oxidized in an oxygen plasma for 1 minute and exposed to fluorinated silane for 2 hours to provide a surface with low adhesion to PDMS. PDMS is then cast against this negative replica, cured, and peeled away to reveal a positive replica of the original master.[109]

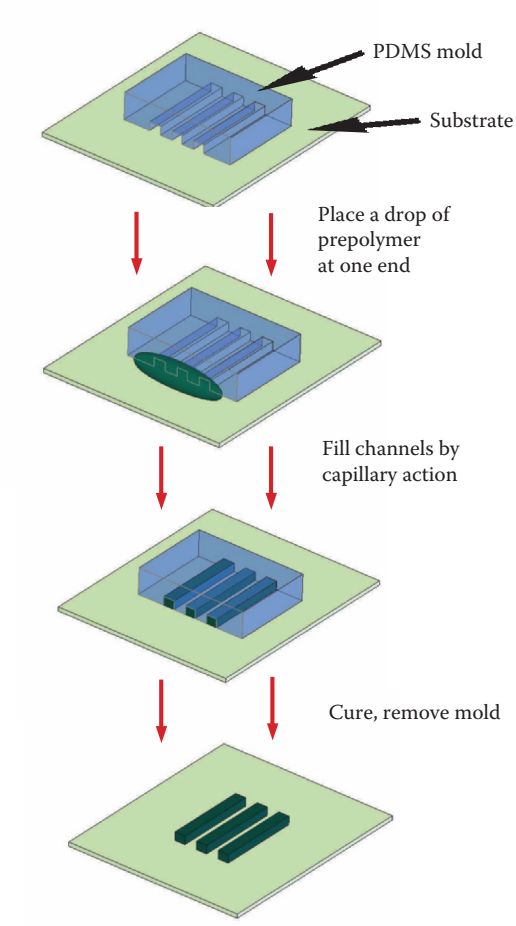


FIGURE 2.56 Schematic description of micromolding in capillaries (MIMIC). (Based on Xia, Y., and G.M. Whitesides. 1998. Soft lithography. *Angew Chem Int Ed* 37:550–75; and Xia, Y., and G.M. Whitesides. 1998. Soft lithography. *Ann Rev Mater Sci* 28:153–84.[113,114])

Soft Lithography Summary

Soft lithography is capable of generating structures as small as 30 nm and is especially attractive as a potential method for fabricating devices on nonplanar substrates. The major advantage of soft lithography is its short turnaround time. It is possible to go from design to production of replicated structures in less than 24 hours. The method is low cost, and, unlike photolithography, soft lithography is applicable to almost all polymers and thus to many materials (e.g., carbon, glasses) that can be prepared from polymeric precursors.

Some of the disadvantages of PDMS-based soft lithography include 1) shrinkage during curing (~1%), 2) swelling by nonpolar solvents such as toluene and hexane, 3) thermal expansion, 4) level of defects in printing SAMs, 5) softness of the material limits the achievable aspect ratio through sagging, and 6) deformation of the soft elastomeric stamps. These factors limit the accuracy in registration across a large area and may limit the practical utility in nanofabrication and multilayer devices.[114] Because soft materials are used, deformation of the stamp/mold, low reproducibility (because of distortion), and defects (yield) are problems that prevent this technology from being a viable manufacturing technique. The technique constitutes a very handy research tool but most likely will not reach the commercial manufacturing stage.

Soft lithography technology inspired new approaches to fabricate nanodevices that do not have commercial potential, i.e., NIL and SFIL (see above). For example, Krauss et al.[115] used NIL to fabricate a 400-Gb/in.^{2} storage device containing sub-10-nm minimum features. As we saw earlier, in NIL the master mold may be an electron beam and RIE-patterned SiO_2 layer (e.g., 10-nm-wide features with 40-nm period and 75-nm tall). This mold is then imprinted into a 90-nm-thick PMMA film on a Si disk. During the hot embossing, the mold and resist-coated disk are heated to 175°C and pressed together at 4.4 MPa for 10 minutes. After cooling to room temperature, the mold is separated from the disk, resulting in replication

of the nano-CD pattern in the PMMA film. Surface Logix, Inc. is the first commercial entity based entirely on the exploitation of soft lithography.

Questions

Most of the questions by Mr. Chuan Zhang, UC Irvine.

2.1: Extreme ultraviolet lithography (EUVL) is one of the most promising future lithography techniques. Explain why EUVL increases the resolution dramatically. The high cost is the major obstacle for the commercialization of this technology; list the major reasons for the high cost of EUVL.

2.2: Compare the differences between traditional projection photolithography and extreme ultraviolet lithography. Why do reflective optics have to be used in an EUVL system?

2.3: Why does x-ray lithography have a high depth of focus? LIGA is less promising as a future MEMS technique due to the requirement of a synchrotron source. List the competing technologies and explain why they might replace LIGA.

2.4: Is the resolution of e-beam lithography given by the spot size of the focused beam? Explain why or why not. What are proximity effects, and how can one minimize their impact on the resolution?

2.5: What are the differences between thermionic emission (also called Schottky emission or SE) and cold field emission (CFE)? What are the advantages of low-energy electron beam lithography?

2.6: What is the difference between direct write e-beam lithography and electron projection lithography? Describe the working principle of SCALPEL (scattering with angular limitation projection electron beam lithography).

2.7: Describe the major procedures in nanoimprint lithography.

2.8: How can a diblock copolymer be used in lithography? List some of the merits of block copolymer lithography.

2.9: How does a Fresnel zone plate work? What are the major merits of Fresnel zone plates over diffractive optics lenses? Can they be applied in lithography?

2.10: What is the definition of scanning proximal probe lithography? Which technologies can be counted as scanning proximal probe lithography?

2.11: How does near-field optics break Abbe's diffraction limit? What is the difference between NSOM and ANSOM lithography?

2.12: What are the major steps in soft lithography? Explain the reasons why they have not been widely used in industry.

2.13: How would you make a conical shaped PMMA structure that is 150 μm tall and features a top angle of 45°?

2.14: Design a miniaturized device incorporating both a lift-off process and a self-aligned mask step in its manufacture.

2.15: Why can only proximity masking be used in the case of x-ray lithography? What about projection printing with x-rays? Sketch the process for fabricating an x-ray mask. What are some of the positive attributes of x-ray lithography? What are the negative attributes?

2.16: Compare UV, x-ray, ion-beam, and electron-beam lithography. Summarize in a comparison Table. Which techniques are used mostly in the IC industry today? How are the photons or charged particles created in each case?

2.17: Most lithographies provide projected shapes. A truly 3D shape (e.g., a contact lens) is harder to make. Which lithographies are capable of providing 3D shapes?

2.18: How could one make scanning tunneling microscopy (STM) into a mass production ready lithography technique? What type of resists would you use?

2.19: Compare masks for the next-generation lithography. Consider the need for vacuum, materials, feature size, cost, etc.

2.20: What are the advantages and disadvantages of using e-beam lithography compared to typical photolithography?

2.21: What difficulties do you expect in trying to implement immersion lithography commercially?

References

1. Gibbs, W. W. 1997. Gordon E. Moore: Parts 1–3: The birth of the microprocessor. *Scientific American* (September 22, 1997); see also Moore, G. E. 1999. A pioneer looks back at semiconductors. *IEEE Design & Test of Computers* 16:8–14.
2. Semiconductor Industry Association. 1997. *The National technology roadmap for semiconductors: technology needs*. http://www.rennes.supelec.fr/ren/perso/gtourneu/enseignement/roadmap97.pdf
3. Singh, R. N., A. E. Rosenbluth, G. L.-T. Chiu, and J. S. Wilczynski. 1997. High-numerical aperture optical designs. *IBM J Res Dev* 41:39–48.
4. Kurzweil, R. 1999. *The age of the spiritual machine.* New York: Penguin Group.
5. Romankiw, L. T. 1993. Think small: one day it may be worth a billion. *Interface* Summer:17–57.
6. King, W. P., T. W. Kenny, K. E. Goodson, G. L. W. Cross, M. Despont, U. Düürig, H. Rothuizen, G. K. Binnig, and P. Vettiger. 2002. Design of atomic force microscope cantilevers for combined thermomechanical writing and ready in array operation. *Journal of Microelectromechanical Systems* 11: 765–74.
7. Terris, B. D., H. J. Mamin, M. E. Best, J. A. Logan, D. Rugar, and S. A. Rishton. 1996. Nanoscale replication for scanning probe data storage. *Appl Phys Lett* 69:4262–64.
8. Kraus, P. R., and S. Y. Chou. 1997. Nano-compact disks with 400 Gbit/in^2 storage density fabricated using nanoimprint lithography and read with proximal probes. *Appl Phys Lett* 71:3174–76.
9. Kurzweil, R. 2005. *The singularity is near.* New York: Viking.
10. Bjorkholm, J. E. 1998. EUV lithography: the successor to optical lithography. *Intel Technol J* Q3:1–8.
11. Wu, B. and A. Kumar. 2009. *Extreme Ultraviolet Lithography*. Columbus, OH: McGraw-Hill.
12. Bakshi, V., ed. 2009. *EUV Sources for Lithography*. SPIE press monograph. Bellingham, WA: SPIE Publications.
13. Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication*. New York: Plenum Press.
14. Arden, W., and K. H. Muller. 1989. Light vs. x-rays: how fine can we get? *Semicond Int* 12:128–31; see also Vora, K. D., B. Y. Shew, E. C. Harvey, J. P. Hayes, and A. G. Peele. 2008. Sidewall slopes of SU-8 HARMST using deep x-ray lithography. *J Micromech Microeng* 18:1–7.
15. Cataldo, A. 1999. IBM uses x-ray lithography to build prototypes. *EE Times*. http://www.eet.com/story/OEG19990414S0031
16. Dejule, R. 1999. Next-generation lithography tools: The choices narrow. *Semiconductor International*. http://www.semiconductor.net/article/197958-Next_Generation_Lithography_Tools_The_Choices_Narrow.php
17. Bourdillon, A., and Y. Vladimirsky. 2006. *X-ray lithography—On the sweet spot*. San Jose, CA: UHRL.
18. Ehrfield, W., P. Hagmann, A. Maner, and D. Münchmeyer. 1986. Fabrication of microstructures with high aspect ratios and great structural heights by synchrotron radiation lithography galvanoforming and plastic moulding (LIGA process). *Microelectronic Engineering* 4:35–56.
19. Ehrfeld, W. and A. Schmidt. 1998. Recent developments in deep x-ray lithography. *J Vac Sci Technol* B16:3526–34; see also Vora, K. D., B. Y. Shew, E. C. Harvey, J. P. Hayes, and A. G. Peele. 2008. Sidewall slopes of SU-8 HARMST using deep x-ray lithography. *J Micromech Microeng* 18:1–7.
20. Wollersheim, O., H. Zumaque, J. Hormes, J. Langen, P. Hoessel, L. Haussling, and G. Hoffman. 1994. Radiation chemistry of poly(lactides) as new polymer resists for the LIGA process. *J Micromech Microeng* 4:84–93.
21. Kouba, J., R. Engelke, M. Bednarzik, G. Ahrens, H. U. Scheunemann, G. Gruetzner, B. Loechel, H. Miller, and D. Haase. 2007. SU-8: Promising resist for advanced direct LIGA applications for high aspect ratio mechanical microparts. *Microsystem Technologies* 13:311–17.
22. O'Brien, M. J., and D. S. Soane. 1989. *Microelectronics processing*. Eds. D. W. Hess, and K. F. Jensen. Washington, DC: American Chemical Society.
23. Bley, P., W. Menz, W. Bacher, K. Feit, M. Harmening, H. Hein, J. Mohr, W. K. Schomburg, and W. Stark. 1991. Three-dimensional mechanical microstructures. *The 4th International Symposium on Microprocess Conference*, pp. 384–391, Kanazawa, Japan.
24. Levinson, H. J. 2005. *Principles of Lithography* (2nd edition). Bellingham, WA: SPIE Publications.
25. Rosolen, G. C. 1999. Automatically aligned electron beam lithography on the nanometer scale. *Appl Surface Sci* 144–45:467–71.
26. Lindquist, J., D. Ratkey, and P. Fischer. 1990. Do you have the right beam for the job? *Res Dev* June:91–98.
27. Wolf, B. 1995. *Handbook of Ion Sources*. New York: CRC Press.
28. Brewer, G., ed. 1980. *Electron-beam technology in microelectronic fabrication.* New York: Academic Press.
29. Whelan, C. S., D. M. Tanenbaum, D. C. La Tulipe, M. Isaacson, and H. G. Craighead. 1997. Low energy electron beam top surface image processing using chemically amplified AXT resist. *J Vac Sci Technol* B15: 2555–60.
30. Streblechenko, D., and M. R. Scheinfein. 1998. Magnetic nanostructures produced by electron beam patterning of direct write transition metal fluoride resists. *J Vac Sci Technol* A16:1374–79.
31. Farrow, R. C., J. A. Liddle, S. D. Berger, H. A. Huggins, J. S. Kraus, R. M. Camarda, R. G. Tarascon, C. W. Jurgensen, R. R. Kola, and L. Fetter. 1992. Alignment and registration schemes for projection electron lithography. *J Vac Sci Technol* B10:2780–83.
32. Harriott, L. R. 2001. Limits of lithography. *Proceedings of IEEE* 89:366–74; see also Harriott, L. R. 1997. Scattering with angular limitation projection electron beam lithography for sub-optical lithography. *J Vac Sci Technol* B15:2130.
33. Pfeiffer, H. C. PREVAIL: IBM's e-beam technology for next generation lithography. *IBM Micronews* 6:41–44.
34. Bindra, A. E-beam projection prevails in nanometer lithography. *Electronic Design*. http://electronicdesign.com/Articles/Index.cfm?AD=1&ArticleID=1345
35. Lee, Y.-H., R. Browning, and R. F. W. Pease. 1992. E-beam lithography at low voltages. *Proc SPIE* 1671:155–165.
36. Zlatkin, A., and N. Garcia. 1999. Functional scanning electron microscope of low beam energy with integrated electron optical system for nanolithography. *Microelectron Eng* 46:213–17.
37. Editorial. 1993. Novel electron-beam lithography system being explored at Cornell's NNF. *Solid State Technol* 36:25–26.
38. Marrian, C. R. K., and E. A. Dobisz. 1992. Scanning tunneling microscope lithography: a viable lithographic technology? *SPIE* 1671:166–75.

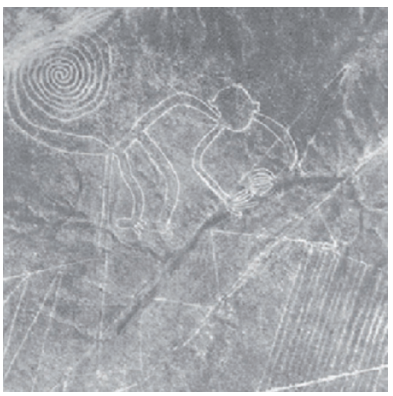
39. Temple, D., W. D. Palmer, L. N. Yadon, J. E. Mancusi, D. Vellenga, and G. E. McGuire. 1998. Si field emitter cathodes: fabrication, performance, and applications. *J Vac Sci Technol* A16:1980–90.
40. Editorial. 1991. Ion beam focused to 8-nm width. *Res Dev* September: 23.
41. Langford, R. M., P. M. Nellen, J. Gierak, and Y. Fu. 2007. Focused ion beam micro- and nanoengineering, pp. 417–23, in *Focused Ion Beam Microscopy and Micromachining*, Volkert, C. A., and A. M. Minor, eds. *MRS Bulletin*. Special issue.
42. Seliger, R. L., J. W. Ward, V. Wang, and R. L. Kubena. 1979. A high-intensity scanning ion probe with submicrometer spot size. *Appl Phys Lett* 34:310–12.
43. Sanchez, J. L., J. A. van Kan, T. Osipowicz, S. V. Springham, and F. Watt. 1998. A high resolution beam scanning system for deep ion beam lithography. *Nucl Instruments Methods Phys Res* B136–38:385–89.
44. van Kan, J. A., J. L. Sanchez, B. Xu, T. Osipowicz, and F. Watt. 1999. Micromachining using focused high energy ion beams: deep ion beam lithography. *Nucl Instruments Methods Phys Res* B148:1085–98.
45. Seidel, P., J. Canning, S. MacKay, and W. Trybula. 1998. Next generation lithography. *Semiconductor Fabtech* 7:147.
46. Tejeda, R., G. Frisque, R. Engelstad, E. Lovell, E. Haugeneder, and H. Löschner. 1999. Finite element simulation of ion-beam lithography mask fabrication. *Microelectron Eng* 46:485–88.
47. Colburn, M., S. Johnson, M. Stewart, S. Damie, T. C. Bailey, B. Choi, M. Wedlake, T. Michaelson, S. V. Sreenivasan, J. Ekerdt, and C. G. Wilson. 1999. Step and flash imprint lithography: a new approach to high-resolution patterning. *Proc SPIE* 3676:379–89.
48. Mutschler, A. S. 2007. Toshiba validates 22-nm CMOS imprint litho. *Electronic News*. http://www.edn.com/article/CA6491602.html
49. http://www.hitachi.com/ICSFiles/afieldfile/2007/08/07/r2007_technology_tp.pdf
50. Calvert, J. M., C. S. Dulcey, M. C. Peckerar, J. M. Schnur, J. H. J. Georger, G. S. Calabrese, and P. Sricharo-Enchaikit. 1991. New surface imaging techniques for sub-0.5 micrometer optical lithography. *Solid State Technol* 34:77–82.
51. Stroeve, P., and E. Franses, eds. 1987. *Molecular engineering of ultrathin polymeric films.* London: Elsevier.
52. Lehn, J.-M. 1990. Perspectives in supramolecular chemistry: from molecular recognition towards molecular information processing and self-organization. *Angew Chem Int Ed Engl* 29:1304–19.
53. Biebuyck, H. A., N. B. Larsen, E. Delamarche, and B. Michel. 1997. Lithography beyond light: microcontact printing with monolayer resists. *IBM J Res Dev* 41:159–70.
54. Marrian, C. R. K., and E. S. Snow. 1996. Proximal probe lithography and surface modification. *Microelectron Eng* 32:173–89.
55. Müller, H. U., C. David, B. Völkel, and M. Grunze. 1995. Nanostructuring of alkanethiols with ultrasharp field emitters. *J Vac Sci Technol* B13:2846–49.
56. Nuzzo, R. G., and D. L. Allara. 1983. Adsorption of bifunctional organic disulfides on gold surfaces. *J Am Chem Soc* 105:4481–83.
57. Bain, C. D., E. B. Troughton, Y. T. Tao, J. Evall, G. M. Whitesides, and R. G. Nuzzo. 1989. *J Am Chem Soc* 111: 321.
58. Delamarche, E., B. Michel, H. A. Biebuyck, and C. Gerber. 1996. Golden interfaces: the surface of self-assembled monolayers. *Adv Mater* 8:719–29.
59. Calvert, J. M., W. J. Dressick, C. S. Dulcey, M. S. Chen, J. H. Georger, D. A. Stenger, T. S. Koloski, and G. S. Calabrese. 1994. *Polymers for microelectronics: resists and diaelectrics.* Eds. L. F. Thompson, C. G. Willson, and S. Tagawa. Washington, DC: American Chemical Society.
60. Hamley, I. W., ed. 2004. *Developments in block copolymer science and technology.* Hoboken, NJ: John Wiley and Sons.
61. Flory, P. J. 1941. Thermodynamics of high polymer solutions. *J Chem Phys* 10:51–61.
62. Matsen, M. W., and F. S. Bates. 1996. Unifying weak- and strong-segregation block copolymer theories. *Macromolecules* 29:1091–98.
63. Mansky, P., P. Chaikin, and E. L. Thomas. 1995. Monolayer films of diblock copolymer microdomains for nanolithography applications. *J Mater Sci* 30:1987–92.
64. Black, C. T., R. Ruiz, G. Breyta, J. Y. Cheng, M. E. Colburn, K. W. Guarini, H.-C. Kim, and Y. Zhang. 2007. Polymer self assembly in semiconductor microelectronics. *IBM J Res Dev* 51:605–33.
65. Stix, G. 2004. *Nano patterning: IBM brings closer to reality chips that put themselves together.* http://www.fractal.org/Fractal-Research-and-Products/Nano-patterning.htm.
66. Segalman, R. A. 2005. Patterning with block copolymer thin films. *Mater Sci Eng* 48:191–226.
67. Hawker, C. J., and T. P. Russell. 2005. Block copolymer lithography: merging "bottom-up" with "top-down" processes. *MRS Bull* 30:952–66.
68. Stoykovich, M. P., and P. F. Nealey. 2006. Block copolymers and conventional lithography. *Materials Today* 9:20–29.
69. Stein, G. E., E. J. Kramer, X. Li, and J. Wang. 2007. Single-cystal diffraction from two-dimensional block copolymer arrays. *Phys Rev Lett* 98:086101.
70. Smith, H. I. 1996. A proposal for maskless, zone-plate-array nanolithography. *J Vac Sci Technol* B14:4318–22.
71. Menon, R., A. Patel, D. Gil, and H. I. Smith. 2005. Maskless lithography. *Mater Today* 8:26–33.
72. Angelo, M. D., M. V. Chekhova, and Y. H. Shih. 2001. Two-photon diffraction and quantum lithography. *Phys Rev Lett* 87:013602.
73. Boto, A. N., P. Kok, D. S. Abrams, S. L. Braunstein, C. P. Williams, and J. P. Dowling. 2000. Quantum interferometric optical lithography: exploiting entanglement to beat the diffraction limit. *Phys Rev Lett* 85:2733–36.
74. Einstein, A., B. Podolsky, and N. Rosen. 1935. Can quantum-mechanical description of physical reality be considered complete? *Phys Rev Lett* 47:777–80.
75. Parthenopoulos, D. A., and P. M. Rentzepis. 1989. Three-dimensional optical storage memory. *Science* 245:843–45.
76. Wu, E.-S., J. Strickler, R. Harrell, and W. W. Webb. 1992. Two-photon lithography for microelectronic application. *SPIE Proc* 1674:776–82.
77. Prasad, P. N. 2004. *Nanophotonics*. Hoboken, NJ: Wiley-Interscience.
78. Tanaka, T., H. B. Sun, and S. Kawata. 2002. Rapid sub-diffraction limit laser micronanoprocessing in a threshold material system. *Appl Phys Lett* 80:312–14.
79. Kik, P. G., S. A. Maier, and H. A. Atwater. 2002. Plasmon printing—a new approach to near-field lithography. *Mat Res Soc Symp Proc* 705:66–71.
80. Fang, N., and X. Zhang. 2003. Imaging properties of a metamaterial superlens. *Appl Phys Lett* 82:161.

81. Fang, N., H. Lee, C. Sun, and X. Zhang. 2005. Sub-diffraction-limited optical imaging with a silver superlens. *Science* 308:524–37.
82. Watanabe, F., M. Arita, T. Motooka, K. Okano, and T. Yamada. 1998. Diamond tip arrays for parallel lithography and data storage. *Jpn J Appl Phys* 37:L562–64.
83. Behm, R. J., N. Garcia, and H. Rohrer. 1989. *Proceedings: NATO ASI*. Dordrecht, the Netherlands: Kluwer.
84. Stroscio, J., and W. Kaiser, eds. 1993. *Scanning tunneling microscopy.* Boston: Academic Press.
85. Magno, R., and B. R. Bennett. 1997. Nanostructure patterns written in III-V semiconductors by an atomic force microscope. *Appl Phys Lett* 70:1855–57.
86. Wong, E. W., P. E. Sheehan, and C. M. Lieber. 1997. Nanobeam mechanics: elasticity, strength, and toughness of nanorods and nanotubes. *Science* 277:1971–75.
87. Sugimura, H., O. Takai, and N. Nakagiri. 1999. Multilayer resist films applicable to nanopatterning of insulating substrates based on current-injecting scanning probe lithography. *J Vac Sci Technol* B17:1605–08.
88. Soh, H. T., K. W. Guarinin, and C. F. Quate. 2001. *Scanning probe lithography.* Dordrecht, the Netherlands: Kluwer.
89. Wilder, K., C. F. Quate, B. Singh, and D. F. Kyser. 1998. Electron beam and scanning probe lithography: a comparison. *J Vac Sci Technol* B16:3864–73.
90. Völkel, B., A. Gölzhäuser, H. U. Müller, C. David, and M. Grunze. 1997. Influence of secondary electrons in proximal probe lithography. *J Vac Sci Technol* B15:2877–81.
91. Sugimura H., and N. Nakagiri. 1997. AFM lithography in constant current mode. *Nanotechnology* 8:A15–A18.
92. Gölzäuser, A. 1999. Nanolithography and electron holography with ultrasharp field emitters. *Appl Surface Sci* 141:264–73.
93. Spindt, C. A., I. Brodie, L. Humphrey, and E. R. Westerberg. 1976. Physical properties of thin-film emission cathodes with molybdenum cones. *J Appl Phys* 5284.
94. de Heer, W. A., A. Chatelain, and D. Ugarte. 1995. A carbon nanotube field-emission electron source. *Science* 270:1179–80.
95. Stroscio, J. A., and D. M. Eigler. 1991. Atomic and molecular manipulation with the scanning tunneling microscope. *Science* 254:1319–26.
96. Drexler, K. E. 1992. *Nanosystems: molecular machinery, manufacturing, and computation.* New York: Wiley.
97. Wallis, T. M., N. Nilius, and W. Ho. 2002. Electronic density oscillations in gold atomic chains assembled atom by atom. *Phys Rev Lett* 89:236802-1–2-4.
98. Nazin, G. V., X. H. Qiu, and W. Ho. 2003. Visualization and spectroscopy of a metal-molecule-metal bridge. *Science* 302:77–81.
99. Piner, R. D., J. Zhu, F. Xu, S. Hong, and C. A. Mirkin. 1999. Dip pen nanolithography. *Science* 283:661–63.
100. Ginger, D. S., H. Zhang, and C. A. Mirkin. 2004. The evolution of dip-pen lithography. *Angew Chem Int Ed* 43:30–45.
101. Salaita, K., Y. Wang, J. Fragala, R. A. Vega, C. Liu, and C. A. Mirkin. 2006. Massively parallel dip-pen nanolithography with 55000-pen two-dimensional arrays. *Angew Chem Int Ed* 45:7220–23.
102. Lim, J. H., D. S. Ginger, K. B. Lee, J. Heo, J. M. Nam, and C. A. Mirkin. 2003. Direct-write dip-pen nanolithography of proteins on modified silicon oxide surfaces. *Angew Chem Int Ed* 42:2309–12.
103. King, W. P., T. W. Kenny, K. E. Goodson, G. L. W. Cross, M. Despont, U. Durig, H. Rothuizen, G. Binnig, and P. Vettiger. 2001. Atomic force microscope cantilevers for combined thermomechanical data writing and reading. *Appl Phys Lett* 78:1300–02.
104. King, W. P., and K. E. Goodson. 2002. Thermal writing and nanoimaging with a heated atomic force microscope cantilever. *J Heat Transfer* 124:597.
105. Synge, E. H. 1928. A suggested method for extending microscopic resolution into the ultra-microscopic region. *Phil Mag* 6:356–62.
106. Likodimos, V., M. Labardi, L. Pardi, M. Allegrini, and M. Giordano. 2003. Optical nanowriting on azobenzene side-chain polymethacrylate thin films by near-field scanning optical microscopy. *Appl Phys Lett* 18:3313–15.
107. Yin, X., N. Fang, and X. Zhang. 2002. Near-field two-photon nanolithography using an apertureless optical probe. *Appl Phys Lett* 81:3663–65.
108. Brook, J., and R. Dandliker. 1989. Submicrometer holographic photolithography. *Solid State Technol* 32:91–94.
109. Omar, B., S. Clube, F. Hamidi, M. D. Struchen, and S. Gray. 1991. Advances in holographic lithography. *Solid State Technol* September:89–94.
110. Dunin-Borkowski, R. E., and T. Kasama. 2004. The prospect of three-dimensional induction mapping inside magnetic nanostructures by combining electron holography with electron tomography. *Microsc Microanal* 10:1010–11.
111. Schueller, O. J. A., S. T. Brittain, and G. M. Whitesides. 1999. Fabrication of glassy carbon microstructures by soft lithography. *Sensors Actuators A* A72:125–39.
112. Xia, Y., J. A. Rogers, K. E. Paul, and G. M. Whitesides. 1999. Unconventional methods for fabricating and patterning nanostructures. *Chem Rev* 99:1823–48.
113. Xia, Y., and G. M. Whitesides. 1998. Soft lithography. *Angew Chem Int Ed* 37:550–75.
114. Xia, Y., and G. M. Whitesides. 1998. Soft lithography. *Ann Rev Mater Sci* 28:153–84.
115. St. John, P. M., R. Davis, N. Cady, J. Czajka, C. A. Batt, and H. G. Craighead. 1998. Diffraction-based cell detection using a microcontact printed antibody grating. *Anal Chem* 70:1108–11.
116. Zhao, X.-M., Y. Xia, and G. M. Whitesides. 1996. Fabrication of three-dimensional micro-structures: microtransfer molding. *Adv Mater* 8:837–40.
117. Jeon, N. L., I. S. Choi, B. Xu, and G. M. Whitesides. 1999. Large-area patterning by vacuum-assisted micromolding. *Adv Mater* 11:946–50.
118. Rogers, J. A., Z. Bao, and V. R. Raju. 1998. Nonphotolithographic fabrication of organic transistors with micron feature sizes. *Appl Phys Lett* 72:2716–18.
119. Duffy, D. C., J. C. McDonald, O. J. A. Schueller, and G. M. Whitesides. 1998. Rapid prototyping of microfluidic systems in poly(dimethylsiloxane). *Anal Chem* 70: 4974–84.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Part II

Pattern Transfer with Subtractive Techniques



Stretching across the Nazca plains—like a giant map or blueprint left by ancient astronauts lie the famous Nazca Lines of Peru.

What subtractive, removing techniques were involved in drawing the monkey above?

Marc J. Madou

Introduction to Part II

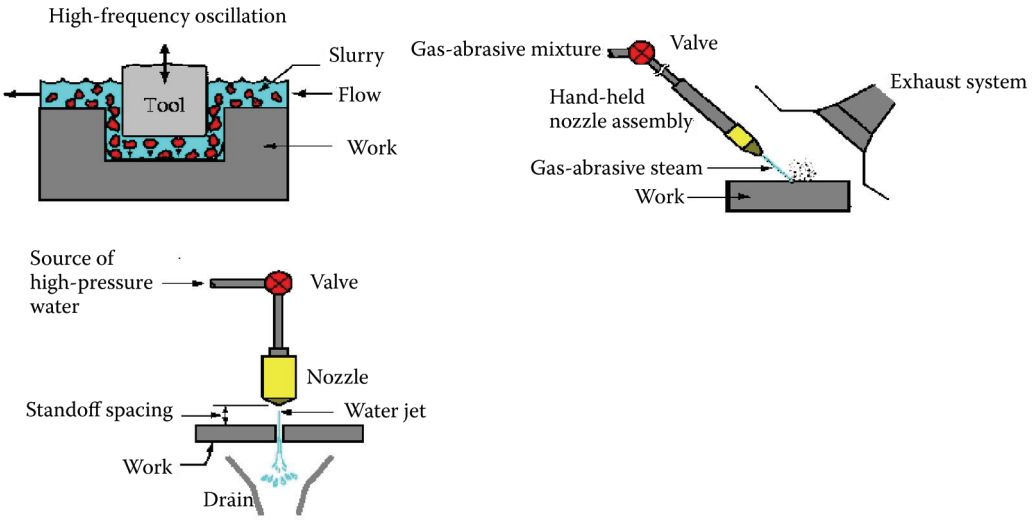
Chapter 3 Dry Etching
Chapter 4 Wet Chemical Etching and Wet Bulk Micromachining: Pools as Tools
Chapter 5 Thermal Energy-Based Removing
Chapter 6 Mechanical Energy-Based Removing

Introduction to Part II

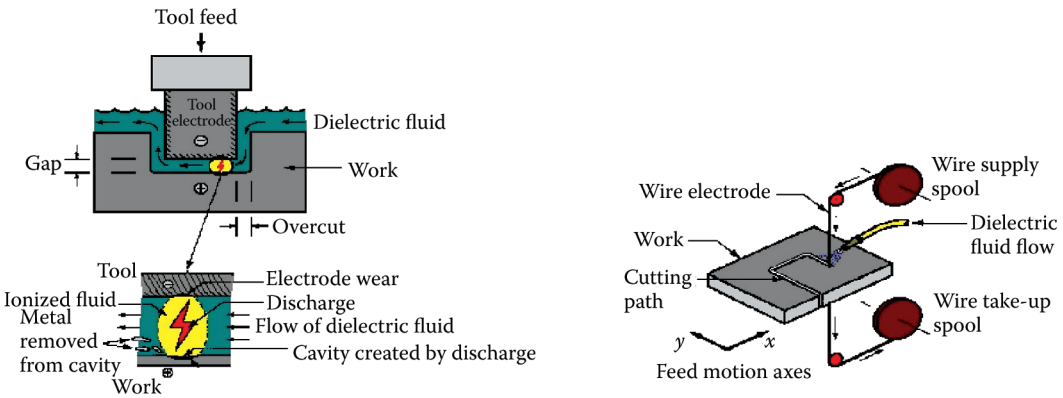
In manufacturing, removing and forming are the two primary processes. Removing processes destroy cohesion among particles as exemplified in wet chemical etching, electrical discharge machining (EDM), traditional mechanical turning, and laser drilling. Forming creates an original shape from a molten mass, a gaseous state, a solution, a plasma, or solid particles. During forming processes, cohesion is created among individual particles; examples include plastic molding, thin film deposition by evaporation or sputtering, and electrochemical metal deposition. In this book, we are interested in forming and removing processes enabling the manufacture of precision miniature machines. Materials are either removed from (subtracted) or added to (formed) a work piece or device under construction (usually in a selective manner).

Manufacturing techniques are often categorized according to the energy used at the work piece as mechanical, chemical, or thermal. Whereas mechanical machining mostly involves material removal,*

* An example of a mechanical additive technique is physical vapor deposition (see Chapter 7) where material is deposited by direct line-of-sight impingement deposition. The depositing material might be generated through the application of thermal energy (e.g., thermal evaporation) or mechanical energy to the source material (e.g., sputter deposition).

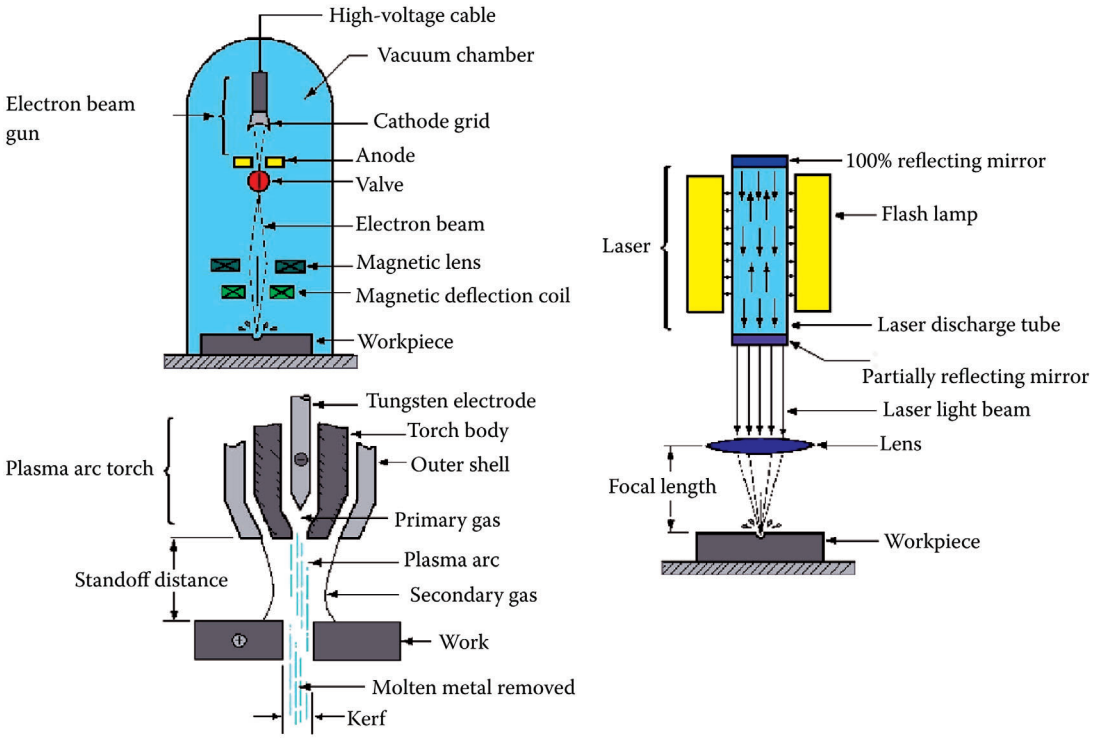


Mechanical machining: ultrasonic machining (S), abrasive jet-machining (S), and water-jet machining (S).

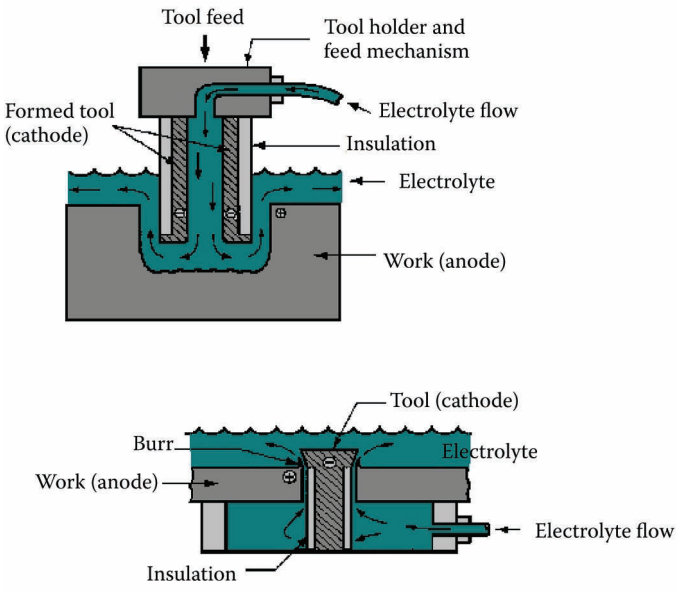


Thermal machining: electrical discharge machining (EDM) (S) and electrical discharge wire cutting (ED-WC) (S).

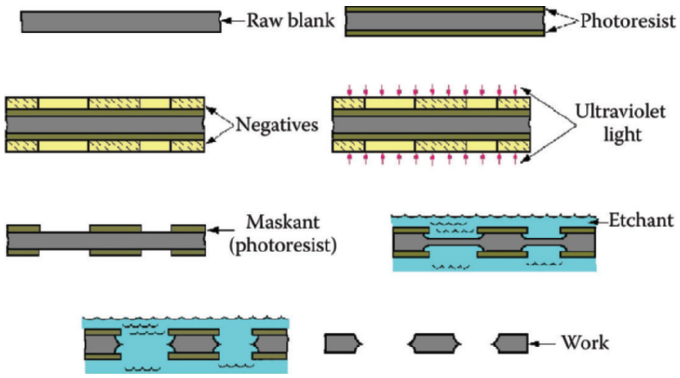
FIGURE II.1 Examples of manufacturing processes grouped according to the energy used at the work piece as mechanical, chemical, and thermal. S = subtractive, A = additive.



Thermal machining: electron-beam machining (EBM) (S shown here but can also be A), laser-beam machining (LBM) (S shown here but can also be A); also called high-energy beam machining.



Chemical machining: electrochemical machining (ECM) (S shown here but can also be A).



Chemical machining: chemical milling (CM) (S).

FIGURE II.1 (*Continued*)

chemical and thermal can be additive (A) or subtractive processes (S). In Part II of this book we deal with material removal processes only; additive technologies are the focus of Part III.

Some examples of subtractive techniques are illustrated in Figure II.1. In categorizing machining techniques, some authors group e-beam, laser, and ion-beam machining under high-energy beam machining.

Table II.1 lists the most important subtractive processes encountered in miniaturization science and includes wet and dry etching, laser machining, ultrasonic drilling, electrical discharge machining (EDM), traditional precision machining, and focused ion-beam (FIB) milling. The manufacturing category, applications (which often involve Si), typical material removal rates, relevant references, and a few remarks on the technique supplement the list.

One of the most widely used subtractive techniques for integrated circuits (ICs) and micro- and nanomachining listed in this table is etching. Etching can be described as pattern transfer by chemical/physical removal of a material from a substrate—often in a pattern defined by a protective masking layer such as a resist or an oxide. Etch parameters—dry or wet—include (for more details see Table 3.5):

1. Etch rate: Rate at which material is removed from the substrate
2. Uniformity: Etch rate constancy across the substrate
3. Throughput: Amount of substrates/wafers etched during one process cycle
4. Directional control: Control of the horizontal and vertical etch rate—isotropic or anisotropic
5. Selectivity of etch:
   a. Controlled by etchant composition
   b. Controlled by etch rate

From the etching processes listed in Table II.1, we review gas and vapor phase dry etching processes, including chemical, physical (purely mechanical), and chemical-physical (mostly chemical with some mechanical assist) in Chapter 3.

Chapter 4 consists of two parts: one deals with wet *chemical etching* and the other with *wet bulk micromachining*, the latter also being chemical etching but optimized for the manufacture of ICs, nano-, and micromachines. Chemical etching techniques cover purely chemical processes as in *chemical milling* (CM), *photochemical machining* (PCM), *electrochemical machining* (ECM) (including μ-ECM), and combinations thereof. Of these chemical etching techniques, chemical milling is the simplest: it only requires a liquid solution that selectively dissolves a specific material (pools as tools).

Chapter 5 discusses thermal removing techniques where thermal energy, provided by a heat source, melts and/or vaporizes material from the work piece. Examples of thermal techniques covered in this chapter are EDM (including μ-EDM), laser- and e-beam machining (LBM and EBM), and plasma arc cutting.

In Chapter 6, we cover traditional and nontraditional mechanical removing techniques. Traditional mechanical material removing technologies covered are mechanical precision machining and abrasive wheel machining. In nontraditional mechanical machining we deal with ultrasonic drilling, electrolytic in-process dressing (ELID) grinding, water jet, abrasive water jet, abrasive jet machining, and focused ion beam (FIB) milling. In Chapter 6, we also highlight recent developments toward miniaturized mechanical manufacturing equipment: the so-called desktop factories.

Table II.1 Subtractive Processes Important in ICs, Micro-, and Nanomachining

Category	Subtractive Technique	Applications	Typical Etch Rate	Remark	References
Chemical	Dry chemical etching	Resist stripping, isotropic features	Si etch rate: 0.1 μm/min (but with more recent high-density plasmas up to 6 μm/min)	Resolution better than 0.1 μm, loading effects	1, 2 Chapter 3
	Physical/ Chemical etching	Very precise pattern transfer	Si etch rate: 0.1–1 μm/min (but with more recent high-density plasmas up to 6 μm/min)	Most important dry etching technique	2, 3 Chapter 3
	Vapor phase etching with XeF_2 (and other interhalogens)	Isotropic Si etching in presence of Al and SiO_2	Si etch rate: 1–3 μm/min	Very selective, fast, and very simple	4 Chapter 3
	Chemical machining (CM)	Shallow removal (up to 12 mm) on large flat or curved surfaces; blanking of thin sheets; in Si spheres, domes, grooves, etc.	General 0.025–0.1 mm/min; in case of Si polishing 50 μm/min with stirring (RT, acid)	Little control, simple; low tooling and equipment cost; suitable for low production cost	5 Chapter 4
	Anisotropic wet chemical etching	Si angled mesas, nozzles, diaphragms, cantilevers, bridges	1 μm/min on (100) Si (90°C, alkaline)	With etch-stop more control, simple	5 Chapter 4
	Wet photoetching	Etches p-type layers in p-n junctions	Etches p-Si up to 5 μm/min (acid)	No electrodes required	6 Chapter 4
Electrochemical	Electrochemical etching (ECM)	Complex shapes with deep cavities; etches p-Si and stops at n-Si in an p-n junction); etches n-Si of highest doping (in n/n+)	Highest rate of material removal. V: 5–25 dc; A: 1.5–8 A/mm^2; 2.5–12 mm/min, depending on current density; etches p-Si at 1.25–1.75 μm/min [(100) plane]	Complex (requires contacts) expensive tooling and equipment; high-power requirement; medium to high production quantity	7, 8 Chapter 4
	Electrochemical grinding (ECG)	Cutting off and sharpening of hard materials, such as tungsten-carbide tools. Also used for honing	Higher removal rates than grinding; A: 1–3 A/mm^2		9 Chapter 4
	Photoelectrochemical etching	Etches n-Si in p-n junctions, production of porous Si	Si etch rate: 5 μm/min (acid)	Complex, requires electrodes and light	9 Chapter 4
Thermal	Laser machining (with and without reactive gases)	Cutting and hole making in thin materials; circuit repair, resistor trimming, hole drilling, labeling of Si wafers	Drilling a hole in Si with a Nd:YAG (400-W laser): 1 mm/s (3.5 mm deep and 0.25-mm diameter)	Expensive equipment, does not require vacuum, heat-affected zone; laser beams can focus to a 1-μm spot, etch with a resolution of 1 μm^3	10 Chapter 5
	Electron-beam machining	Cutting and small hole making in thin films; etching of circuits in microprocessors, large number of simple holes in hard to machine materials	Material removal rate is ~0.1 mg/s	Vacuum is needed, very expensive equipment, fast, drilling rates of 2000 holes/s	11 Chapter 5

(Continued)

Table II.1 Subtractive Processes Important in ICs, Micro-, and Nanomachining (continued)

Category	Subtractive Technique	Applications	Typical Etch Rate	Remark	References
	Electrical discharge machining (EDM)	Shaping and cutting complex parts made of hard materials; some surface damage may result; drilling holes and channels in hard brittle metals	V: 50–380 dc; A: 0.1–500; typical removal rate for metals: 0.3 cm^3/min	Poor resolution (>50 μm), only conductors, expensive tooling and equipment	12 Chapter 5
Mechanical	Mechanical turning, drilling, and milling	Almost all machined objects surrounding us	Removal rates of turning and milling of most metals: 1–50 cm^3/min, for drilling: 0.001–0.01 cm^3/min	Prevalent machining technique	13 Chapter 6
	Physical dry etching, sputter etching, and ion milling	Si surface cleaning, unselective thin film removal	Si etch rate: 300 Å/min	Unselective and slow, plasma damage	3 Chapter 3
	Water-jet machining (WJM)	Cutting all types of nonmetallic materials 25 mm and thicker	Varies considerably with work piece material		14 Chapter 6
	Abrasive water-jet machining (AWJM) and abrasive jet machining (AJM)	Single or multilayer cutting of metallic and nonmetallic materials; cutting, slotting, deburring, flash removal, and cleaning of metallic and nonmetallic materials	Up to 7.5 m/min	Tends to round off sharp edges; some hazards because of airborne particles	14 Chapter 6
	Abrasive grinding				15 Chapter 6
	Electrolytic in-process dressing (ELID) grinding				16 Chapter 6
	Ultrasonic drilling	Holes in quartz, silicon nitride bearing rings	Typical removal rate of Si: 1.77 mm/min	Especially useful for hard, brittle materials	17 Chapter 6
	Focused ion-beam (FIB) milling	Microholes, circuit repair, microstructures in arbitrary materials	Typical Si etch rate: 1 μm/min	Long fabrication time: >2 h including setup	18 Chapter 6

Different-colored entries correspond to different chapters where more details can be found. RT, room temperature; chem. etch., chemical etching; iso., isotropic; anis., anistropic.
Source: Kalpakjian, S., and S. R. Schmidt. 2003. *Manufacturing processes for engineering materials*, 4th ed. Upper Saddle River, NJ: Prentice Hall. See also http://www.engineershandbook.com.

References

1. Flamm, D. L. 1993. Feed gas purity and environmental concerns in plasma etching: Part I. *Solid State Technol* October:49–54.
2. Pandhumsoporn, T., M. Feldbaum, P. Gadgil, M. Puech, and P. Maquin. 1996. *Micromachining and microfabrication process technology II* Austin, TX: SPIE.
3. Manos, D. M., and D. L. Flamm, eds. 1989. *Plasma etching: an introduction.* Boston: Academic Press.
4. Chang, F. I., R. Yeh, G. Lin, P. B. Chu, E. Hoffman, E. J. Kruglick, K. S. J. Pister, and M. H. Hecht. 1995. Gas-phase silicon micromachining with xenon difluoride. *Proceedings: SPIE microelectronic structures and microelectromechanical devices for optical processing and multimedia applications* 2641:117–28.
5. Kern, W., and C. A. Deckert. 1978. *Thin film processes.* Eds. J. L. Vossen, and W. Kern. Orlando, FL: Academic Press.
6. Yoshida, T., T. Kudo, and K. Ikeda. 1992. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '92)*. Travemunde, Germany: IEEE.
7. Jackson, T. N., M. A. Tischler, and K. D. Wise. 1981. An electrochemical P-N junction etch-stop for the formation of silicon microstructures. *IEEE Electron Device Lett* EDL-2:44–45.
8. Meek, R. L. 1971. Electrochemically thinned N/N+ epitaxial silicon-method and applications. *J Electrochem Soc* 118:437.
9. Engineer's Handbook.
10. Watanabe, Y., Y. Arita, T. Yokoyama, and Y. Igarashi. 1975. Formation and properties of porous silicon and its applications. *J Electrochem Soc* 122:1351–55.
11. Chryssolouris, G. 1991. *Laser machining*. New York: Springer Verlag.
12. Kalpajian, S. 1984. *Manufacturing processes for engineering materials.* Reading, MA: Addison-Wesley.
13. DeVries, W. R. 1992. *Analysis of material removal processes.* New York: Springer Verlag.
14. Lorincz, J. November 2009. Waterjets: evolving from macro to micro. *Manufacturing Engineering*. Society of Manufacturing Engineers and http://www.drolsenslab.com/
15. Salmon, S. February 2010. What is abrasive machining? *Manufacturing Engineering*. Society of manfacturing Engineers.
16. Ohmori, H., and I. D. Marinescu, eds. 2011. *Handbook of ELID technologies*. Boca Raton, FL: CRC Press, *The 3rd International conference on Ultraprecision and ELID-Grinding*, 8–9 September, 2010, Toledo, OH and www.nexsys.co.jp/contents_e/technology.htm
17. Moreland, M. A. 1992. *Engineered materials handbook*. Ed. S. J. Schneider. Metals Park, OH: ASM International.
18. Vasile, M. J., C. Biddick, and S. Schwalm. 1994. Microfabriction by ion milling: the lathe technique. *J Vac Sci Technol* B12:2388–93.

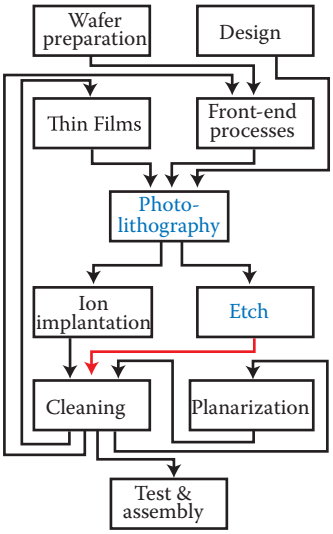
Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

3

Dry Etching

Tek-Vac Model DRIE-1200-LL-ICP: High Density Plasma Etching System for 200 mm substrates

Etching in IC, micro- and nano-fabrication



I think there is a world market for maybe five computers.

Thomas Watson
Chairman, IBM, 1943

Science has a way of getting us to the future without consulting the futurists and visionaries.

Robert Park
Voodoo Science

Outline

Introduction

Dry Etching: Definitions and Jargon

Plasmas or Discharges

Physical Etching: Ion Etching or Sputtering and Ion-Beam Milling

Plasma Etching (Radical Etching)

Physical/Chemical Etching

Deep Reactive Ion Etching

Vapor-Phase Etching without Plasma (XeF_2)

Dry Etching Models—In Situ Monitoring

Examples

Questions

References

Introduction

Lithography steps precede a number of subtractive and additive processes. Materials are either removed or added to a device, usually in a selective manner, transferring the lithography patterns onto integrated circuits (ICs) micro- or nanomachines. This chapter deals with material removal by dry etching processes. In dry etching, the solid substrate is etched by gaseous species. While introducing resist stripping in Chapter 1, we learned that the popularity of dry stripping of resists (ashing) and the need for better control of critical dimension (CD) precipitated major research and development efforts in all types of plasma-based dry etching

processes. The preference of dry etching over wet etching methods is based on a variety of advantages: fewer disposal problems, less corrosion of metal features in the structure, and less undercutting and broadening of photoresist features; that is, better CD control (nanometer dimensions compared with 3 μm or larger for wet etching) and, under the right circumstances, a cleaner resulting surface. With the new generations of ICs, all with sub-100-nm geometries, surface tension precludes a wet etchant from reaching down between photoresist features, whereas dry etching avoids this problem. In Figure 3.1, dry anisotropic and wet isotropic etchings for patterning a thin film of SiO_2 on a Si substrate are compared. In step (a) the photoresist protects parts of the underlying SiO_2 from dry or wet etching. In steps (b) and (c) the etching proceeds until the underlying Si substrate is reached, undercutting the photoresist in wet etching but not in dry etching. In step (d) the resist is stripped. As in IC manufacture, dry etching has evolved into an indispensable technique in miniaturization science. After the commercial successes with wet-etched Si micromachined products in the 1980s and early 1990s, the maturation of flexible, generic, and fast anisotropic deep reactive ion etching (DRIE) techniques in the late 1990s spawned a second wave of Si miniaturized commercial products.

Dry Etching: Definitions and Jargon

In dry etching, substrates are etched without wet chemicals or rinsing and enter and leave the etching system in a dry state. The term *dry etching* covers a family of methods by which a solid surface is etched in the gas or vapor phase physically by ion bombardment,

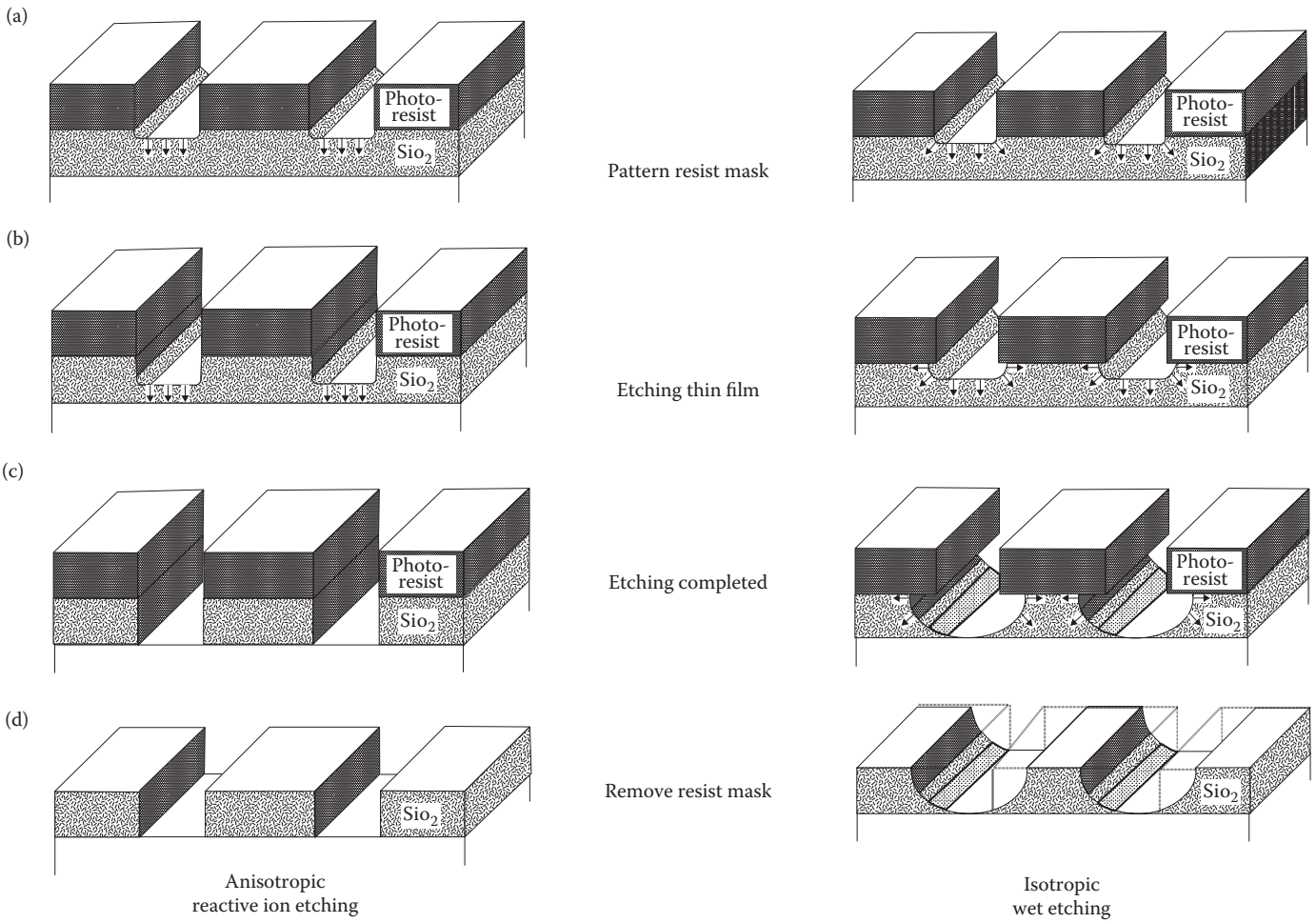


FIGURE 3.1 Comparison of dry anisotropic and wet isotropic etching to pattern a thin SiO_2 film. In step (a) the photoresist protects parts of the underlying SiO_2 from dry or wet etching. In steps (b) and (c) the etching proceeds until the underlying Si substrate is reached, undercutting the photoresist in wet etching but not in dry etching. In step (d) the resist is stripped.

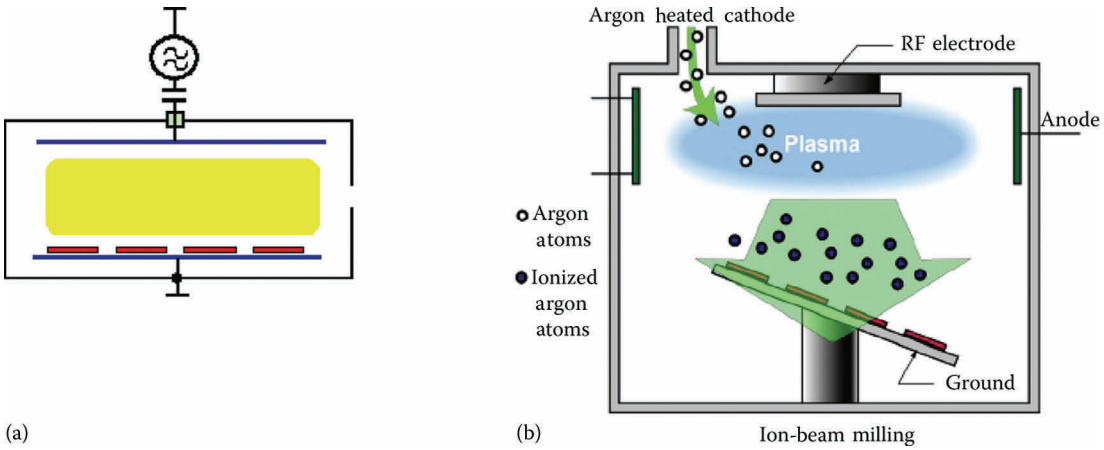


FIGURE 3.2 Plasma-assisted dry etching is categorized according to the specific setup as either *glow discharge* etching (diode setup) (a) or *ion-beam* etching (triode setup) (b).

chemically by a chemical reaction through a reactive species at the surface, or by combined physical and chemical mechanisms or reactive ion etching (RIE). A key ingredient in dry etching involves the creation of plasma. Plasma-assisted dry etching is categorized according to the specific setup as either *glow discharge* etching (diode setup) or *ion-beam* etching (triode setup). Using glow discharge techniques, a broad plasma is generated in the same part of the vacuum chamber where the substrate is located; when using ion-beam techniques, plasma is generated in a separate chamber or in another part of the same vacuum chamber from which ions are extracted and directed in a beam toward the substrate by a number of electrode grids. In Figure 3.2 we compare a dry etching diode and dry etching triode setup.[1] It is also common to differentiate between 1) physical sputter/ion etching and ion-beam etching (IBE)[2] or also ion-beam milling (IBM), 2) chemical plasma etching (PE), and 3) synergetic reactive ion etching (RIE). Each of these techniques will be discussed in detail further below, but here we present a short dry etching synopsis:

1. In physical sputter/ion etching and ion-beam milling (IBE), etching occurs as a consequence of a purely physical effect, namely, momentum transfer between energetic Ar^+ ions and the substrate surface. Some type of chemical reaction takes place in all the other dry etching methods, but not in this one.
2. In chemical plasma etching (PE), neutral chemical species such as chlorine or fluorine atoms generated in the plasma diffuse to the substrate, where they react to form volatile products with the layer to be removed. The only role of the plasma is to supply gaseous, reactive etchant species.
3. In the case of physical/chemical etching (RIE), line-of-sight-impacting ions damage the surface, inducing highly anisotropic chemical reactions of the surface with plasma neutrals, or, alternatively, a passivating layer is cleared by the ion bombardment, clearing horizontal surfaces only and again resulting in highly anisotropic etching.

We can also distinguish plasmas produced in a DC field (0 Hz), an AC field (50 kHz and up), a radiofrequency (RF) field (13.6–27 MHz), and in a microwave field (300 MHz–10 GHz). Finally, dry plasma etching systems can also be divided into single-wafer and batch reactors.

In Table 3.1, we review the most common dry etching methods and clarify the associated jargon in the legend. Of the techniques listed, PE and RIE are the most widely used in IC manufacture and micro- and nanomachining.

In selecting a dry etching process, the desired shape of the etch profile and the selectivity of the etching process require careful consideration. Figure 3.3 shows different possible dry etch profiles, where, depending on the etching mechanism used, nondirectional, i.e., isotropic, or directional, i.e., vertical, or anisotropic etch profiles are obtained. In dry etching, anisotropic etch profiles—directional or vertical—can be generated in single crystalline, polycrystalline, and amorphous materials. The anisotropy of dry etching is not a result of aniso-

Table 3.1 Some Popular Dry Etching Systems

	CAIBE	RIBE	IBE	MIE	MERIE	RIE	Barrel Etching	PE
Pressure (Torr)	~10^{-4}	~10^{-4}	~10^{-4}	10^{-3}–10^{-2}	10^{-3}–10^{-2}	10^{-3}–10^{-1}	10^{-1}–10^{0}	10^{-1}–10^{1}
Etch Mechanism	Chemical/ physical	Chemical/ physical	Physical	Physical	Chemical/ physical	Chemical/ physical	Chemical	Chemical
Selectivity	Good	Good	Poor	Poor	Good	Good	Excellent	Good
Profile	Aniso-tropic or isotropic	Aniso-tropic	Aniso-tropic	Aniso-tropic	Aniso-tropic	Isotropic or aniso-tropic	Isotropic	Isotropic or aniso-tropic
Commercial equipment	Ionfab 300 Plus	Ionfab 300 Plus	Ionfab 300 Plus	MIE-710, Materials Research	Tegal 6000 MERIE	Unaxis Plasma-Therm 790	Polaron K1050X from Quorum Techno-logies	Technics PE-IIA

CAIBE, chemically assisted ion-beam etching; RIBE, reactive ion-beam etching; MIE, magnetically enhanced ion etching; MERIE, magnetically enhanced reactive ion etching; RIE, reactive ion etching; PE, plasma etching. See text.

tropy in the etching rate in different crystallographic directions, as in the case of anisotropic wet chemical etching in single crystal Si (Chapter 4); rather, the degree of anisotropy is controlled by the plasma conditions. Selectivity of a dry etch refers to the difference in etch rate between the mask and the film to be etched, again controllable by plasma conditions.

At low pressures, in the 10^{-3}–10^{-4} Torr range, anisotropic etching is easy, but accomplishing selectivity is difficult. The reason is that etching in this regime is mechanical in nature: the substrate is hit directionally, and atoms are ejected like a billiard ball. Physical ion-beam etching (IBE) at those pressures results in tapered profiles as shown in Figure 3.3a, low selectivity, and low etch rates because the mechanical energy transfer is inefficient. At higher pressures (1 Torr), in plasma etching (PE), chemical effects dominate, leading to better selectivity and higher etch rates than in the case of ion beams, but etched features are isotropic (Figure 3.3b). Chemical etching reactions are fast and efficient compared with mechanical etching, where atoms are ejected by mechanical force. Reactive ion etching (RIE) enables profile control as a result of the synergistic combination of physical ion sputtering with the chemical activity of neutral reactive species with a high etch rate and high selectivity. In the extreme case of anisotropic etching, perfect vertical sidewalls result (no lateral undercut) with perfect retention of

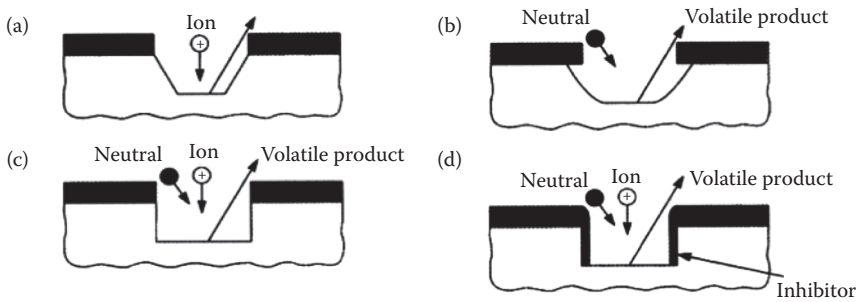


FIGURE 3.3 Important dry etching profiles associated with different dry etching techniques. (a) Sputtering. (b) Chemical etching in plasma at low voltage and relative high pressure leads to isotropic features and lateral undercuts. (c) Ion-enhanced etching: physical-chemical is the most perfect image transfer as the undercutting is limited by the combined action of physical and chemical etching. The low pressure and high voltage lead to directional anisotropy. (d) Ion-enhanced inhibitor. Sidewalls are protected from undercutting by a surface species (e.g., a polymer), which starts etching when hit by a particle such as an ion, a photon, or an electron. Because very few particles hit the sidewalls (high voltage and high pressure), undercutting is suppressed. (From Flamm, D. L. 1993. Feed gas purity and environmental concerns in plasma etching: Part I. *Solid State Technol* October:49–54. With permission.[75])

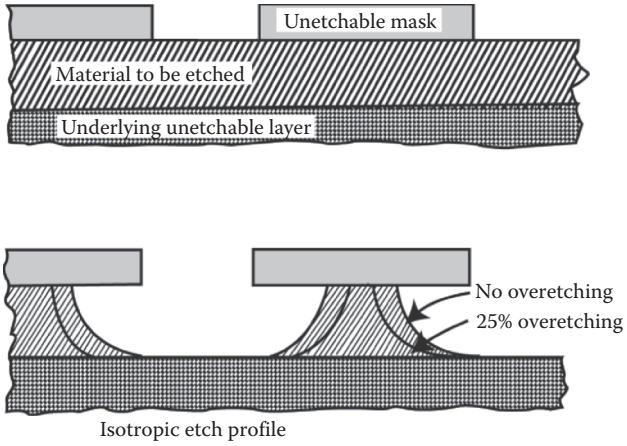


FIGURE 3.4 Example illustrating the loss of dimensionality incurred by an isotropic etch when the depth is of the order of the lateral dimensions involved. Making the initial lithography smaller compensates for the loss of dimensionality by overetching.

the CDs (see Figure 3.3c and d). In the case of isotropic etching, dimensionality losses are incurred with an etch depth the order of the lateral dimensions as illustrated in Figure 3.4. To ensure uniform etch rates over the whole wafer and between wafers, isotropic overetching may be required, further compromising the dimensionality. Figure 3.4 compares the etch profile of no overetching with a 25% overetch. Making the initial lithography smaller compensates for the loss of dimensionality incurred by overetching.

Even though isotropic etching leads to nonvertical sidewalls and loss of dimensionality, there remain some advantages to their use. Often a vertical sidewall is not the desired edge profile. For example, with line-of-sight deposition methods, such as resistive evaporation (see Chapter 7), a tapered sidewall is easier to cover than a vertical wall. Another example in which a nonvertical process has an advantage is illustrated in Figure 3.5. To take away the layer shown on the vertical walls, an anisotropic etchant would require extensive overetching, whereas an isotropic etchant will remove the material quickly.

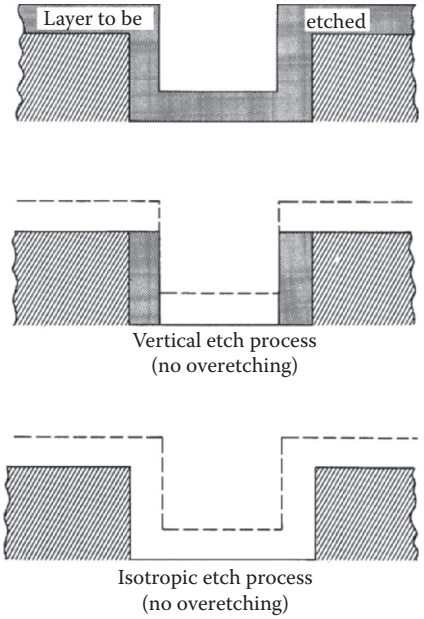


FIGURE 3.5 An example of dry etching illustrating a case where an isotropic etch is preferred to a vertical etch process. To clear the layer on the vertical walls, an isotropic etch only requires a minimal overetch, whereas an anisotropic etch requires a substantial amount of overetching.

Plasmas or Discharges

Introduction

Plasmas, defined as ionized gases that are neutral with equal numbers of ions and electrons, are commonly referred to as the fourth state of matter. A more precise definition is that plasma is a quasineutral gas of charged and neutral particles exhibiting collective behavior. The different states of matter generally found on earth are solid, liquid, and gas, and we have learned to work, play, and rest using these familiar states of matter. But in 1879, Sir William Crookes, an English physicist, identified a fourth state of matter, now called plasma. Plasmas in the stars and in the tenuous space between them make up more than 99% of the visible universe and perhaps most of which is not visible. Here on Earth, plasmas are important in film etching, sputter deposition, fluorescent lamps, welding arcs, and gas lasers (to name only a few).

Most dry etching systems also find their common base in plasmas or discharges, areas of high energy electric and magnetic fields that rapidly dissociate a suitable feed gas (e.g., sulfur hexafluoride, SF_6 for Si etching) to form neutrals, energetic ions (SF_5^+ and F^-), photons, electrons, and highly reactive radicals ($F^\bullet$). A surface in contact with plasma is exposed to fluxes of atoms, molecules, ions, electrons, and photons. This bombardment stimulates the production of outgoing fluxes of neutrals, ions, electrons, and photons, leading to a modified or etched surface layer.[3] We will start our foray into

the physics and chemistry of plasmas by looking at the simplest plasma setup—a DC-diode glow discharge in a low-pressure argon atmosphere. After a discussion of the spatial zones in a glow discharge, we explain the Paschen curve and then introduce radiofrequency (RF) plasmas.

Physics of DC Plasmas

The simplest plasma reactor consists of opposed parallel-plate electrodes in a chamber maintainable at low pressure, typically in the order of 1 mbar, as illustrated in the reactor shown in Figure 3.6a. The electrical potentials established in the reaction chamber, filled with an inert gas such as argon at a reduced pressure, determine the energy of ions and electrons striking the surfaces immersed in the discharge. To explain the generation of a DC plasma consider the setup shown in Figure 3.6a where we have simply put two electrodes (anode and cathode) at opposite ends of an evacuated glass tube and applied 1.5 kV between them. With the electrodes separated by 15 cm this results in a 100 V/cm field. Electrical breakdown (rapid ionization of a medium following the application of an overvoltage) of the argon gas in this reactor will occur when electrons, accelerated in the existing field, transfer an amount of kinetic energy greater than the argon ionization potential (i.e., 15.7 eV) to the argon neutrals. Elastic collisions deplete very little of the electron's energy and do not significantly influence the molecules because of the great mass difference between electrons and gas molecules. Inelastic collisions, on the other hand, excite the molecules of the gas or ionize them by completely removing an electron. Thus, such energetic inelastic collisions may generate a second free electron and a positive ion for each successful strike. Both free electrons re-energize, creating an avalanche of ions and electrons that results in a gas breakdown emitting a characteristic beautiful blue glow (in the case of argon; for air or nitrogen a pink color results from excited nitrogen molecules). The excitation/relaxation processes are responsible for the plasma glow. Avalanching requires the ionization of 10–20 gas molecules by one secondary electron. At the beginning of a sustained gas breakdown, a current, typically several hundred milliamps, starts to flow, and the voltage between the two electrodes decreases from 1.5 kV to about 150 V. Newly produced electrons are accelerated toward the anode, and with sufficient voltage, the gas rapidly becomes filled with positive and negative particles throughout its volume, i.e., it becomes ionized. The discharge current builds up to the point where the voltage decrease across a current-limiting resistor is equal to the difference between the supply voltage (1.5 kV) and the electrode potential difference (150 V). To sustain plasma, a mechanism to generate additional free electrons must exist after the plasma-generating electrons have been captured at the positively charged anode. Plasma-sustaining electrons are generated through the emission of secondary electrons (Auger electrons) when the

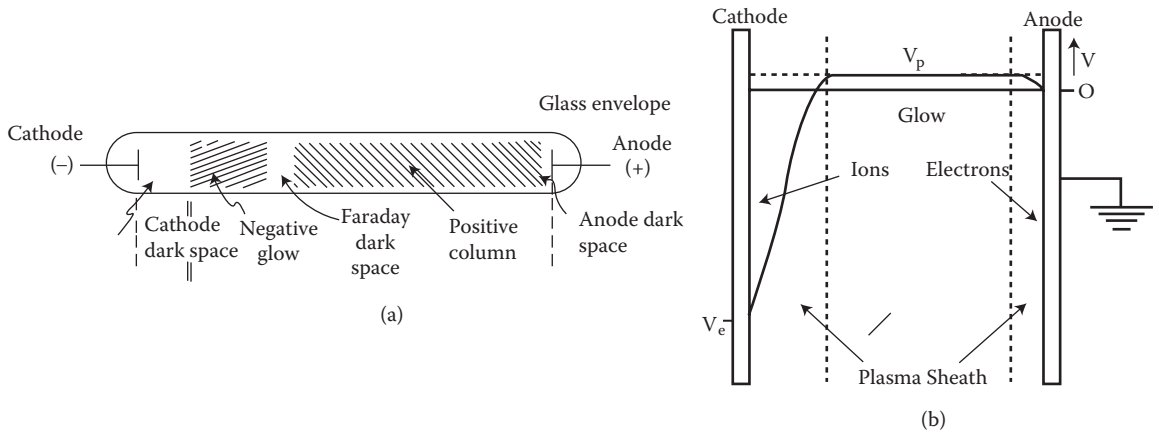


FIGURE 3.6 Voltage distribution in a DC diode discharge in equilibrium. (a) Structure of the glow discharge in a long tube-based DC diode system. (b) Voltage distribution in a real DC diode system where the anode is brought much closer to the cathode than in (a) and is positioned in the negative glow region (so-called obstructed glow).

cathode is struck by ions. The continuous generation of those "new" electrons prompts a sustained current and a stable plasma glow. The discharge reaches a steady state when electron generation and loss processes balance each other. Electrons are lost by 1) drift and diffusion to the chamber walls, 2) recombination with positive ions, and 3) attachment to neutral molecules to form negative ions. Once this equilibrium is reached, the glow region of the plasma, being a good electrical conductor, sustains a very low field only, and its potential is almost constant. Most of the space between the electrodes is now filled with a soft glow as a result of the relaxation of excited species; however, certain dark spaces are observed near the electrodes. Most of the potential drops reside at the electrode surfaces where electrical double layers are formed in "sheath fields," counteracting the loss of electrons from the plasma (Figure 3.6b). The most significant dark space is seen at the cathode. This dark space, called the "ion sheath," is a positive space charge region, which is key to most dry etching processes. The cathode sheath, the first few millimeters in front of the cathode, corresponds to the cathode dark space (also called *Crookes* or *Hittorf* region). Positive ions entering this sheath move slower toward the cathode than electrons move away from it, leaving this sheath with an excess of positive charges. Thus, if a conductor is placed on the cathode, it is bombarded with ions and may be etched. If an insulating material, such as a dielectric or an oxide-coated silicon wafer, is placed on the cathode (or the anode for that matter), the process stops. The dielectric rapidly charges up, and the field in the gas drops, causing the generation of electrons to cease. For this reason, DC plasmas are rarely useful for etching silicon wafers. The solution to this problem lies in the use of an AC power source of sufficiently high frequency to reverse polarity before the dielectric becomes fully charged. Typically, radiofrequency (RF) is used, although the minimum effective frequency is in the range of >50 kHz (see below under RF plasmas).

In Table 3.2 we summarize the cathode dark space characteristics.

The light distribution shown in Figure 3.6a is only observed when the interelectrode distance is large compared with the size of the electrodes. The "Faraday dark space" and the "positive column" are characterized by a low electric field to conduct the electrons toward the anode. These two regions are often present

Table 3.2 Cathode Dark Space Characteristics

Characteristics	Remark
Fewer electrons	Highly mobile electrons are accelerated quickly away from the negative cathode.
More positive ions	As electrons escape, they leave a high concentration of positive ions. These are attracted to the negative cathode.
Low conductivity	Positive ions move slowly. The result is that the field is higher across the cathode dark space.
Ions are accelerated	When the ions drift into the field of the dark space, they are accelerated toward the cathode.
Ions collide with the cathode and with neutral species on their way to the cathode	Therefore, the cathode is bombarded both with ions and with neutral species such as radicals and gas molecules. The ionic bombardment releases secondary electrons, which sustain the glow discharge. At low pressures and high voltages, the bombardment can cause sputtering.
Secondary electrons are created here	These electrons quickly accelerate to the ionization energy. As they leave the high field region and encounter ionizing collisions, they lose energy and then only have enough energy to excite. This is the edge of the negative glow region.
The dark space increases as the pressure is reduced (longer mean free path, higher energies, fewer collisions)	Eventually the dark space extends nearly across the chamber; the mean free path becomes too large to maintain a sufficient number of collisions and the glow is extinguished.
A conductor placed on the cathode	Will be subjected to ion bombardment.
An insulator placed on the cathode	The material charges up, and all the applied voltage exists across the material, and the plasma is extinguished.

Source: Based on a class by Dr. Lynn Fuller and Dr. Richard Lane, Microelectronic Engineering, Rochester Institute of Technology.

in glow discharges used as lasers ("positive column lasers") and as fluorescence lamps. However, for most of the other applications of glow discharges (e.g., sputtering, deposition, chemical etching, analytical chemistry), the distance between cathode and anode is very short, so that normally (i.e., at typical discharge conditions) only a short anode zone is present beside the cathode dark space and the negative glow region, where the slightly positive plasma potential returns back to zero at the anode (see Figure 3.6b). This situation is termed an *obstructed glow*. Virtually all processing plasmas used in industry operate with a short interelectrode distance (e.g., twice that of the cathode dark space width) using the negative glow as principal plasma.[4] For electrodes of 12 cm in diameter, the interelectrode distance will typically be less than 1 cm. In the negative glow region, the potential is nearly constant and slightly positive, and hence the electric field is very small. It is this type of discharge that we depict in Figure 3.6b, with only a cathode dark space, negative glow, and anode dark space. If the interelectrode distance becomes significantly more than twice the dark space thickness, the discharge is extinguished. The latter characteristic can be applied to prevent discharge from contacting areas in a reactor where it is not desired.[5]

A striking characteristic of plasma is its permanent positive charge with respect to the electrodes—a result of the random motion of electrons and ions. The positive charge of plasma can be understood from kinetic theory, which predicts that for a random velocity distribution, the flux of ions, j_i, and electrons, j_e, on a surface is given by:

$$j_{i,e} = \frac{n_{i,e} \; \langle v_{i,e} \rangle}{4} \tag{3.1}$$

where n and $\langle v \rangle$ are the densities and average velocities, respectively. Because ions are heavier than electrons (typically 4,000–100,000 times as heavy), the average velocity of electrons is larger. Consequently, the electron flux is larger than the ion flux, and the plasma loses electrons to the walls, thereby acquiring a positive charge. The bombarding energy of ions is proportional to the potential difference between the plasma and the surface being struck by the ions. The reason behind the asymmetric voltage distribution at the anode and cathode as seen in Figure 3.6b is as follows: electrons near the cathode rapidly accelerate away from it because of their relatively light mass; ions, being more massive, accelerate toward the cathode more slowly. Thus, on average, ions spend more time in the Crookes dark space, and, at any instant, their concentration is greater than that of electrons. The net effect is a very large field in front of the cathode compared with that in front of the anode and in the glow region itself. Consequently, the greatest part of the voltage between the anode and the cathode, V_e, is dropped across the Crookes dark space where charged particles (ions and electrons) experience their largest acceleration. This region is also called the *cathode-fall* region.

Glow discharges represent nonequilibrium conditions as the average electron energy, $\langle v_e \rangle$, also called *electron temperature*:

$$\langle v_e \rangle = kT_e \text{ (e.g., 1–10 eV)} \tag{3.2}$$

is considerably higher, in the range of 1–10 eV, than the average ion energy, $\langle v_i \rangle$, also called *ion temperature*:

$$\langle v_i \rangle = kT_i \text{ (e.g., 0.04 eV)} \tag{3.3}$$

which is much colder, typically having energies of 0.02–0.1 eV. Thus, a discharge or plasma cannot be described adequately by one single temperature but rather by the ratio of electron and ion temperature [T_e (electrons)/T_i (ions) = 10–100]. High-temperature electrons in a low-temperature gas are possible because of the small mass of electrons (m_e) compared with ions and neutrals. In collisions between electrons and argon atoms (M_A), as a result of an m_e/M_A ratio of 1.3×10^{-5}, electrons have a very poor energy transfer and stay warmer longer than the heavier ions and neutrals. In plasma with an ionization fraction of 10^{-4}, there are many more cold gas atoms than ions, and the ions quickly reach the background gas temperature as a result of the many collisions between ions and neutrals.[4] Electrons can attain a high average energy, often many electronvolts (equivalent to tens of thousands of degrees above the gas temperature), permitting electron-molecule collisions to excite high-temperature-type reactions, forming free radicals in the low-temperature neutral gas. Generating the

same reactive species without plasma would require temperatures in the ~10^3–10^4 K range, destroying resists and damaging most inorganic films.

Plasma is typically weakly ionized: the number of ions is small compared with the number of reactive neutrals such as radicals. The ratio between ionized and neutral gas species in a glow discharge plasma is in the order of 10^{-6}–10^{-4}. This fact, as we will see below, is crucial in understanding which entities are responsible for the etching of a substrate placed in a glow discharge. There are 1 million times more radicals than ions or electrons because radicals form more easily and their lifetime is much longer. We will see that ions do not etch, but radicals do. Ions affect the process by energetic (physical) bombarding of the surface, influencing the chemical processes of etching. The degree of ionization in plasma depends on a balance between the rate of ionization and the rate at which particles are lost by volume recombination and by losses to the walls of the apparatus. Wall losses generally dominate volume recombination. Accordingly, the occurrence of a breakdown in a given apparatus depends on the gas pressure (particle density), the type of gas, electric field strength (electron velocity), surface-to-volume ratio of the plasma, and distance between the anode and the cathode (see also Paschen's law in the section below). In air at atmospheric pressure (760 Torr) the mean free path is very short, and it is difficult for electrons to acquire enough energy to ionize gases. In such case, only extremely high electric fields may create a plasma in the form of arcing (as in lightning; see Figure 3.7) instead of a steady glow discharge.

FIGURE 3.7 Discharge in lightning.

The choice of a dry etching technique depends on the efficiency or "strength" of the particular plasma. Plasma strength is evaluated by parameters such as average electron energy, $\langle v_e \rangle$ (Equation 3.2), average ion energy (Equation 3.3), electron density (e.g., between 10^9 and 10^{12} cm^{-3}), plasma ion density (e.g., 10^8–10^{12} cm^{-3}), neutral species density (e.g., 10^{15}–10^{16} cm^{-3}), and ion current density (e.g., 1–10 mA/cm^2). A quantity of particular use in characterizing plasma's average electron or ion energy is the ratio of the electrical field to the pressure:

$$kT_{i,e} \sim \frac{E}{p} \tag{3.4}$$

With increasing field strength, the velocity of free electrons or ions increases because of acceleration by the field (~E), but velocity is lost by inelastic collisions. Because an increase in pressure decreases the electron or ion mean free path, resulting in more collisions, the electron or ion energy decreases with increasing pressure (~1/P).

To use the described DC plasmas for 1) dry etching and 2) sputter deposition, one uses one of the following setups:

1. In the etching arrangement, the substrate to be etched is placed on the cathode of the DC diode system. An argon plasma in this setup produces sufficiently energetic ions (between 200 and 1000 eV) to induce physical etching, that is, ion etching or sputtering of the material on the cathode.
2. The same setup is used for sputter deposition where atoms from the target (cathode) are ejected like a billiard ball and deposited on the substrate laying on the anode (see Chapter 7).

The low pressure in sputter etching (10–200 mTorr) gives rise to a long mean free path, and the externally applied voltage concentrates across the cathode plasma sheath where ions are accelerated before hitting the cathode. Each ion will collide numerous times with other gas species before transversing the plasma sheath, as the sheath thickness is larger than the mean free ion path. As a result of these collisions, ions lose much of their energy and move across the sheath with a drift velocity that is

less than "free fall" velocity. But these vertical velocities are still very large compared with the random thermal velocities in the glow discharge region. The resulting bombarding ion flux, j_i, is given by:

$$j_i = qn_i\mu_i E \qquad (3.5)$$

where E = electric field
n_i = ion density
μ_i = ion mobility
q = charge

Reducing the pressure in the reactor increases the mean free path because particles have more of a chance to accelerate before colliding with another particle or wall. Consequently, ions accelerating toward the cathode at lower pressures can gain more energy before a collision takes place. A typical sputtering profile was illustrated earlier in the schematic in Figure 3.3a. In the opposite high pressure mode, reactive species such as radicals generated by the DC plasma may combine with the substrate to form volatile products, chemically etching the substrate in so-called *plasma etching*. In this case, ions striking the surface have lower energy (e.g., less than 10–20 eV) as a result of the lower sheath voltages and short mean free path because of the relatively high pressures (typically >0.1 Torr). Aside from generating the radicals through energetic collisions in the plasma, ions contribute very little or nothing to the etching process itself. The fast etching radicals reach the etch site via diffusion from the plasma rather than by electromigration. A typical plasma etching profile was shown in Figure 3.3b.

Spatial Zones in a DC Glow Discharge

Introducing Figure 3.6a we noticed that the structure of plasma discharges in tubes that are long compared to the electrode sizes is complex. When a DC glow discharge is sustained in a long, narrow cylindrical tube filled with a rare gas at low pressure, the visible light emitted from the glow is distributed as shown in Figure 3.6a and, in more detail, in Figures 3.8 and 3.9. In Figure 3.9 we also show the potential distribution and the ion and electron density along the length of such a discharge tube (a DC "gas diode"). Regions described as "glows" emit

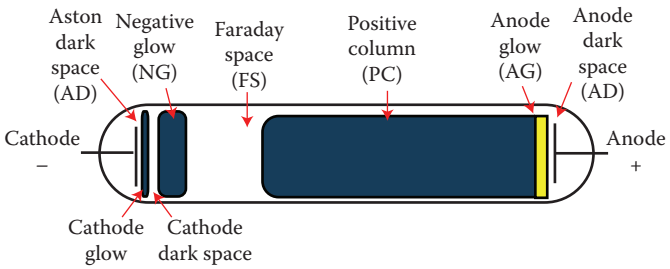


FIGURE 3.8 Detailed structure of a glow discharge in a DC gas diode.

significant light; regions labeled as "dark spaces" do not. Photons are created by the excitation/relaxation of the gas molecules, and when excitation/relaxation is the dominant process the plasma glows. The glow discharge can be subdivided into several regions between the anode and the cathode. Besides the Crookes or Hittorf dark space (cathode DS), the Aston DS, the Faraday DS, and the anode DS, a cathode glow can be observed, a negative glow (NG) region, the positive column (PC), and the anode glow (AG). The "dark" regions are not really dark, just less luminous than the bright regions. As the discharge becomes more extended (i.e., stretched horizontally in the geometry of the figure), the positive column may become striated, and for monatomic gases, the cathode glow may become striated as well. That is,

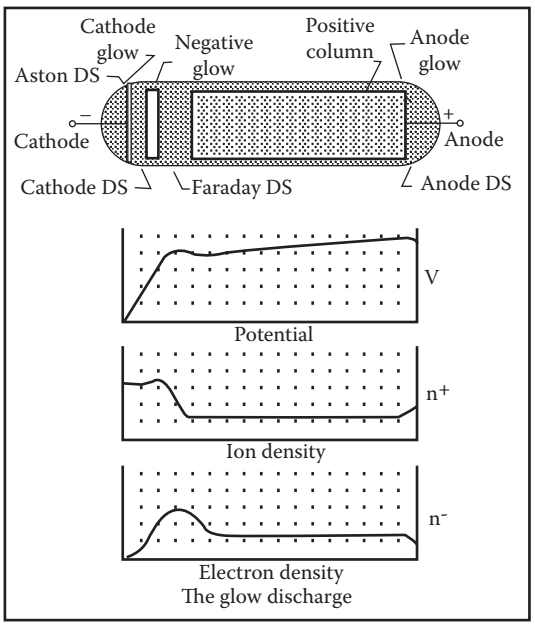


FIGURE 3.9 Detailed structure of a glow discharge in a DC diode with potential distribution and the ion and electron density along the length of the discharge tube (see also Figure 3.8).

additional alternating dark and bright regions may form. Alternatively, compressing the discharge horizontally results in fewer regions. The positive column will be compressed, whereas the negative glow will remain the same size, and, with small enough gaps, the positive column will disappear altogether.

The cathode dark space has the strongest electric field in the discharge. This region is filled with a positive space charge of ions moving in the opposite direction of the electrons. Electrons leaving the cathode with a few electronvolts of energy do not have the capacity of excitation/relaxation, resulting in a very thin first dark layer at the cathode (Aston dark space-AD). However, electrons emitted from the cathode rapidly attain energy suitable for excitation/relaxation, giving rise to a small bright layer; that is, the first cathode layer or cathode glow. After this thin excitation layer, the electrons are exhausted, and a second dark layer arises. Electrons then re-energize in the field and may form a second thin cathode layer, vaguer than the first, and so on. In Figures 3.8 and 3.9 we show one Aston dark space and one cathode glow layer only. If there is only one cathode glow, the second Aston dark space corresponds to the Crookes, Hittorf, or cathode dark space. A plurality of Aston dark spaces and thin cathode glow layers are only clearly observable in monoatomic gases; in polyatomic gases other energy dissipation processes take over.[6] In the main part of the cathode DS (Crookes or Hittorf dark space), electrons have gained too much energy from the electric field for efficient excitation, and luminescence becomes very weak.

The negative glow is the bright segment adjacent to the cathode dark space. In this brightest region of the glow discharge, there are many more electrons (a result of charge multiplication), and they do not gain any further energy, as this region is equipotential or field free.

The electron energy in the negative glow region is much more suitable for excitation/relaxation, and a bright light with a color characteristic for the discharge gas (blue in the case of argon) results. The electrons are exhausted by this effort and drift out into the Faraday dark space and the positive column at low velocity.

If the distance between anode and cathode is sufficiently large, a long positive column is present. There are now much fewer ions around, so the electrons can be accelerated in the gentle field. As they speed up, their number density goes down (the main losses are through diffusion to the walls). When they have sufficient energy to excite the molecules of the gas, the emitted light becomes evident. This light is produced with minimal excitation and consists only of light emitted by neutral atoms that have been excited above their resonance energy, which explains its different character from the light emitted by the negative glow. The positive column is less bright than the negative glow, and, although of another color, it is also characteristic of the discharge gas (purple in the case of argon). This region is of no importance in thin film processing technology but, as mentioned before, is very important for plasma discharges used as light sources (e.g., a fluorescent Hg-Ar discharge lamp) and in lasers.[4]

Close to the anode, electrons are attracted and accelerated, and positive ions are repelled. The accelerating electrons excite gas molecules, and an anode glow results.

The described discharge regions vary in radiation intensity, potential and electrical field distribution, space charge, and current density, and their occurrence and position depend on the discharge parameters such as pressure, voltage, current, kind of gas, and, as we have seen, the distance between cathode and anode. We refer to work by Coe et al.,[7,8] Raizer et al.,[9] and Bogaerts et al.[10–12] for a more in-depth study of the various phenomena involved in low-pressure noble gas discharges.

Paschen's Law

The preceding section on DC plasmas introduces the concept of avalanche and gas breakdown in a DC diode system. To enhance our understanding of plasmas in general, and specifically to understand better what happens when the electrode distance or pressure in the discharge chamber is changed, we now introduce Paschen's law.

When the electric field between a pair of metal electrodes in air (at room temperature and atmospheric pressure) increases to more than, roughly, 3 MV m^{-1}, *electrical breakdown* occurs. Analogous to the process described above for the argon-filled chamber,

free electrons accelerate in the electric field, and if they pick up enough energy before hitting air molecules, ionization results and more ions and electrons are created. This can lead to yet further ionization or an avalanche destroying the insulating properties of the air between the electrodes and leading to the passage of a high current, accompanied by the emission of heat, light, and sound—that is, a spark discharge. A spark discharge in a localized region (e.g., near a sharp point) is called a *corona* discharge.

The voltage for electrical breakdown depends on the geometry of the electrode gap, the gas, and the pressure. In 1889, Paschen published a paper describing a generalized form for that relationship, and this has become known as *Paschen's law.*[13] The law essentially states that the breakdown characteristics of an electrode gap are a function of the product of the gas pressure and the gap distance, usually written as V = f (P × d), where the breakdown voltage V is a function of the pressure P and the gap distance d. As the pressure or gap is reduced, the breakdown voltage decreases slowly to a minimum and then rises steeply, as illustrated in Figure 3.10, representing the Paschen's law curve in air and other gases. The sharp increase in the breakdown voltage on the left side in Figure 3.10 (low $P \times d$ side) occurs because the electrode spacing is too small for ionization to occur. The number of gas molecules in the space between the electrodes is proportional to $P \times d$. When the pressure is too low or the distance too small, most electrons reach the anode without colliding with gas molecules. As a consequence, the lower the pressure or the shorter the d, the higher the value of V required to give electrons enough energy to cause breakdown of the gas. For air, the minimum breakdown voltage (V_{min}) is about 300 V at ~8-μm gap distance (see also Table 3.3). On the right side of this plot (high $P \times d$ side), the slow increase in breakdown voltage with increasing electrode distances and higher pressures arises because electron collisions with gas atoms become more and more frequent so that for the electrons to gather sufficient energy to overcome the ionization potential of the inert gas, more potential needs to be applied.

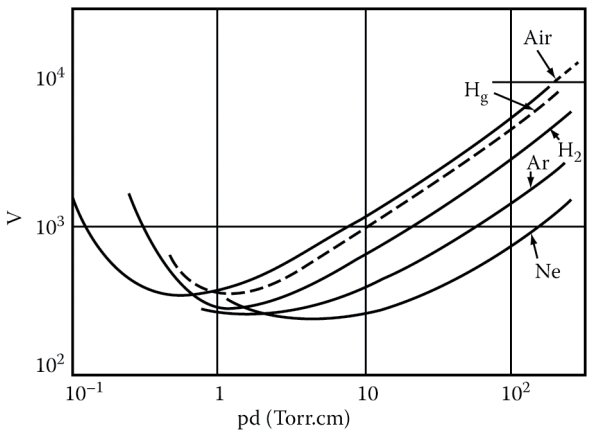


FIGURE 3.10 A DC breakdown voltage as a function of gas pressure P and electrode spacing d for plane parallel electrodes in air and some other gases. Such curves are determined experimentally and are known as *Paschen curves*.

TABLE 3.3 Minimum Breakdown Voltage (V_{min}) and Pressure/Distance Product (P × d) at That Minimum Voltage for Various Gases as Deduced from Their Paschen Law Curves

Gas	V_{min} in Volts (V)	P × d at V_{min} (Torr cm)
Air	327	0.567*
Ar	137	0.9
H_2	273	1.15
He	156	4.0
CO_2	420	0.51
N_2	251	0.67
N_2O	418	0.5
O_2	450	0.7
SO_2	457	0.33
H_2S	414	0.6

*At 1 atm = 760 mmHg or 760 Torr, d = 7.460 10^{-4} cm or ~8 μm.
Source: Naidu, M. S., and V. Kamaraju. 1995. *High voltage engineering*. New York: McGraw-Hill.[14]

Some microdevices such as electrostatic micromotors, microionizers for ion mobility spectrometers, and microswitches operate in air without sparking, as they operate on the left side of the Paschen curve (see Volume III, Chapter 8). Our intuition would tend to predict that the smaller the gap distance, the smaller the breakdown voltage. This typifies how linear scaling can be misleading when predicting the behavior of microdevices. Microstructures, with dimensions of a few micrometers, operate in air as if surrounded by a reduced pressure environment.

Different gases exhibit similar "Paschen" behavior, with the curves more or less shifted from the air curve, depending on the mean free path of the gas molecules involved (see Figure 3.10). Paschen's law—V = f(P × d)—should be more correctly stated

as V = f(δ × d), where δ is the density of gas molecules, which is, of course, affected by the temperature and the pressure of the gas. Much research has been done to provide a theoretical basis for Paschen's law and to develop a greater understanding of the breakdown mechanism. This has proven to be a formidable task, as many factors have an effect on the breakdown in a gap, such as radiation, dust, and surface irregularities. The electrode material also affects the sparking voltage, with cathodes of barium and magnesium having higher voltages than aluminum, for example.

In Table 3.3 we list the minimum breakdown voltage (V_{min}) and the product of the pressure and electrode spacing ($P \times d$) at that such voltage for various common gases as deduced from Paschen's law curves.[14]

A good text and reference book on electrical breakdown of gases, although old, is written by Meek and Craggs.[15]

Physics of RF Plasma Reactors

AC voltages overcome the problem of charge accumulation on a dielectric in a DC plasma system. The positive charge that accumulates as a result of ion bombardment during one-half of the AC cycle can be neutralized by electron bombardment during the next half cycle. The frequency of AC must be high enough so the half period will be shorter than the charge-up time of the dielectric. Although this time will vary because of conditions and dielectric materials, for most applications the frequency must be greater than 50 kHz. At low frequencies, both electrons and ions can follow the changing field, so that a glow discharge is the same as DC, except that the polarity reverses twice each cycle. At high frequency, the massive ions cannot respond to the frequency changes, whereas electrons can. By far the most common RF used is 13.56 MHz, allowed by the Federal Communications Commission (because of its noninterference with radio-transmitted signals). The RF power supply is rated between 1 and 2 kW. In RF-generated plasma, a radiofrequency voltage applied between the two electrodes causes free electrons to oscillate and collide with gas molecules, leading to a sustainable plasma. The RF-excited discharges can be sustained without relying on the emission of secondary electrons from the cathode. Electrons pick up enough energy during oscillation in an RF field to cause ionization, thus sustaining the plasma at lower pressures than in DC plasma (e.g., 10 vs. 40 mTorr). The RF breakdown voltage of an AC plasma exhibits the Paschen behavior of a DC plasma; that is, a minimum in the required voltage as a function of the pressure with the mean free path of electrons, λ_e, substituting the spacing, d, between electrodes.

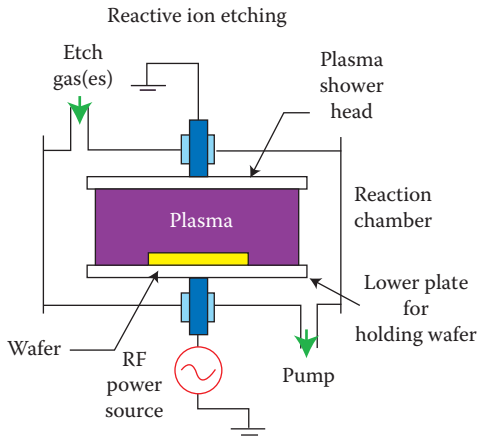


FIGURE 3.11 Two-electrode setup (diode) for RF ion sputtering or sputter deposition. For ion sputtering, the substrates are located on the cathode (target); for sputter deposition (see Chapter 7), the substrates to be coated are located on the anode.

In the simplest case of RF ion sputtering, the substrates to be etched are laid on the cathode (target) of a discharge reactor, for example, a planar parallel-plate reactor as shown in Figure 3.11. The reactor consists of a grounded anode and powered cathode or target enclosed in a low-pressure gas atmosphere (e.g., 10^{-1}–10^{-2} Torr of argon). The RF plasma, formed at low gas pressures, consists of positive cations, negative anions, radicals, and photons (the UV photons create the familiar plasma glow). As with DC plasma, neutral species greatly outnumber electrons and ions—the degree of ionization only being on the order of 10^{-4}–10^{-6} for parallel-plate gas discharges. With one of the two electrodes capacitively coupled to the RF generator, this electrode automatically develops a negative DC bias and becomes the cathode with respect to the other electrode (see Figure 3.12a). This DC bias (also called *self-bias* V_{DC}) is induced by the plasma itself and is established as follows. When initiating an AC plasma arc, electrons, being more mobile than ions, charge up the

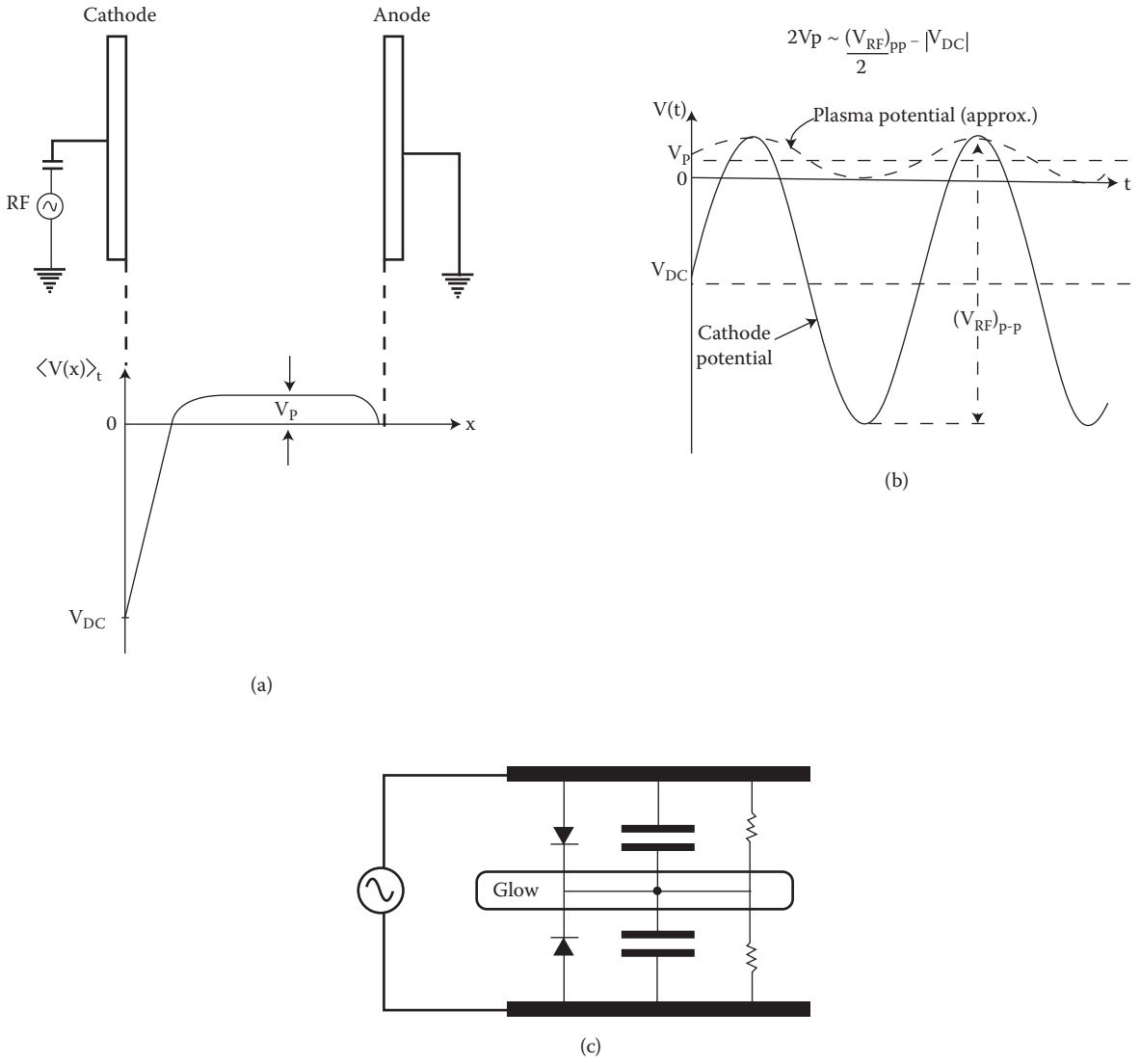


FIGURE 3.12 An RF plasma. (a) Approximate time-averaged potential distribution for a capacitively coupled planar RF discharge system. (b) Potential distribution in glow discharge reactors—V_p: plasma potential; V_{DC}: self-bias of cathode electrode; $(V_{RF})_{pp}$: peak-to-peak RF voltage applied to the cathode. (c) Equivalent electrical circuit of an RF plasma.

capacitively coupled electrode; because no charge can be transferred over the capacitor, the electrode surface retains a negative DC bias.

The energy of charged particles bombarding the surface in a glow discharge is determined by three different potentials established in the reaction chamber: the plasma potential V_p, that is, the potential of the glow region, the self-bias V_{DC}, and the bias on the capacitively coupled electrode $(V_{RF})_{pp}$ (see also Figure 3.12b). The following analysis clarifies how these potentials relate to one another and how they contribute to dry etching.

The self-bias buildup, V_{DC}, on an insulating electrode by the electron flux (Equation 3.1) is given by:

$$V_{DC} = \frac{kT_e}{2e} \ln \frac{T_e m_i}{T_i m_e} \tag{3.6}$$

where T_e and T_i are the electron and ion temperatures defined by Equations 3.2 and 3.3, and m_e and m_i are the electron and ion masses, respectively.[16] In the plasma, there are equal amounts of hot electrons ($T_e = 2 \times 10^4$ K) and cold positive ions ($T_i = T_{gas} = 500$ K). For a common RF power density of 4 kWm^{-2} the self-bias is about 300 V, so that the powered electrode (the cathode) is bombarded with 300 eV of positive ions.[17]

The electron loss creates an electric sheath field in front of any surface immersed in the plasma,

counteracting further electron losses. This sheath or dark space forms a narrow region between the conductive glow region and the cathode (the Crookes dark space) where most of the voltage (V_{DC} in Figure 3.12a) is dropped as in the DC plasma case. The cathode, being capacitively coupled, effectively acts as an insulator for DC currents. The thickness of the sheath is typically 0.01–1 cm, depending on pressure, power, and frequency. The other electrode is grounded and conductive (no charge buildup and no voltage buildup) and automatically becomes the anode with respect to the capacitively coupled electrode.

The time-average of plasma potential V_p, DC cathode potential (self-bias potential) V_{DC}, and peak-to-peak RF voltage $(V_{RF})_{pp}$ applied to the cathode are approximately related as (see Figure 3.12b):

$$2V_p \sim \frac{(V_{RF})_{pp}}{2} - |V_{DC}| \tag{3.7}$$

Clearly, the magnitude of the self-bias depends on the amplitude of the RF signal applied to the electrodes. In the RF discharge, the time-averaged RF potential of the glow region, referred to as the *plasma potential* V_p, is more significantly positive with respect to the grounded electrode than in the case of a DC plasma.

Positive argon ions from the plasma are extracted by the large field at the cathode and sputter that electrode at near-normal incidence, with energies ranging from a few to several hundred electronvolts, depending on plasma conditions and chamber construction. One of the most important parameters determining plasma condition is total reactor pressure. As pressure is decreased to less than 0.05–0.1 Torr, the total ion energy, E_{max}, increases as both the self-bias voltage and the mean free path of the bombarding ions increase. Subsequently, ion-substrate bombardment energy increases sharply with decreasing pressure. The maximum energy of positive ions striking a substrate placed on the cathode is proportional to:

$$E_{max} = e(|V_{DC}| + V_p) = eV_T \tag{3.8}$$

with $V_T = |V_{DC}| + V_p$, whereas the maximum energy for a substrate on the grounded electrode (i.e., the anode) is proportional to:

$$E_{max} = eV_p \tag{3.9}$$

Typical values are 300 eV for the cathode and less than 20 eV for the anode.[17] Equations 3.8 and 3.9 can be deduced from an inspection of Figure 3.12a. The situation where the wafers are placed on the cathode is referred to as *reactive ion etching* or *reactive sputter etching*. Chamber construction, especially the ratio of anode to cathode area, influences the rates of V_T/V_p and, consequently, as calculated from Equations 3.8 and 3.9, the energy of the sputtering ions on these respective electrodes. Etching of the anode should be avoided, that is, V_p should be kept small (e.g., <20 eV). To deduce the influence of the geometry of the anode and cathode on dry etching, we compare the sheath voltages in front of the anode and cathode using the Child-Langmuir equation, expressing the relationship between the ion-current flux, j_i, the voltage drop, V, over the sheath thickness of the dark space, d, and the mass of the current-carrying ions, m_i. The relation can be deduced from Equation 3.5, assuming the presence of a space-charge limited current:

$$j_i = \frac{KV^{\frac{3}{2}}}{\sqrt{m_i}\,d^2} \tag{3.10}$$

in which K is a constant.[18] The current density of the positive ions must be equal on both the *anode* $j_i(P)$ and cathode $j_i(T)$; that is, $j_i = j_i(P) = j_i(T)$, resulting in the following relation for the sheath voltages:

$$(\text{cathode})\frac{V_T^{\frac{3}{2}}}{d_T^2} = (\text{anode})\frac{V_p^{\frac{3}{2}}}{d_p^2} \tag{3.11}$$

The plasma behaves electrically as a diode (large blocking voltage drop toward the capacitively coupled cathode and small voltage drop on the anode/plasma interface), in parallel with the sheath capacitance. As soon as any electrode tends to become positive relative to the plasma, the current increases dramatically, causing the plasma to behave as if a diode were present in the equivalent electrical circuit of the plasma. Hence a planar, parallel setup also is called a *diode* setup. The equivalent electrical circuit representing an RF plasma is represented in Figure 3.12c. The dark spaces in plasma are areas

of limited conductivity and can be modeled as capacitors:

$$C \sim A/d \tag{3.12}$$

The plasma potential is determined by the relative magnitudes of the sheath capacitances, which, in turn, depend on the relative areas of anode and cathode. An RF voltage will split between two capacitances in series according to:

$$\frac{V_T}{V_p} = \frac{C_p}{C_T} \tag{3.13}$$

Using Equations 3.12 and 3.13, we can write:

$$\frac{V_T}{V_p} = \left(\frac{A_p}{d_p}\right)\left(\frac{d_T}{A_T}\right) \tag{3.14}$$

and substituting into Equation 3.11, we obtain:

$$\frac{V_T}{V_p} = \left(\frac{A_p}{A_T}\right)^4 = R^4 \tag{3.15}$$

where A_p is the anode area and A_T the cathode area. If there were two symmetric electrodes, both blocked capacitively, sputtering would occur on both surfaces. If the area of the cathode were significantly smaller than the other areas in contact with the discharge, the plasma potential would be small, and little sputtering would occur on the anode, whereas the cathode would sputter very effectively. Because the cathode in a setup such as that represented in Figure 3.11 usually is large (>1 m^2), allowing many silicon wafers or other substrates to be etched simultaneously, the grounded area needs to be larger yet. In sputtering systems, the anode and the entire sputtering chamber are grounded, creating a very small anode dark space where hardly any sputtering takes place. Other researchers have assumed conservation of the positive ion current; that is, $j_i(P) \times A_p = j_i(T) \times A_T$ instead of current density, and they find an exponent 2 in deriving Equation 3.15. The exponential in Equation 3.15 in practical systems is usually closer to 2.[5]

In a setup where the electrode dimensions are large compared with the electrode distance, the anode and cathode areas are considered equal, and little difference is made between V_T and V_p.[5]

Higher ion energies (V_T large) translate into lower etch selectivity and can be a cause of device damage (Table 3.4). Consequently, a key feature for good etch performance (Table 3.5) is effective ionization to produce very high quantities of low-energy ions and radicals at low pressures. Selectivity as high as 40:1 might be required in the future for ICs and even more for micromachines. It is generally believed that "radiation" damage of plasma can be minimized by keeping ion energies low. It is also generally believed that a lot of the radiation damage can be annealed out. In reality, very little is understood of the damage plasma can do. The higher the etch rate, the better the wafer throughput. Good selectivity, uniformity, and profile control are more easily achieved at lower etch rates. Trenches of various depths are made in Si in the manufacture of metal oxide semiconductor (MOS) devices. Shallow

Table 3.4 Device Damage

Some "flavors" of device damage:
• Alkali (sodium) and heavy metal contamination
• Catastrophic dielectric breakdown
• Current-induced oxide aging
• Particulate contamination
• UV damage
• Temperature excursions that can activate metallurgical reactions
• "Rogue" stripping processes that simply do not remove all the residue
• Plasma-induced charges, surface damage, ion implantation

Table 3.5 Etch Performance*

Etch performance is judged in terms of the following:
1. Etch rate: Rate at which material is removed from the substrate
2. Uniformity: Etch rate constancy across the substrate
3. Throughput: Amount of substrates/wafers etched during one process cycle
4. Directional control: Control of the horizontal (CD) and vertical etch rate (profile)—isotropic or anisotropic
5. Surface quality (residue, microloading effects)
6. Device damage
7. Particle control
8. Reproducibility
9. Selectivity of etch:
a. Controlled by etchant composition
b. Controlled by etch rate

*See also Introduction to Part II.

trenches (0.5–1.5 μm) are used to aid in reducing the effects of lateral diffusion during processing and to make flat structures (planarization). Deeper trenches (1.5–10 μm) are used to create structures that become capacitors and isolation regions in integrated circuits. For the shallowest trenches, photoresist is adequate as a mask. For deeper etching, a silicon dioxide mask may be needed. In micromachining, one would like to obtain aspect ratios of 100 and beyond; thus, the difficulty in masking keeps mounting. With aspect ratios above 2:1 or 2.5:1, *the etching action* at the bottom of a trench tends to slow down or stop altogether.

In answer to these identified needs, equipment builders have come up with low-energy, high-density plasmas, for example, magnetrons, inductively coupled plasmas (ICPs),[19] and electron cyclotron resonances (ECRs)[20] (see below Figure 3.36).

Physical Etching: Ion Etching or Sputtering and Ion-Beam Milling

Introduction

Bombarding a surface with inert ions (e.g., argon ions), in a setup as shown in Figure 3.11, translates into ion etching or sputter etching. With ions of sufficient energy impinging vertically on a surface, momentum transfer (sputtering) causes bond breakage and ballistic material ejection, throwing the bombarded material across the reactor to deposit on an opposing collecting surface, provided the surrounding pressure is low enough. The kinetic energy of the incoming particles largely dictates which events are most likely to take place at the bombarded surface, that is, physisorption, surface damage, substrate heating, reflection, sputtering, or ion implantation. At energies less than 3–5 eV, incoming particles are either reflected or physisorbed. At energies between 4 and 10 eV, surface migration and surface damage results. At energies >10 eV (e.g., from 5–5000 eV), substrate heating, surface damage, and material ejection, that is, sputtering or ion etching, take place. At yet higher energies, >10,000 eV, ion implantation, that is, doping, occurs. The energy requirements for these various processes are summarized in Table 3.6 (see also Figure 3.15).

Table 3.6 Energy Requirements Associated with Various Physical Processes

Ion Energy (eV)	Reaction
<3	Physical adsorption
4–10	Some surface sputtering
10–5000	Sputtering
10–20 K	Implantation

The deposition phenomenon mentioned above can be used to deposit materials in a process called *sputter deposition* (Chapter 7). A low pressure resulting in a long mean free path is required for material to leave the vicinity of the sputtered surface without being backscattered and redeposited.

In what follows, we shall consider two purely physical dry etching techniques at energies >10 eV, that is, sputtering or ion etching and ion-beam milling.

Sputtering or Ion Etching

The gradient in the potential distribution around the target lying on the cathode (see Figure 3.11) accelerates the ions, prompting them to impinge on the substrate in a direction normal to the surface. The etch rate in the direction of the impinging ions (V_z) becomes a strong function of E_{max} (Equations 3.8 and 3.20). The impinging ions erode or sputter etch the surface by momentum transfer. This offers some advantages: volatility of the etch products is not as critical as for dry chemical etching, and only a billiard ball effect plays a role in physical etching. As a consequence, differences in etch rates for various materials are not large (sputter yields for most are within a factor of three), and the method entails directional anisotropy. Physical etching is inherently nonselective because the ion energy required to eject material is large compared with differences in chemical bond energy and chemical reactivity. When no reactive etching processes are available for a given material, physical etching is always an option. Directional anisotropy remains as long as the dimensions of the surface topographical structures are small compared with the thickness of the sheath between the bulk plasma and the etched surface. However, etch rates are slow, typically a hundred to a few hundred angstroms per minute. The use of a magnetron can help improve the etch rate

and minimize heating of substrates. In both DC and RF diode sputtering, most electrons do not cause ionization events with Ar atoms. They end up being collected by the anode, substrates, and so on, where they cause unwanted heating. A magnetron adjusts this situation by confining the electrons with magnetic fields near the target surface; consequently, current densities at the target may increase from 1 mA/cm^2 to 100 mA/cm^2.

As sputter etching is nonselective, it introduces a masking problem. Another hurdle is the need for rather high gas pressure to obtain sufficiently large ion currents, resulting in a short mean free path, λ_i, of the ions. With a mean free path smaller than the interelectrode spacing, considerable redeposition of sputtered atoms on the etching substrate lying on the cathode can occur. Electrical damage from ion bombardment also must be taken into account when critical electronic components reside on the substrate.

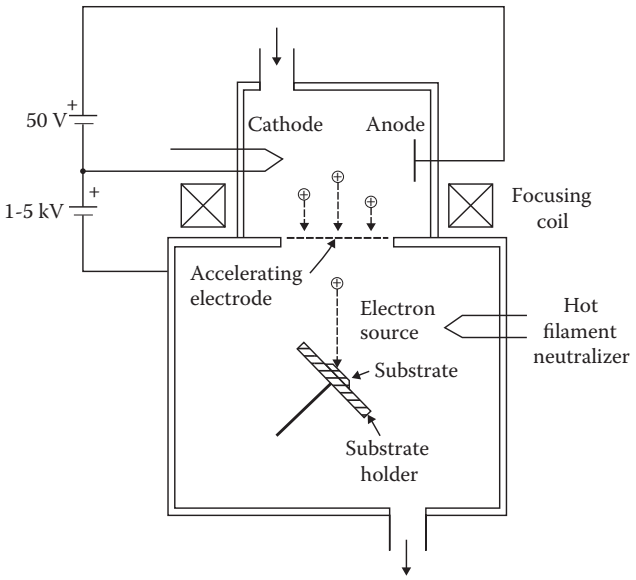


FIGURE 3.13 Ion-beam etching (IBE) apparatus (triode). In an ion-beam apparatus, the beam diameter is approximately 8 cm. Substrates are mounted on a moveable holder allowing etching of large substrates. Coils focus the ion beam and densify the ion flux in magnetically enhanced ion etching (MIE).

Ion-Beam Etching or Ion-Beam Milling

Ion-beam milling (IBM) is a process wherein the plasma source for ion etching is decoupled from the substrate, which is placed on a third electrode. Decoupled plasma is also called *remote* plasma, and the equipment needed is called a *triode* setup. The plasma source can be an RF discharge or a DC source (Penning source). In a typical DC triode setup, as shown in Figure 3.13, control of the energy and flux of ions to the substrate in the substrate chamber happens independently. The sample being etched can be rendered neutral by extracting low-energy electrons from an auxiliary thermionic cathode (i.e., a hot filament neutralizer), thus making this DC equipment usable for sputtering insulators, as well as conductors. In ion-beam milling, as in ion etching, noble gases are generally used as they exhibit higher sputtering yields as a result of heavy ions and avoid chemical reactions. The argon pressure in the upper portion of the chamber can be low (10^{-4} Torr), resulting in a large mean free ion path. Adding a hot wire (Kaufman source) or a hollow cathode can enhance ionization at those low pressures. In a Kaufman source, electrons are emitted by a hot filament (typically tantalum or tungsten) and accelerated by a potential difference between the cathode filament and an anode. A hollow cathode is a metal cylinder with a hole several millimeters in diameter. When the pressure is decreased, the hollow cathode starts to emit extremely strong light (negative glow) when the cathode sheath thickness reaches about the same dimensions as the diameter of the hole. The hollow cathode bounces electrons back and forth within the negative glow, thereby enhancing the plasma density considerably beyond that of a normal discharge.[5]

The discharge voltage in IBE must be larger than the gas ionization potential (15.7 eV for argon) and typically is operated at several times this value, about 40–50 eV, to establish a glow discharge. Ions are extracted from the upper chamber by sieve-like electrodes (extraction grids), formed into a beam, accelerated, and fired into the lower chamber where they strike the substrate. Achievable ion energies range from about 30 eV to several keVs. Beam divergence through collisions with residual gas molecules in the substrate chamber usually limits the lowest practical ion energy to a few hundred electronvolts. Typical etch rates with argon ions of 1 keV of energy and an ion current density of 1.0 mA/cm^2 are in the range of 100–3000 Å/min for most materials such as silicon,

polysilicon, oxides, nitrides, photoresists, and metals. Inert ion-beam etching (IBE) in an ion miller is, in principle, capable of very high resolution (<100 Å), but aspect ratios are usually less than or equal to unity, and selectivity is poor. The use of the extraction grid also increases the potential for contaminating the ion beam with sputtered grid material.

Higher plasma ion densities may be created by using magnetic coils in a magnetron sputtering machine to increase the electron path length. Just as the ionospheric van Allen belts are confined by the Earth's dipole field, plasma remains confined within the magnetic envelope generated by the coils, thus preventing electron loss to plasma-exposed surfaces. This principle is exploited, for example, in magnetically enhanced ion etching (MIE) (see coils in Figure 3.13). A magnetic field is applied so that electrons cannot pass directly from the anode to the cathode, but rather follow helical paths between collisions, greatly increasing their path length and ionization efficiency. To understand this enhancement in ionization efficiency, consider an electron moving with a velocity v_e. This velocity will be affected by both the electric and the magnetic fields, and the force on the electron is given by the Lorentz force (see Volume I, Equation 3.64):

$$F = e\,(V_e \times B) \qquad \text{(Volume I, Equation 3.64)}$$

in which all quantities, except for the electron charge, are vectors. The direction of the force vector on the electron is perpendicular to both the magnetic field and the direction of the velocity; consequently, the electron will spiral around the magnetic field lines (see Figure 3.14).

With an angle ϕ between the momentary electron velocity vector and the magnetic field vector, the radius r of the electron orbit is:

$$r = \frac{m_e v_e}{eB\sin\phi} \tag{3.16}$$

where m_e is the mass of the electron. Equation 3.16 is obtained by equating the magnetic and centripetal forces. When the electron is moving at right angles to the magnetic field, this simplifies to:

$$r = \frac{m_e v_e}{eB} \tag{3.17}$$

The circular motion of electrons in the presence of a magnetic field can also be described in terms of an angular frequency ω_c (also the cyclotron frequency), which can be calculated from the electron velocity v_e and the radius r of the electron helix as described in Equation 3.17:

$$\omega_c = \frac{v_e}{r} \tag{3.18}$$

At an electron energy of 2 eV, the mean electron velocity is 9×10^5 m s^{-1} (see Equation 3.2). According to Equation 3.17, with a magnetic field B of 5×10^{-3} Tesla (T) applied, electrons will be forced into a circular path with a radius of about 1 mm.[21] At a higher electron temperature of 100 eV, the cyclotron radius of an electron in a field of 0.01 T (or 100 Gauss) is 3.2 mm. These cyclotron radii are of the order of the mean free path and smaller than the smallest dimension of the average dry etch reactor. By lengthening the electron path considerably, electrons are prevented from reaching the reactor walls where

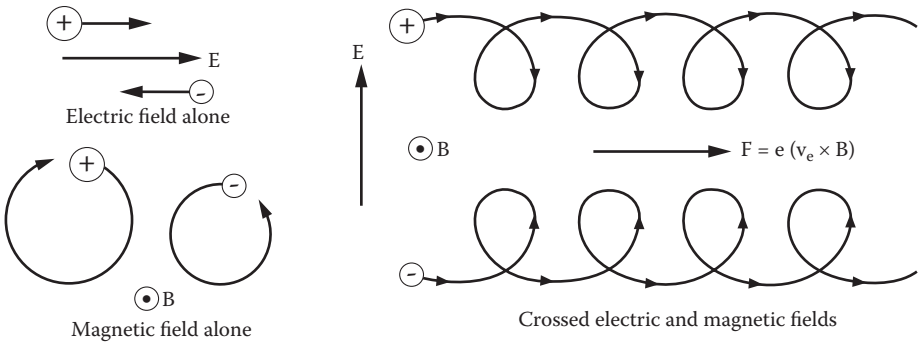


FIGURE 3.14 Electron path with electric field alone and with magnetic field alone and finally with crossed electric and magnetic fields.

they would be lost. Any velocity component of an electron parallel to the magnetic field will change the circular path into a helix. In principle, ions also will move in helical paths; however, the radius r (see Equation 3.17), because of the heavy mass of ions, is so large that the effect can be neglected. Clearly, from the above, a magnetic field can lower wall recombination by confining the electrons, hence making relatively low field strengths adequate for obtaining high plasma densities. When plasma density is increased with magnetic confinement, the degree of ionization is between 10^{-2} and 10^{-4}, compared with 10^{-4}–10^{-6} for simple plate discharges, resulting in an increased etch rate.

The cyclotron resonance frequency of an electron estimated from Equation 3.18 is about 10^8 s^{-1}, which is in the microwave range. If the RF excitation frequency is in the same range, resonance takes effect, causing an additional increase in the RF energy transfer efficiency. Microwave energy may be injected into plasma to energize the ECR ion source as reviewed further below (Figure 3.36b).

When inert argon ions in a triode setup as shown in Figure 3.13 are replaced with reactive ions, reactive ion-beam etching (RIBE) occurs. The ions not only transfer momentum to the surface but they also react directly with the surface, which means a chemical/physical mechanism is involved. Direct reactive ion etching of a substrate is the exception rather than the rule, and it is the radicals generated in the plasma that usually dominate the chemical reactions at the surface.

Systems using a hot filament to produce a plasma have a relatively short lifetime in the presence of reactive gases, as conventional filament materials tend to etch away (e.g., tungsten in the presence of halides forms volatile tungsten halides).[21] The latter type of plasma source problem is avoided by using microwave excitation.

IBE may cover an area of 3–8 cm. This also is referred to as *showered ion etching.* Yet another type of IBE, a maskless technique called *focused ion beam* (FIB), in which the beam is made extremely narrow and used as a direct writing tool, is discussed in more detail in Chapter 6.

Because of its low etch rate and low selectivity, IBE is used more for fundamental mechanistic studies rather than for patterning. For example, in IBE, the substrate can be tilted relative to the direction of ion bombardment, and, as the ion energy at the substrate and the angle of incidence are known, fundamental information, such as sputter yields, can be obtained.[21] The sputter yield S is defined as the mean number of atoms removed from the surface per incident ion. The sputtering yield depends on the incident ion energy, the mass of the incident ion, the mass of the substrate atom to be etched away, the angle of the incident ion with respect to the substrate, the crystallinity and the crystal orientation of the substrate, the temperature of the substrate during etching, and the partial pressure of oxygen in the residual gas. The sputter yield as a function of the incident ion energy is shown in Figure 3.15. The sputtering yield increases with the incident ion energy and reaches a broad maximum at energies between 5 and 50 keV. Beyond 50 keV, the sputtering yield decreases because of the deeper penetration of the ions in the substrate—that is, ion implantation.[22] The angular dependence of the sputtering yield S will be considered below when we discuss etching profiles in physical etching.

In Chapter 7 we will learn more about sputter deposition. In sputter deposition the substrate on the cathode is replaced by a sputter target cathode, and materials sputtered from that target are deposited on the substrate now placed on the anode.

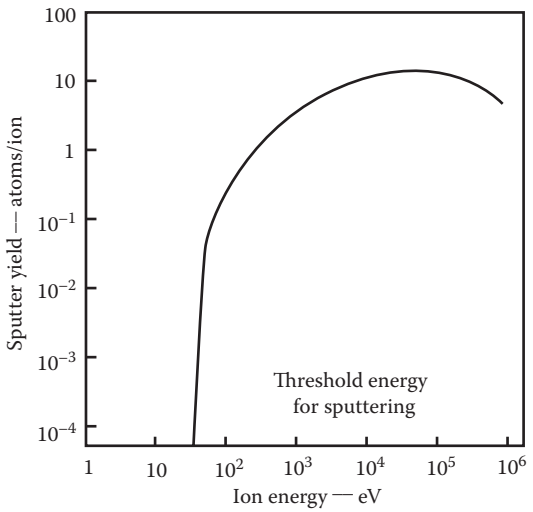


FIGURE 3.15 Sputter yield as a function of the energy of the bombarding ion.

Etching Profiles in Physical Etching

Introduction

The ideal result in dry or wet etching is usually the high-fidelity transfer of the mask pattern onto the substrate, with no distortion of critical dimensions (CDs). Isotropic etching (dry or wet) always enlarges features and thus distorts CDs. Chemical anisotropic wet etching is crystallographic. As a consequence, CDs can be maintained as long as features are strategically aligned along certain lattice planes (see Chapter 4 on wet chemical etching and wet bulk micromachining: pools as tools). With sputtering, the anisotropy is controllable by the plasma conditions.

As can be seen from Figure 3.3a, ion etching and ion milling do not lead to undercutting of the mask, but the walls of an etched cut are not necessarily vertical. A variety of factors contribute to this loss of fidelity in pattern transfer, and they are either caused by involatile sputtering reaction products or by special ion-surface interactions. We shall briefly review these dry physical ion etching problem areas.

Faceting Due to Angle-Dependent Sputter Rate

In Figure 3.15 sputter yield S is shown as a function of the energy E of the bombarding ion. The sputter yield is a complex function of the binding energy U of the material and its atomic number $Z_{material}$ as well as the nature of the sputtering gas (with its own atomic number Z_{gas}). S increases for heavier gases and also depends on the angle of incidence of the ions (θ). The dependence of S on different factors is summarized in Table 3.7. Many different phenomena may occur when an ion hits a substrate/target.

TABLE 3.7 Sputtering Yield S Dependence on Various Factors

Factor		Effect on Sputter Yield S
Target material	Binding energy U	Higher U means lower S
	Atomic number $Z_{material}$	Higher Z means lower S
Sputtering gas	Atomic number Z_{gas}	S increases for heavier gases
	Incident energy E	S increases for higher energy
Geometry		S reaches a maximum 20–30° of glancing

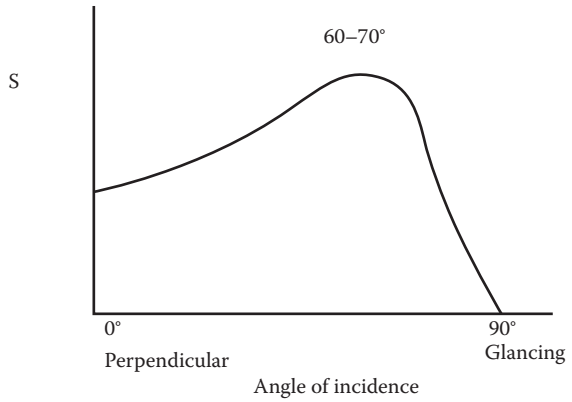


FIGURE 3.16 Sputter yield as a function of angle of incidence of bombarding ions.

The ions may bounce back, implant, absorb on the surface, or reflect by a grazing collision sequence.

In Figure 3.16 we plot the sputter yield S as a function of θ, the angle at which ions are directed at the surface (0° is perpendicular and 90° is glancing). Sputtering is often most efficient at about 20–30° off glancing. In sputtering (at energies <50 keV), the incoming ions may be modeled as hard spheres colliding with the substrate, with 95% of the incident energy going into the substrate (thus making cooling of the target absolutely required) and 5% of the incident energy carried off by substrate/target atoms. The atoms come off with a cosine distribution. It intuitively can be understood that 1) at a glancing angle the incoming ions do not eject atoms from the substrate/target ($S = 0$), 2) hitting the substrate perpendicularly with ions leads to fewer atoms being ejected from the substrate/target, and 3) that in between there are a maximum number of atoms being kicked out.

Even when starting out with a vertical mask sidewall, ion sputtering exhibits a tendency to develop a facet on the mask edge at the angle of maximum etch rate. This corner faceting is detailed in Figure 3.17a. The corner of the mask, always a little rounded even when the mask walls are very vertical, etches faster than the rest and is worn off. Faceting at the mask corner arises because the sputter yield S for materials is a function of the angle at which ions are directed at the surface [$S = f(\theta)$, see Figure 3.16]. The sputter-etch rate of resist, for example, reaches a maximum at an incidence angle of about 60°, more than twice the rate at normal incidence.[22] Sloped mask sidewalls may eventually be followed by sloped etch steps in the

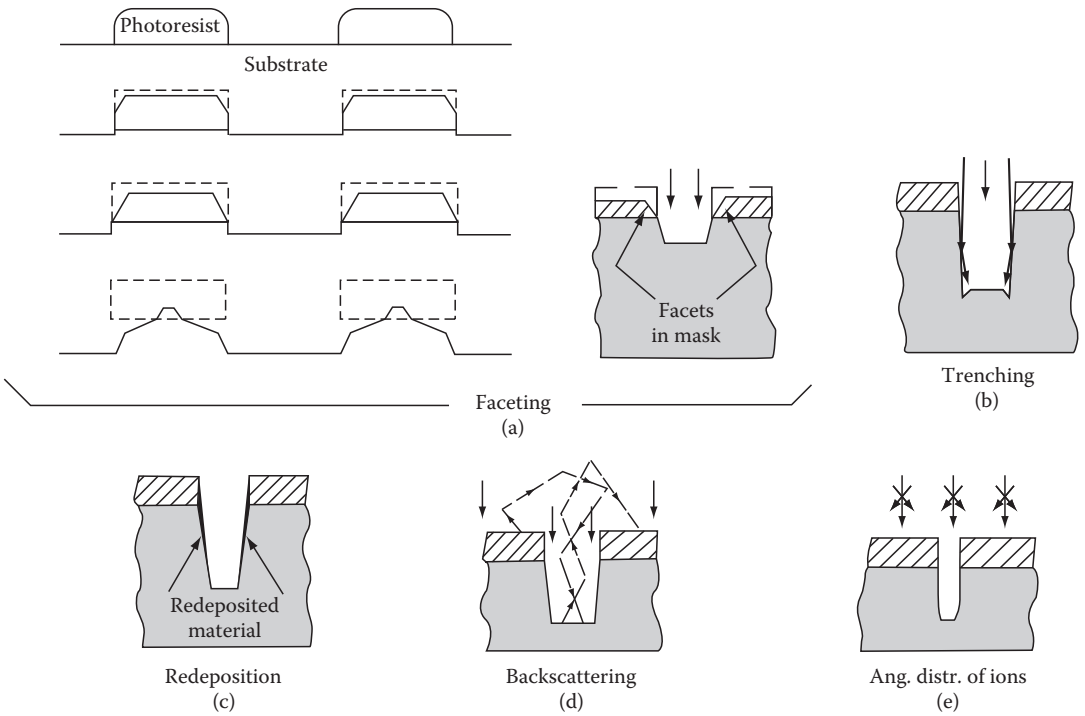


FIGURE 3.17 Limitations of dry physical etching. (a) Faceting. Sputtering creates angled features. An angled facet (~60°) in the resist propagates as the mask is eroded away. Sloped walls may also be created in the underlying substrate. (b) Ditching or trenching as a result of glancing incidence of ions. (c) Redeposition of material sputtered from the bottom of a trench. (d) Backscattering. (e) Angular distribution of incident ions. (Based on Lehmann, H. W. 1991. *Thin film processes II.* Eds. J. L. Vossen, and W. Kern. Boston: Academic Press.[1])

substrate. Faceting of the substrate itself will proceed along its own preferred sputtering direction angle. The faceting is more pronounced with an applied bias as a result of the increased electric field at corners. Usually, faceting affects only the masking pattern, and making the mask sufficiently thick minimizes its influence on the fidelity of the pattern transfer.

It should be noted that some of the disadvantages of physical etching, such as resist corner faceting, can sometimes be exploited. For example, a gently sloping edge is advantageous to facilitate metal coverage or planarization because a tapered sidewall is easier to cover than a vertical wall, especially when using line-of-site deposition methods. The method most often used to obtain such a taper is a controlled resist failure, that is, erodible or sacrificial masks (Figure 3.18). In other words, the "negative effect" of faceting described in connection with Figure 3.17a is put to good use.

Ditching or Trenching

When the slope of the side of the mask is no longer completely vertical, some ions will collide at a glancing angle with the sloping edges before they arrive at the etch surface. This gives a local increase in etch rate, leading to ditches (Figure 3.17b). For this mechanism to be active, there must be a sizable fraction of ions with at least slightly off-vertical trajectories, or the sidewall must have a slight taper as shown in Figure 3.17b.[23] For example, the taper of the sidewall could result from redeposition (see next section) or faceting (see above). Because ditching is a

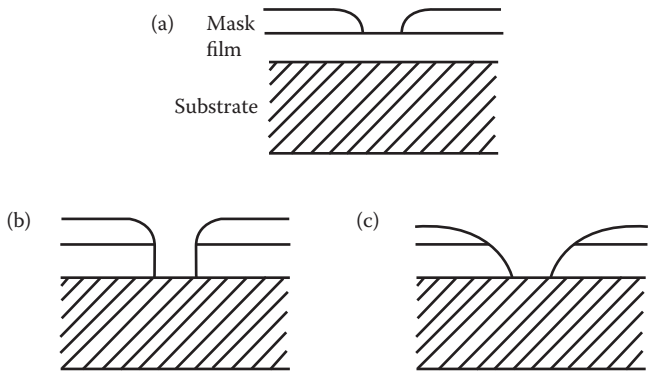


FIGURE 3.18 Erodible or sacrificial masks. (a) Patterned masking structure prior to etching. (b) Nonerodible mask. (c) Partially erodible mask.

small effect (e.g., 5%), it often goes unnoticed unless thick layers are etched.

Redeposition

Another sputtering limitation, already alluded to, is the redeposition of involatile products on step edges (Figure 3.17c). Redeposition involves sputtered involatile species from the bottom of the trench settling on the sidewalls of the mask and etched trench. The phenomenon manifests itself mainly on sloped sidewalls.

By tilting and rotating the substrate during etching (Figure 3.19), etch profiles can be improved. The reasons for tilting and rotating improvements are a combination of shadowing the bottom of the step (to reduce trenching), partially etching the sidewalls of the mask (to reduce redeposition), and gaining more nearly vertical edges on the etched profiles in the substrate. Especially with aspect ratios exceeding unity, redeposition becomes problematic, and reactive gas additives are necessary to generate volatile etch products. With aspect ratios greater than 2:1 or 2.5:1, the etching action at the bottom of too-fine features tends to slow down or stop altogether. Reactive additives bring us into the realm of chemical-physical etching (discussed further below). For micromachining, tilting and rotating of substrates are recurring topics. They are some of the desirable modifications of standard equipment used for micromachining applications (for example, see the fabrication of Spindt field emitters shown in Figure 2.42).

Redeposition may also be used to control sidewall profiles. The material redeposited on the sidewall protects the sidewalls, whereas the material redeposited at the bottom gets removed during the etching process. With the proper control of the redeposition, one can achieve a highly anisotropic profile.

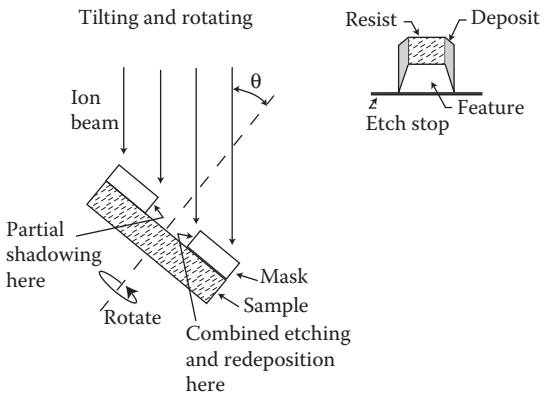


FIGURE 3.19 Left: By tilting and rotating the substrate during etching etch profiles can be improved. Right: Redeposition may be used to control sidewall profiles.

Backscattering

Backscattering, illustrated in Figure 3.17d, is a form of redeposition associated with involatile etch products. A fraction of the sputtered and involatile species from the surface is backscattered onto the substrate after several collisions with gas phase species. This indirect redeposition may involve contaminants from the walls and fixtures in the vacuum chamber. Backscattering fixes the upper pressure limit for ion-enhanced etching. Significant redeposition can take place at pressures as low as 10 mTorr.[1]

Angular Distribution of Incident Ions

Sheath scattering and/or field nonuniformities cause off-vertical ion trajectories. Most ions impinge perpendicularly on the substrate. Scattering of a small fraction of ions in the sheath causes a distribution of impingement angles.[1] This scattering in the sheath is responsible for hourglass-shaped etch profiles that have been observed in trench etching of silicon (see Figure 3.17e). Besides sheath scattering, there is a second mechanism active that may lead to skewing of ion directionality; namely, inhomogeneities in the electrical field at the substrate surface. With etching conductors, the bending of electric field lines caused by surface topography has the effect of enhancing ion flux at feature edges and leads to ditching. On the other hand, with etching insulators, charging effects may cause appreciable ion fluxes to the sidewalls of a trench and contribute to lateral etching. The latter will again lead to hourglass-shaped trenches. The importance of the effect is very much related to the electrical conductivity of the masking and etching surfaces, with the greatest significance for strongly insulating materials.[24]

Temperature Control during Dry Etching

During dry etching, heavy ion bombardment of the substrate generates a lot of heat. To protect photoresist masks, temperature control is required. This could in principle be achieved by water cooling the

wafer-chuck (pedestal or cathode) although the low pressure makes for poor heat transfer between the substrate and the wafer-chuck, so helium backside cooling is required. Alternatively a clamp ring or electrostatic chuck (E-chuck) is used to hold the wafer.

Physical Etching Summary

In physical etching, ion etching or sputtering, and ion-beam milling, argon or other inert ions extracted from the glow discharge region are accelerated in an electrical field toward the substrate, where etching is purely impact controlled. Sputtering is inherently nonselective because large ion energies compared with the differences in surface bond energies and chemical reactivities are involved in ejecting substrate material. The method is slow compared with other dry etching means, with etch rates limited to several hundreds of angstroms per minute compared with thousands of angstroms per minute, and higher for chemical and ion-assisted etching (as high as 6 μm/min with the newest deep RIE equipment). Sputter etching tends to form facets, ditches, and hourglass-shaped trenches and frequently redeposits material in high-aspect-ratio (>2:1) features. Electrical damage to the substrate from ion bombardment and implantation can be problematic. However, some of the reversible plasma damage can be removed by a thermal anneal. With the continuing increase in device complexity, which includes layers of different chemical composition, inert ion etching and ion-beam sputtering will continue to find new applications.

Plasma Etching (Radical Etching)

Introduction

In reactive plasma etching (PE), which corresponds to a high-pressure RIE extreme, reactive neutral chemical species such as chlorine or fluorine atoms and molecular species generated in the plasma diffuse to the substrate where they form volatile products with the layer to be removed. The only role of the plasma is to supply gaseous, reactive etchant species. Consequently, if the feed gas were reactive enough, no plasma would be needed. At pressures of $>10^{-3}$ Torr, the neutrals strike the surface at random angles, leading to isotropic, rounded features. A dry chemical etching regime can be established by operating at low voltages, eliminating impingement of high-energy ions on the sample, and facilitating surface etching almost exclusively by chemically active, neutral species formed in the plasma. The reaction products—volatile gases—are removed by the vacuum system. The volatility of the formed reaction product introduces a major difference with sputtering, where involatile fragments are ejected like a billiard ball and may be redeposited close by.

Reactor Configurations

Reactive plasma etching is one extreme of RIE and basically follows the same process we encountered when discussing dry stripping of resists (Chapter 1, Dry Stripping). Three different popular configurations for plasma etching (barrel reactor, downstream etcher, and parallel-plate system) were shown in Figure 1.36. In resist etching, the process is called *ashing*. Depending on the configuration, high-energy ion bombardment of the substrate can be prevented, more or less, and plasma-induced device damage avoided. For example, in a barrel reactor, the substrates are shielded by a perforated metal shield to reduce substrate exposure to charged high energetic species in the plasma. In a downstream stripper, the geometry of the reactor allows reactant generation and stripping to take place in two physically separated zones (triode-type configuration with a remote plasma). In parallel-plate strippers, the substrates are placed inside the plasma source, which leads to higher ion damage compared with the two previous methods.

Reaction Mechanism

The plasma etching process can be broken down into as many as five primary steps, as illustrated in Figure 3.20. The first step is the production of the reactive species in the gas phase (1). In a glow discharge, a gas such as CF_4 dissociates to some degree by impact with energetic particles in the plasma with an average energy distribution between 1 and 10 eV. In the dissociation, reactive species such as

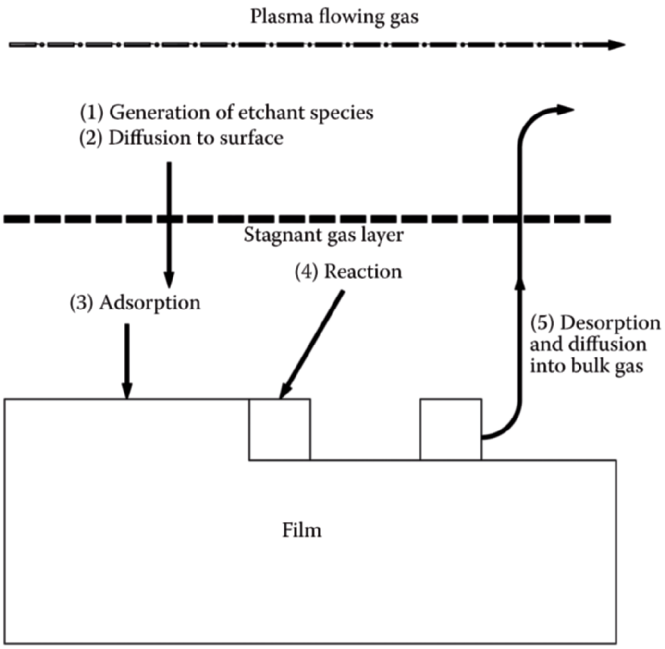


FIGURE 3.20 Primary process steps occurring in a plasma etch process. See text for details.

CF_3^+, CF_3, and F are formed. This step is vital because most of the gases used to etch thin films do not react spontaneously with the film; for example, CF_4 does not etch silicon. In a second step, the reactive species diffuse to the solid (2) where they become adsorbed (3), diffuse over the surface, and react with the surface (4). The etching species must get to the surface to react with a thin film or substrate molecules. The mechanics of getting to the surface can limit aspect ratio and determines undercutting and uniformity. Adsorption also can affect aspect ratio. The reaction is strongly temperature-dependent. Finally, the reaction products leave the surface by desorption and diffusion to the bulk (5). Desorption can stop the etching process if the reacted species are not volatile, and diffusion to the bulk gas phase can lead to nonuniform etching as a result of dilution of unreacted etching species. As in a parallel resistor combination, total resistance is determined by the smallest resistance—the reaction with the smallest rate constant determines the overall reaction rate.

Some of the reaction steps listed above occur in the gas phase and are termed *homogeneous* reactions, whereas others occur at the surface and are called *heterogeneous* reactions. For homogeneous reactions, the plasma only plays the role of creating highly reactive species from the plasma gas. Radicals are more abundant in a glow discharge than ions because they are generated at a lower threshold energy (e.g., <8 eV), which leads to a higher generation rate. Moreover, the uncharged radicals have a longer lifetime. Low-energy ions rarely act as the reactant themselves; instead, neutrals are responsible for most reactive etching (chemical etching) at pressures greater than 0.001–0.005 Torr. Radicals and molecules formed in the plasma are not inherently more chemically reactive than ions, but they are present in significantly higher concentrations. Heterogeneous reactions display even more complexity than homogeneous reactions. In principle, all the species generated in the plasma may influence the reaction rate at the surface; nonreactive species may decrease the reaction rate by blocking surface sites, and adsorption of radicals may enhance the reaction rate. Radicals and other neutrals reach the surface by diffusion, whereas ions are accelerated toward the surface by the negative potential on the substrate electrode. Besides chemically active species and ions, the effect of electron bombardment and irradiation by visible and UV radiation emanating from the plasma requires consideration. Chemical etching occurs at low bias, and because, in principle, no highly energetic ions bombard the surface, sputtering itself is not an important surface-removal mechanism, and radiation damage to the substrate is reduced. Unfortunately, the term *reactive ion etching* is used indiscriminately for all chemical dry etching, even though ions themselves are not the major reactive species. Radicals and molecules also serve as the primary depositing species for all types of films in plasma-enhanced chemical vapor deposition (PECVD) (see Chapter 7). Ions directly participate in chemical etching only in RIBE, where ion reactions at very low pressures can etch at modest rates of less than 400 Å/min. RIBE is an example of physical/chemical dry etching where the same ion has both a physical (ion impact) and a chemical (reactive etching) component. Working with remote plasma in a triode-type system (e.g., a downstream stripper) further reduces ion bombardment of and current flux to the wafer. The steps of reactant generation and etching are separated efficiently because charged species (mainly electrons and positive

O_2^+ ions in the case of pure oxygen plasma used for stripping resist) suffer a much higher loss than reactive neutrals such as atomic oxygen as a result of the presence of plasma excitation.

In the absence of crystallographic effects (typically seen with III–V compounds but not with Si), chemical dry etching leads to isotropic profiles only when ions do not assist the reaction. In the case where ions assist, anisotropy is induced (see below under Energy-Driven Anisotropy or Damage Mechanism). Isotropic etching of the reactive species leads to mask undercutting. In some cases mask undercutting is required—for example, in device fabrication involving lift-off or when layers on sidewalls must be cleared.

Because of the chemical nature of the etching process, a high degree of control over the relative etch rates of different materials (i.e., selectivity) can be obtained by choosing suitable reactive gases. During stripping, oxygen plasma removes photoresists by oxidizing the hydrocarbon material to volatile products. Fluorine compounds are used for silicon etching, and many materials are susceptible to chlorine etching. For example, aluminum is etched in chlorine but not in fluorine because aluminum chlorides are volatile, whereas aluminum fluorides are involatile. Mixtures of gaseous compounds such as CF_4 (fluorocarbon)-O_2 assist in patterning silicon, silicon dioxide, silicon nitride, and so on.

Macro- and Microscopic Loading Effects

Macroscopic Loading Effects

In dry etching, the number of radicals in the plasma is in the same range as the number of atoms to be removed; in wet etching, on the other hand, the number of etchant molecules might be 10^5 times higher than the number of atoms to be removed. A "loading effect" occurs in dry etching as the etchant is depleted by reaction with the substrate material. As a result, the etch rate is inversely proportional to the Si area that is exposed to the plasma.[2] The more purely chemical the etching process, the bigger the loading effect. With lower pressures, the loading effect becomes smaller. Conventional plasmas can sustain enough radicals to etch at rates of 1000 Å/min. With more effective power sources, such as a cyclotron or magnetron, 10^4 Å/min can be achieved.

The loading effect brings an important limitation of dry etching to light: as the etch rate becomes dependent on wafer loading, uniformity is jeopardized. If the supply of reactant limits the etch rate, small variations in flow rate or gas distribution uniformity may lead to nonuniformities in the etch rate. The gas flow F is the most important parameter to control in this regard. Gas flow rates are on the order of 5–200 sccm (standard cubic centimeter per minute), with a maximum flow rate set by the maximum pumping speed and the desired working pressure.[1] The symmetry of the gas flow (i.e., the relative position of gas inlet and pumping port) has to be optimized for a given reactor configuration. Hence a minimum flow rate of the reactant gas prevents the process from being limited by reactant supply. A utilization factor, U (the ratio of rate of formation of etch product to the rate of etch gas flow), may be defined, and $U \geq 0.1$ is suggested for uniform etching. To illustrate: a 500-Å/min etch rate of a 3-in. Si wafer in a CF_4 plasma corresponds to a removal rate of 5×10^{19} Si atoms/min, or a SiF_4 evolution rate of approximately 2 sccm. This means that the CF_4 flow rate should be at least 20 sccm.[1,25]

Closely linked to the utilization factor U is the residence time τ in seconds of the feed gas in the dry etch reactor. The residence time is given by:

$$\tau = \frac{PV}{F} \tag{3.19}$$

where V = reactor volume in liters
P = steady-state pressure in Torr
F = flow rate of the feed gas in Torr-liters per second

(Most gas flow meters are calibrated in standard cubic centimeters per minute and 1.0 Torr-L/s = 79.0 sccm.)

The residence time represents the average length of time a molecule of gas spends in the chamber irrespective of any chemical reaction that might occur.

Etching uniformity is also impacted by the relative reactivity of the wafer surface with respect to the cathode material used. Resulting nonuniformities in this case are referred to as *bull's eyes* because circular interference patterns appear on an etched wafer.

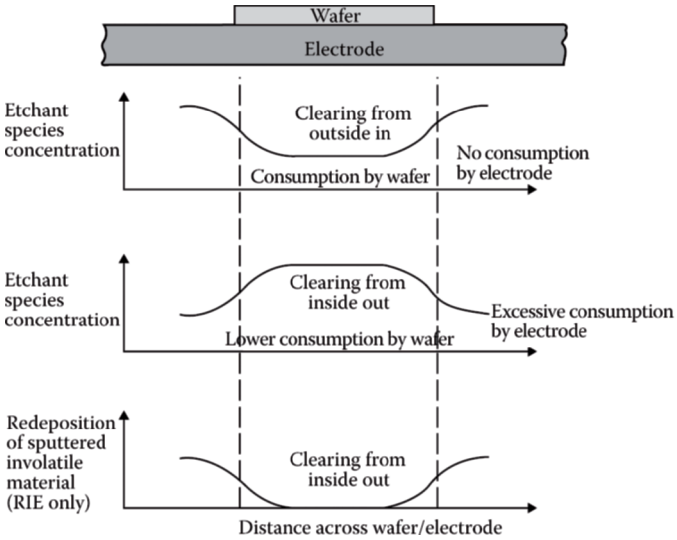


FIGURE 3.21 Chemical and physical effects leading to nonuniform etching across a wafer (bull's eye effect). (From Elwenspoek, M., H. Gardeniers, M. de Boer, and A. Prak. 1994. *Micromechanics*. Report No. 122830, Twente, the Netherlands: University of Twente.[25])

If an aluminum electrode is used for Si or poly-Si etching in SF_6 plasma, a very pronounced bull's eye effect will be evident. The striking nonuniformity in this etching pattern results from a lower consumption of reactant species above the aluminum electrode because aluminum only mildly reacts with reactive species formed by SF_6. By applying an electrode material that consumes the fluorine reactive species as fast as Si itself (e.g., a Si cathode), concentration gradients of reactant species at the edge of the wafer are avoided, resulting in uniform etching (see Figure 3.21).

Microloading Effects during Plasma Etching

The loading effect may also be observed while etching different features on the same wafer. In this case one refers to local loading or microloading effect. For example:

1. Density of the unmasked areas may vary over small distances, resulting in differences in etch rate analogous to macroscopic loading.
2. Etch rates in DRIE show a dependence on the aspect ratio of the feature being etched; the etch rate is diffusion limited and decreases with increasing aspect ratio. In cases where a high-aspect-ratio trench is etched in the vicinity of a wide trench, the latter locally decreases the availability of etching species at the expense of the narrower trench.[26] The effect is also known as "RIE lag" or aspect-ratio-dependent etching (ARDE). Lower etch rate for higher AR trenches (smaller widths) is illustrated in Figure 3.22.

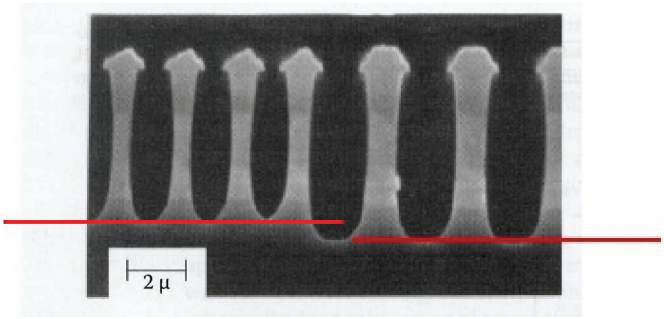


FIGURE 3.22 Illustration of RIE lag. The probability of reactant species reaching the bottom is reduced. (See red lines; wider trenches etch deeper.)

Possible mechanisms for explaining the RIE lag are the depletion or trapping of the reactant species (conductance limitations) as the species travel to the bottom of the trench, thus causing distortion of the ion paths as a result of charging and shadowing effects involving neutrals or ions. In all cases the result is the reduction of the probability of reactant species reaching the bottom.

Silicon Grass or Black Silicon

A common issue of nonuniformity with RIE, especially in the case of deep dry chemical etching (DRIE; see below), is the micrograss or black silicon structures observed, for example, at the bottom of a deep Si trench (Figure 3.23). Jansen et al.[2,27] associated the effect, with all types of micro- and nanomasking from materials deposited or grown on the Si, for example, native oxide or dust (that is, contamination already present on the wafer before etching). Si spikes formed as a result of dirty wafers before etching are easily prevented by a precleaning step, and because the etch is very selective, the native oxide needs to be removed completely. More grass may form during etching because of redeposition of mask material.[2]

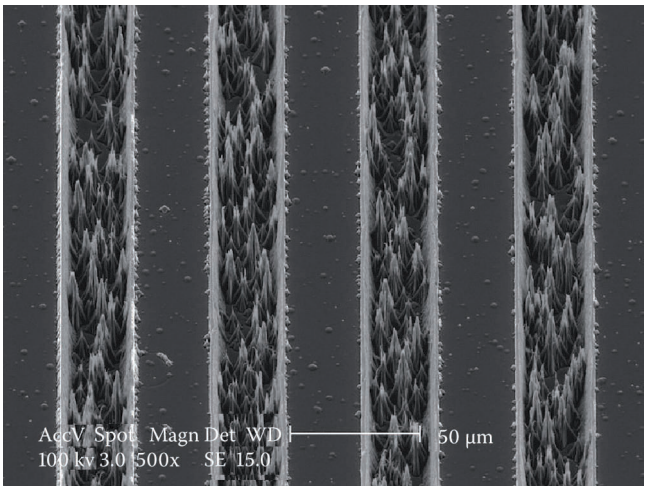


FIGURE 3.23 Silicon "micrograss" structures, which form by plasma etching of silicon in an atmosphere that causes small pieces of silicon dioxide to sputter and then block (mask) etching where they land.

This source of black Si is more pronounced with higher ion energies; therefore, working with a lower self-bias is in general advantageous. Grass formation during RIE was also explained as a consequence of the sharpening of the ion angular distribution with the increasing aspect ratio of a trench during etching.[23] To further minimize the latter types of nonuniformities, wafers should be positioned away from the edges of electrodes to eliminate edge effects caused by changing sheath thicknesses and varying the angles of incidence. A good thermal contact between wafers and cathode is also important. Local temperature variations caused by nonuniform heat-sinking can lead to large etch nonuniformities, particularly in chemically dominated processes.[25]

Atmospheric Plasma Etching

Introduction

Until recently, most plasma science experiments were carried out under vacuum conditions. However, atmospheric-pressure plasmas enjoyed a growth spurt in the 1980s driven by new applications such as high power lasers, MEMS electrostatic switches, novel plasma processing applications and sputtering, EM absorbers and reflectors, remediation of gaseous and other pollutants and waste streams, sterilization, decontamination, and other biomedical applications. Recently, new hardware designs have enabled the stable operation of nonthermal plasmas at atmospheric pressure, and publications on atmospheric-pressure glow discharge plasmas have grown rapidly. We introduce two examples of atmospheric plasmas: one operates in an argon plasma arc, whereas the other is arc free.

The main characteristics of a plasma, such as the breakdown voltage and the voltage current characteristic and structure of the discharge, depend on the geometry of the electrodes and the vessel, the gas used, and the electrode material. In the highly

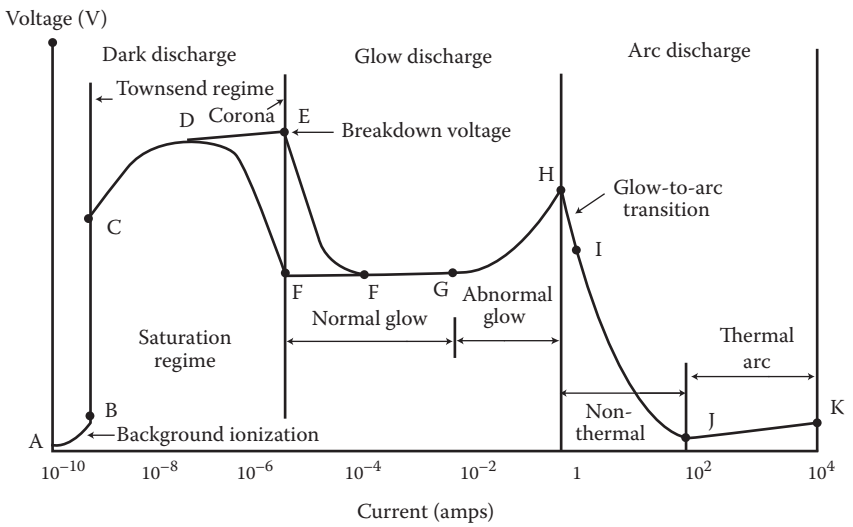


FIGURE 3.24 Voltage current plot of a glow discharge.

nonlinear voltage current plot of plasma, as shown in Figure 3.24, we discern several regimes:

A to B: The applied electrical field E accelerates residual ions and electrons as a result of background radiation in the reactor (e.g., from cosmic rays, radioactive materials in the building). These ions and electrons move toward the electrodes.

B to C: As the applied V increases, eventually all the available residual charges are swept up. This causes the current through the plasma to saturate.

C to E: As the applied V increases beyond C, the E field is now large enough that the remaining electrons can ionize gas. As V increases further, secondary electrons also cause ionization.

D to E: In the so-called Townsend regime, coronal discharges occur at various edges and sharp points of the electrodes in the gas. This is the result of the E field concentration at such sites.

A to E: Electrical breakdown occurs. The voltage is now high enough to cause ionization to occur because of secondary electrons in the gas. Additional secondary electrons are generated at the cathode as a result of ion impact on it. There is a sudden increase of the current by several orders of magnitude, causing an avalanche.

F to G: After breakdown from E to F, the gas enters the normal glow where the voltage V is almost independent of the current over several orders of magnitude. As the current increases, the fraction of the cathode occupied by plasma increases until the plasma covers the entire cathode surface at G.

G to H: In the abnormal glow regime, further increases in the applied voltage cause increases in the current. This is the region of operation for most sputtering and other plasma systems.

H to K: At H, the electrodes have become sufficiently hot that the cathode can now emit electrons thermionically, and a second avalanche can happen in the arc region (this is the regime used in spectrochemistry). If the DC power supply is capable (low internal R), a transition from glow to arc happens in the region from H to I. In this arc regime from I to K, the discharge V decreases as I increases. The point J marks the transition from a nonthermal to a thermal arc (electrons have the same temperature as the gas).

At atmospheric pressures, plasmas are weakly ionized, and the mean-free path for electron-molecule collisions is very short (<1 μm). These plasmas are typically thermal with electrons and gas at the same temperature. At atmospheric pressure a small perturbation can easily cause a glow-to-arc transition. The transition time is very short, and this makes it hard to stabilize an atmospheric glow discharge. Some tricks to prevent an arc include the use of inert gas to lower breakdown voltage, high-frequency power sources and operation in the dielectric barrier discharge mode, and a structured cathode (e.g., a pin cathode). Stabilization of high-pressure (atmospheric-pressure) plasmas is most easily achieved if the plasma is confined to "small" spaces. Interesting pattern formation and self-organization has been observed in these microplasmas.

Atmospheric Downstream Plasma

In atmospheric plasma jet etching, reactive chemical species are generated in a DC arc between two electrodes in a noble gas (such as argon) at atmospheric pressure. A stream of reactant gas (the gas jet) such as CF_4 then flows into the arc and onto the wafer, which is located up to 12 in. away to avoid damage, as shown in Figure 3.25.[28] Although atmospheric plasma sources have been used previously for industrial powder spray and plasma cutting (see Chapters 5 and 9), they have not been used in microelectronic processing for fear of damaging and contaminating the silicon wafer surface. Among the concerns were possible overheating as a result of high power density, contamination from chemical erosion of electrodes, and nonuniform etching.

Atmospheric downstream plasma (ADP) processing overcomes these problems by using a magnetically controlled, inert gas, DC arc-plasma discharge.

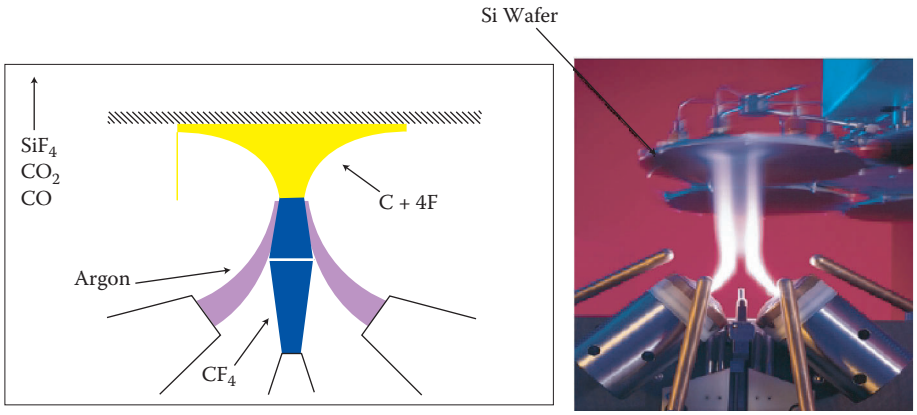


FIGURE 3.25 Atmospheric downstream plasma processing of a Si wafer. (From Tru-Silicon Technologies, http://www.trusi.com/frames.asp?.)

Plasma jet etching is an atmospheric variant of plasma etching capable of producing very high etch rates without substrate damage. By deploying the plasma in a downstream configuration, ADP prevents ion impact damage while keeping wafer temperatures low. This technology may improve the performance and reduce the cost of bulk material removal operations, such as mechanical grinding, lapping, and wet chemical etching [especially chemomechanical polishing (CMP); see Chapter 1, Figure 1.66]. ADP etching for back-side damage removal after grinding and wafer thinning is a potential major application. Other applications include damage-free dicing and via etching (for the latter two applications, an aluminum mask is used). The trend to thinner wafers and thinner chips will not stop with the current 200–250-μm thickness, but is forecast to continue reducing to 100–150 μm and even thinner (see Figure 3.26), driven by the demand for portable electronic devices, smart cards, multichip modules, PC Card (originally PCMCIA Card for People Can't Memorize Computer Industry Acronyms or more seriously for Personal Computer Memory Card Interface Association), and so on. Wafer strength requirements of large-diameter wafers limit back-grinding to thickness of 200–250 μm; therefore, new technology is essential. Tru-Silicon Technologies is one of the companies providing this type of equipment (for a tutorial on applications of ADP, see http://www.trusi.com/frames.asp?).

Atmospheric Arc-Free Plasma Etching

A picture of an atmospheric, arc-free (!), plasma jet invented at the University of California, Los Angeles is shown in Figure 3.27.[29] The "cold flame" (T ~ 75°C) shown here is impinging on a finger (http://www.seas.ucla.edu/prosurf). This jet was developed for

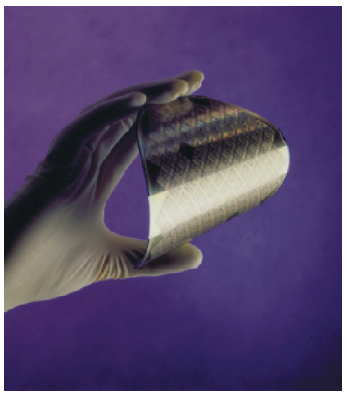


FIGURE 3.26 A 50-μm thin silicon wafer produced using atmospheric downstream plasma (ADP). (From Tru-Silicon Technologies.)

FIGURE 3.27 A picture of an atmospheric, arc-free (!), plasma jet invented at University of California, Los Angeles. The "cold flame" (T ~ 75°C) shown here is impinging on a finger (http://www.seas.ucla.edu/prosurf).

etching materials and depositing materials at atmospheric pressure and temperatures between 100 and 275°C. Gas mixtures containing helium, oxygen, and carbon tetrafluoride are passed between an outer, grounded electrode and a center electrode, driven by 13.56-MHz RF power at 50–500 W. At a flow rate of 51 L/min, a stable, arc-free discharge was produced. The discharge extends out through a nozzle at the end of the electrodes, forming a plasma jet. Materials placed 0.5 cm downstream from the nozzle are etched at high rates: 8.0μm/min for Kapton (O_2 and He only), 1.5 μm/min for silicon dioxide, 2.0 μm/min for tantalum, and 1.0 μm/min for tungsten.[29] The plasma jet produces a large flux of atoms and/or radicals, depending on the gas fed to the device. The setup may also be used for deposition. For example, tetraethoxysilane was used to deposit high-quality silicate glass films. It should be noted that like ADP this is a downstream plasma source. The ions and electrons are rapidly consumed by collisions before impinging on the substrate.

Physical/Chemical Etching

Introduction

The most useful plasma etching is neither entirely chemical nor physical. By adding a physical component to a purely chemical etching mechanism, the shortcomings of both sputter-based and purely chemical dry etching processes can be surmounted.

In physical/chemical techniques, ion-surface interactions promote dry etching. A first type is found in reactive ion beam etching (RIBE). RIBE constitutes a rather exceptional case of energy-driven anisotropy where ions are reactive and etch the surface directly. More common in energy-driven etching is *chemically assisted ion-beam etching* (CAIBE). In this case, ion bombardment induces a reaction by making the surface more reactive for the neutral plasma species, for example, by creating surface damage. In inhibitor-driven anisotropy, ions clear the surface of film-forming reaction products, allowing etching with reactive neutrals to proceed on the cleared areas. We will also learn that silicon dry etch anisotropy may be influenced by local dopant concentrations and that a wide variety of existing gas compositions crucially influences the etch profile. To help the reader summarize matters, a list of simplifying rules is provided at the end of this section.

Energy-Driven Anisotropy or Damage Mechanism

During energy-driven anisotropy, bombardment by ions (<1000 eV) disrupts an unreactive substrate and causes damage, such as dangling bonds and dislocations, resulting in exposed surface atoms, which are more reactive toward etchant species (electrons or photons also can induce surface activation). Ion bombardment is mainly in the vertical direction so that the etch rate in the vertical direction is much higher than in the lateral direction (anisotropy!). Besides "energy-driven anisotropy," the process is also referred to as the *damage mechanism*. In Figure 3.3c, a typical resulting etch profile was shown. Vacuum pumping removes the volatile reaction products. This type of etching is referred to as *reactive ion etching* (RIE) when it involves reactive chemicals in a diode-type reactor, and as *chemically assisted ion-beam etching* (CAIBE) when it involves chemical reactants (e.g., Cl_2) introduced over the substrate surface in a triode-type setup. Figure 3.11 shows a parallel-plate setup for RIE, and a hexode reactor is shown in Figure 3.28. The hexode reactor is designed for batch processing. In this design, the cathode has the shape of a hexagon surrounded by the cylindrical chamber walls forming the anode. As many as 24 100-mm-diameter or 18 150-mm-diameter wafers

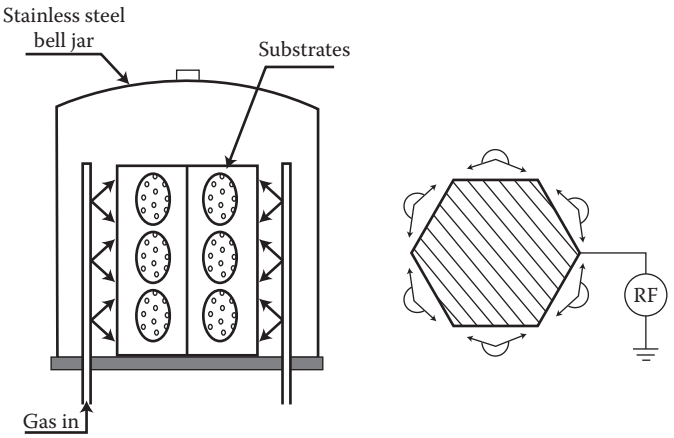


FIGURE 3.28 Schematic diagram of a multiwafer RIE etcher, the so-called hexode reactor.

can be mounted on the slightly tilted sidewalls of the hexagonal cathode. The latter is slowly rotated during operation to increase uniformity.[1]

The etch rate reached by RIE is substantially higher than in ion etching. For example, the etch rate of Si in Ar sputtering hovers around 100 Å/min compared with 2000 Å/min for a reactive gas such as CCl_2F_2. Chemically assisted ion etching can lead to accurate transfers of the mask pattern to the substrate and to a fair selectivity in etching different materials. The directional anisotropy ensues from operating at low pressures and high voltages. Under these conditions, the mean free path of reacting molecules typically grows larger than the depth to be etched, resulting in the horizontal surfaces being hit and etched by reactive neutrals more frequently than the sidewalls. The need for energetic impinging ions, as in the case of physical etching, is reduced as the complex on the surface activates more easily. In Figure 3.29 we illustrate the energy-driven anisotropy process or damage mechanism.

Reactive ion-beam etching (RIBE), like CAIBE, involves chemical reactants in a triode setup. There is an important difference, however. In the case of RIBE, the reactive ions are introduced through the ion source itself; in CAIBE, the reactive gas is fed over the substrate to be etched, and unreactive ions are generated in the ion source. At very low pressures in RIBE systems, reactive ions, substituting Ar ions, can sustain a modest etch rate of less than 400 Å/min. In CAIBE, ion bombardment of a substrate in the presence of a reactive etchant species leads to a synergism where fast directional material removal rates greatly exceed the separate sum of chemical attack and sputtering rates. At one point, this type of etching was thought to be caused by direct chemical reactions between the ions and the surface material (as with RIBE). However, in most practical situations for etch rates ranging from 1,000–10,000 Å/min that is impossible because the ion flux is much lower than the surface removal rates. The synergistic effect of chemical attack and sputtering rates is illustrated in Figure 3.30; the etch rate of Ar^+ ion bombardment alone or the chemical etch rate of XeF_2 alone does not add up to the combined effect.

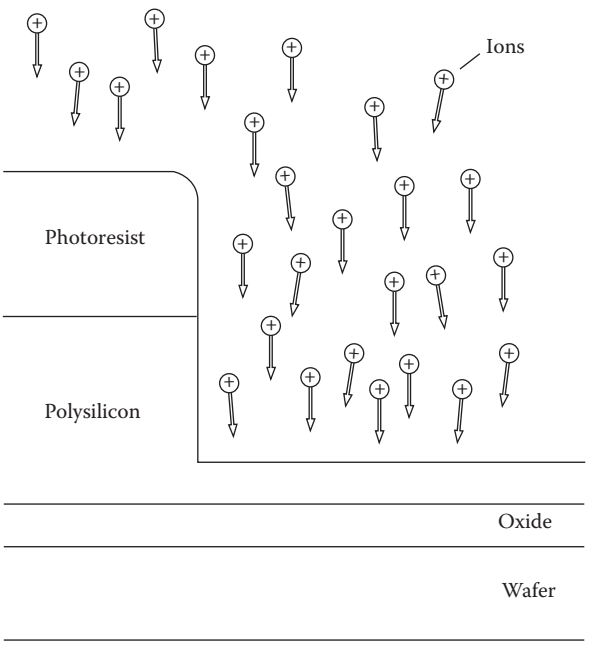


FIGURE 3.29 Energy-driven anisotropy or damage mechanism. (Drawing by Genis Turon.)

CAIBE is particularly effective in achieving highly anisotropic etch profiles because of the independent control of the physical and chemical etching components. To get an idea of how smooth a sidewall can be achieved by CAIBE, consider the work on vertical mirror facets in an InGaN/AlGaN laser diode structure. Atomic force microscope (AFM) measurements show that the laser diode vertical sidewalls etched in a CAIBE system (using Cl_2 and BCl_3 as the reactive species) exhibit a root mean squared roughness of only 40–60 Å, and the inclination angle is within ±2° of vertical.[30] Using $N_2/H_2/CH_4$ chemistry, extremely smooth sidewalls were also observed in CAIBE etching of InP.[31] To reduce surface damage, the aim is to increase the chemical component, decrease the ion acceleration energy, and work with a high ion flux. In CAIBE, as in other ion-beam techniques, the ion acceleration energy and the ion density can be independently controlled. With an ICP or ECR, a highly dense and uniform ion beam (e.g., Ar^+) is generated. ICP and ECR plasmas are discussed below.

The distinction between RIBE and CAIBE is not absolute because the presence of a reactive gas in CAIBE will produce some beam ions from species that back-diffuse into the broad-beam ion source. In a magnetically enhanced RIE (MERIE) system, a higher ion flux is obtained at lower energy through

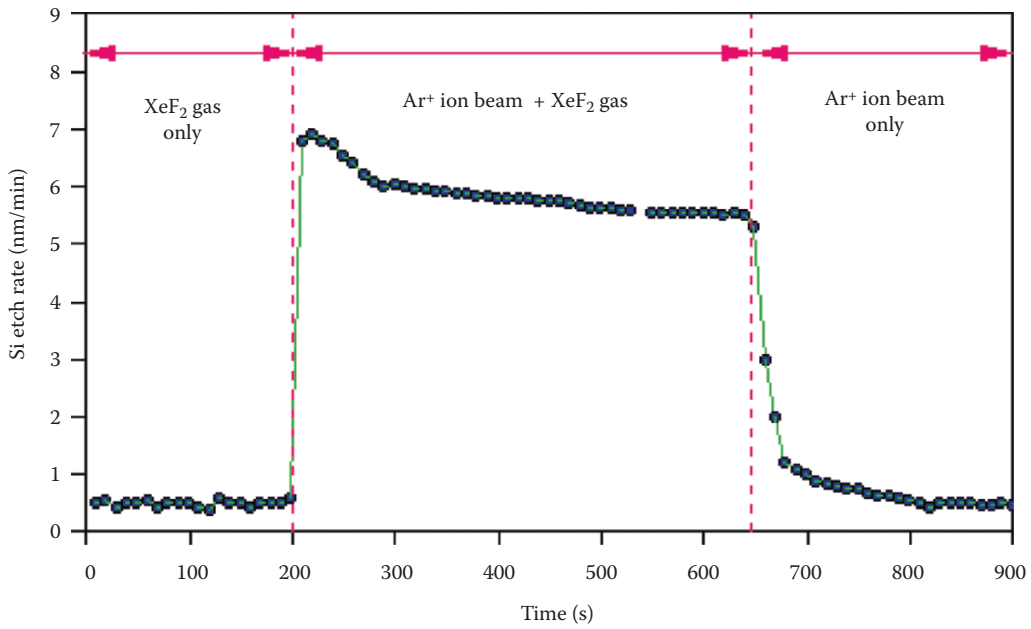


FIGURE 3.30 Etch rate experiment in RIE. Etch rate of Ar^+ bombardment alone or chemical etch of XeF_2 alone does not add up to their combined synergistic effect.

magnetic confinement of the plasma. For typical RIBE and CAIBE performance characteristics also visit http://www.oxfordplasma.de/technols/ibe.htm.

A hypothetical etch profile for Si and SiO_2 etching with a reactive gas and positive ion bombardment is shown in Figure 3.31a. In the absence of ionic bombardment, the etch rate is assumed to be zero for SiO_2; consequently, the SiO_2 film etches with perfect anisotropy (vertical sidewalls). The silicon etch rate in the absence of ionic bombardment, on the other hand, is assumed to be finite but small; consequently, the etched feature ends up having a profile with slanted sidewalls. Mathematically, the above can be expressed in terms of etch depth Z, undercut X, the etch rate under bias V_Z, and the etch rate without bias V_X, as:

$$\frac{V_x}{V_z} = \frac{X}{Z} \tag{3.20}$$

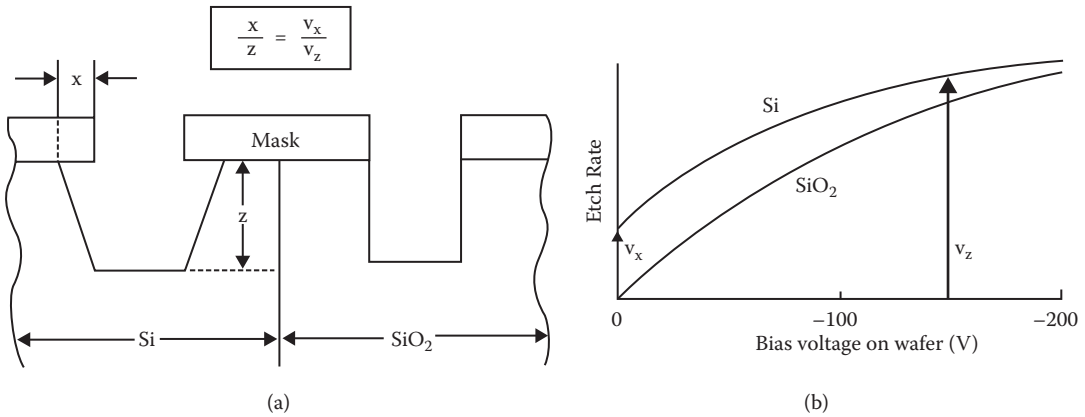


FIGURE 3.31 Relationship between the shape of the etched wall profile and the dependence of the etch rate on the wafer potential. (a) Etch profiles. (b) Etch rate bias dependence. The etch rate of SiO_2 in the z-direction, V_z, increases with increasing wafer bias. Assuming no chemical etching component is present in the SiO_2 reaction, that is, $V_X = 0$ (the etch rate equals 0 at bias = 0), an ideal vertical profile results. The etch rate of Si also is bias-dependent, but some reaction occurs at $V_X = 0$. In other words, a chemical component to the reaction does not need ion bombardment resulting in nonvertical walls (see text). (From Manos, D. M., and D. L. Flamm, eds. 1989. *Plasma etching: An introduction.* Boston: Academic Press.[76])

(See Figure 3.31b.) The anisotropy of RIE can further be enhanced by the use of helium back-side cooling (as low as –120°C might be required). The low-temperature suppression of chemical attack of silicon by fluorine atoms exemplifies how lower temperatures further improve anisotropy and critical dimension control. Although suppressing the isotropic fluorine reaction, the same low temperature influences ion-assisted reactions only slightly and hence improves the anisotropy of profiles; the etch rate at $V_X = 0$ decreases or becomes zero.

Inhibitor-Driven Anisotropy or Blocking Mechanism

Inhibitor-driven anisotropy embodies another example of a physical/chemical etching technique (Figure 3.3d). In this case, etching leads to the production of a surface-covering agent in a so-called blocking mechanism. Ion bombardment clears the "passivation" from horizontal surfaces; reaction with neutrals proceeds on these cleared surfaces only, and the etch proceeds again mainly in the vertical direction as shown in Figure 3.32. The protective film may originate from involatile etching products or from film-forming precursors that adsorb during the etching process. Passivating gases, such as BCl_3 and some halocarbons (Freons™ such as CCl_4 and CF_2Cl_2), are sources of inhibitor-forming species. In the latter case, the reactive gas component appears to be adsorbed on the surface (for example, a polymer is formed), where it is subsequently dissociated by electron, ion, or photon bombardment, clearing the surface for reaction with the reactive neutrals.

An example of etch profile manipulation with inhibitor-driven chemistry using an idealized Si sample is illustrated in Figure 3.33. At the top of this figure, we show the silicon etch rate as a function of percentage of H_2 in CF_4 for a biased and an unbiased wafer. We remember from earlier in this chapter that the silicon etch rate increases as the wafer bias is increased; thus, the etch rate curve for the biased wafer lies above the one for the unbiased wafer. If H_2 is added to a CF_4 feed gas, the Si etch rate decreases, and at some value of H_2 concentration, the nonbombarded surface etch rate decreases to zero ($V_X = 0$ at 10% H_2 in Figure 3.33), whereas the bombarded surface continues to etch. The decrease in etch rate stems from an increase in the amount of passivating polymerization. Aggressive fluorine reacts with the hydrogen so that carbon compounds polymerize more readily. As before, Equation 3.20 can be used to calculate the ratio of underetch X to etch depth Z. At $V_X = 0$, that ratio, of course, is zero.

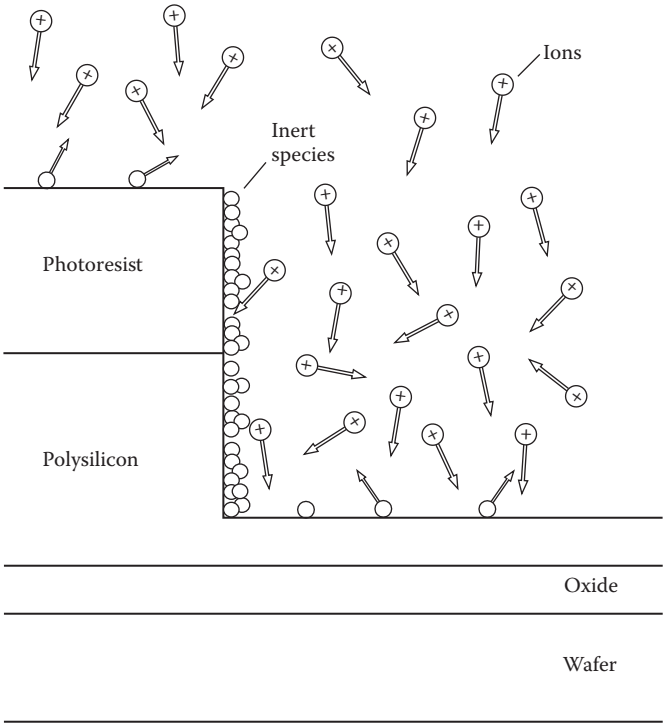


FIGURE 3.32 Inhibitor-driven anisotropy or blocking mechanism. (Drawing by Genis Turon.)

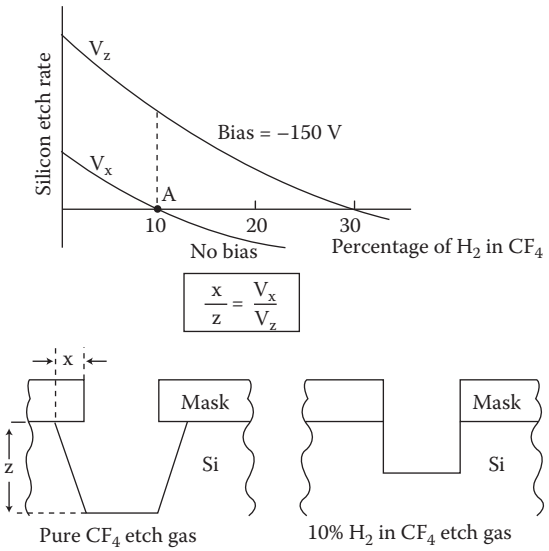


FIGURE 3.33 Trench profile manipulation by decreasing the fluorine-to-carbon ratio (through hydrogen introduction). (From Manos, D. M., and D. L. Flamm, eds. 1989. *Plasma etching: An introduction.* Boston: Academic Press.[76])

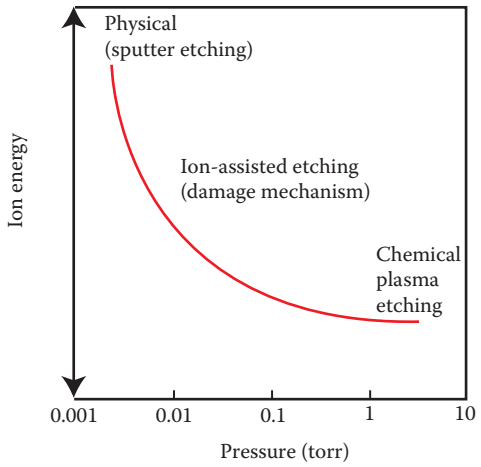


FIGURE 3.34 Ion energy versus pressure for a plasma. (Redrawn by Genis Turon from Flamm, D. L. 1992. Dry plasma resist stripping part I: overview of equipment. *Solid State Technol* 35:37–39. With permission.[36])

Similarly, aluminum etching in $CCl_4 + Cl_2$ or $CHCl_3 + Cl_2$ plasmas presents a good example of an inhibitor system. Even though Cl_2 gas and chlorine atoms formed in these plasmas are rapid chemical etchants for clean aluminum, aluminum in these plasma mixtures can afford near-vertical profiles with excellent line-width control.

Another typical example of a sidewall mechanism is the etching of phosphorus-doped Si, which etches isotropically in Cl_2 plasma but etches anisotropically when C_2F_6 is added to the source gas. Qualitatively the effect is accounted for by assuming that the two gases dissociated in the plasma are:

$$C_2F_3 + e \rightarrow 2CF_3 + e \qquad \text{Reaction 3.1}$$

$$Cl_2 + e \rightarrow 2Cl + e \qquad \text{Reaction 3.2}$$

and the possible surface reactions are:

$$xCl + Si \rightarrow SiCl_x \text{ (etching)} \qquad \text{Reaction 3.3}$$

$$CF_3 + Cl \rightarrow CF_3Cl \text{ (recombination)} \qquad \text{Reaction 3.4}$$

Reaction 3.3 results in etching of the silicon surface, and Reaction 3.4 results in recombination without material removal. Ion bombardment enhances Reaction 3.3, and Reaction 3.4 is assumed to be dominant on the sidewalls. The recombination reaction acts as a sidewall passivant.

In practice, the anisotropy brought about by the directed ion flux in plasma allows the final etched feature to be within 10% of its dimensions on the mask, and submicron resolutions become feasible. A desire to move away from sidewall passivating chemistries to avoid CD loss is prominent, especially in the IC industry. Thick sidewall coatings not only reduce CD control, but they also prove difficult to strip. Processes involving little or no sidewall thickness increase are sought for tight CD control in submicron devices. In this respect, wafer cooling (to as low as –120°C) presents an attractive way of obtaining sidewall protection without the need for additional chemistries. At low substrate temperatures, reaction products are involatile and can serve as very thin sidewall inhibitors. Generally, at lower temperatures, lateral etch rate can be suppressed while using simpler chemistries. In micromachining applications where feature size is in general larger than in ICs and aspect ratios are more extreme, applying passivating chemistries is almost a necessity, but wafer cooling can help the process.

Dopant-Driven Anisotropy

Although not commonly acknowledged, silicon plasma etch anisotropy may be influenced by local dopant concentrations. For example, Li et al.[32] have demonstrated that Cl-based plasma etch can be used to etch lightly doped p- or n-type silicon anisotropically and heavily doped n-type silicon isotropically. These authors, by forming buried n^+ layers beneath a lightly doped epitaxial layer, were able to selectively undercut structures above the n^+ regions.[32] Schwartz and Schaible[33] also described the dopant dependence of dry plasma etching.

It is more commonly known that the Si etch rate depends on the electronic properties of the Si substrate.[34] The n-type silicon (e.g., doped with P, As, and so on) etches faster than intrinsic silicon, which etches faster than p-silicon (e.g., doped with B, Ga, and so on). Dopant concentration effects on etch rates only become apparent at concentrations greater than 10^{19} cm^{-3}. The doping effect is not chemical in nature because it is absent if the dopants are not electrically activated. The effect depends on the electronic structure of the surface, and band-bending effects at the semiconductor surface have explained it. Using ECR etching, Juan et al.[35] demonstrated

that the n^{++} Si etch increases, but the p^{++} Si etch rate decreases with respect to lightly doped silicon. Thus, it was established that, with a high microwave power or high RF power, as well as an increased temperature, a large difference in etch rates of p^{++} and n^{++} Si results.

Ion Energy versus Pressure Relationship in a Plasma: Comparison of Different Dry Etch Technologies

In the preceding sections, we described two extreme cases of dry etching—ion etching (purely physical) and plasma etching (purely chemical)—and the intermediate case of chemical/physical etching. In general, three factors control the etch rate in a plasma reactor: 1) neutral atom and free radical concentration, 2) ion concentration, and 3) ion energy. Ion and radical concentrations control the reaction rate, whereas ion energy provides the necessary activation and controls the degree of anisotropy. The respective contribution to chemical and physical action of plasma can be manipulated by varying voltage and gas pressure. For good CD control, anisotropy is required, and positive ions need to be formed at low pressures ($<10^{-2}$ Torr) to strike the surface at normal incidence. Etching at low pressures, with a long mean free path length of the ions, λ_i, is inherently more anisotropic (i.e., directional) and less contaminating because etch reaction byproducts are more volatile at lower pressures and are easier to remove. At these lower pressures, ion density drops off quickly, causing a lower etch rate and lower wafer throughput. Increasing the power or wafer bias increases the etch rate, as the remaining ions become more energetic. Higher ion energies can cause additional problems in terms of device damage. What is needed is a plasma source operating at low pressure with very high ion density. Operating at low pressures also reduces the effects of chemically induced macro- and microloading (see above). With high pressure, short λ_i, and low voltage, one gets isotropic chemical etching. The pressure/voltage relationship for a plasma is schematically represented in Figure 3.34.[36] Thus, sputtering and dry chemical etching represent the two extremes of a continuous dry etching spectrum (Table 3.8) with physical etching by sputtering with inert argon ions at one end and chemical etching with reactive neutral species at the other.

In Tables 3.9 and 3.10 we contrast the features of different etching techniques, highlight the underlying mechanism, and give a typical application. Entries concerning endpoint detection will be clarified further below.

Gas Compositions in Dry Etching

Most dry bulk etching of silicon is accomplished using fluorine free radicals, generated from fluorine-containing gases and by forming volatile SiF_4 during the etch (the byproducts in dry etching should always have the highest possible vapor pressure). As opposed to chlorine- and bromine-based processes, fluorine plasma reactions with silicon usually proceed spontaneously, not requiring ion bombardment. This method results in high etch rates, and by themselves fluorine radicals produce profiles that are nearly isotropic. Often, chlorofluorocarbons (CFCs) are used to produce polymer deposition in parallel with the etching process. Oxidizing additives increase etchant concentration and suppress excessive polymerization. The addition of oxygen to CF_4 in Si etching typifies the procedure.[37] Oxygen, at concentrations less than 16%, reacts with CF_X radicals to enhance F atom formation [increase of fluorine-to-carbon ratio (F/C)] and eliminate polymerization. At yet higher concentrations, adsorbed oxygen on

Table 3.8 Dry Etching Spectrum

Types of Etching	Methods	Geometry	Selectivity	Excitation Energy	Pressure
Gas/vapor etching	Chemical	Isotropic	Very high ↓	None ↓	High (760–1 Torr) ↑
Plasma etching	Chemical	Isotropic	High ↓	10–100s of watts ↓	Medium (>100 mTorr) ↑
Reactive ion etching	Chemical and physical	Directional	Fair ↓	100s of watts ↓	Low (10–100 mTorr) ↑
Sputtering etching	Physical	Directional	Low ↓	100–1000s of watts ↓	Low (~10 mTorr) ↑

1 Torr = 1 mmHg.

TABLE 3.9 Comparison of Different Dry Etch Technologies

	Chemical Etch	RIE	Physical Etch
Examples	Stripping	Plasma patterned etches	Argon sputtering
Etch rate	High to low	High, controllable	Low
Selectivity	Very good	Reasonable, controllable	Very poor
Etch profile	Isotropic	Anisotropic, controllable	Anisotropic
Endpoint	By time or visual	Optical	By time

the surface depresses the etch rate.[37] Radical scavengers, such as hydrogen, increase the concentration of inhibitor former and reduce etchant concentration of fluorine in the etchant (decrease of F/C) (see also example in Figure 3.33). Heinecke[38] realized that the etch ratio of SiO_2/Si could be increased either by adding H_2 to the CF_4 feed gas or by using CHF_3 or, in general, fluorocarbons displaying a smaller F/C ratio than CF_4. Adding hydrogen to fluorocarbon gases helps promote CF_X film growth (such as CF_2, C_2F_4) for selective SiO_2 etching. Carbon accumulates less on oxide surfaces than on Si surfaces, as SiO_2 surfaces directly react with hydrogen. By adding hydrogen, carbon blocking increases (especially on the Si) because hydrogen scavenges fluorine, forming HF and preventing reaction with carbonaceous species on the surface.

The same approach helps to optimize the selectivity on other oxide/nonoxide systems (e.g., SiO_2/resist, SiO_2/Si_3N_4, and Ta_2O_5/Ta). The F/C ratio should not be made smaller than 2, or contamination of the whole system with C_xF_y polymer sets in. Factors that tend to control polymerization and selectivity are temperature, hydrogen concentration, pressure, and ion bombardment. Some etchants, such as Cl atoms, do not readily etch through thin native oxide films on Si, Nb, and Al, for example. These native oxides prevent the onset of etching unless small amounts of native oxide etchants, such as BCl_3, are added. Inert gases such as argon or helium help stabilize the plasma, enhance anisotropy, improve uniformity, or reduce the etching rate by dilution. Because helium has a high thermal conductivity, it also improves heat transfer between wafers and the supporting electrodes.[25]

From the above, the F/C ratio in dry etching is a key parameter to control. With an F/C ratio >3, etching is dominant, and with an F/C <2, polymerization takes over. In Figure 3.35 we show the F/C ratio for a number of different etchant gases (for CF_4, the F/C ratio is obviously 4 and for C_2F_4 it is 2) and different applied voltages. A negative bias on a surface exposed to the plasma increases, at a constant F/C ratio, the etching behavior over

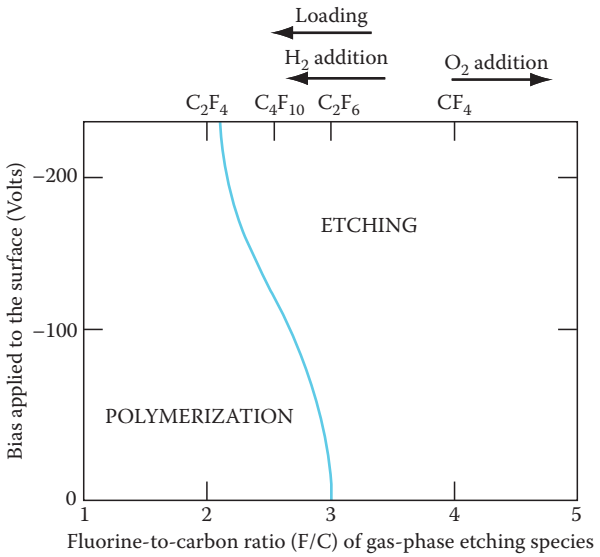


FIGURE 3.35 F/C ratio, DC bias, and polymerization. (Drawing by Genis Turon.)

TABLE 3.10 Dry Etch Mechanism and Their Application

Pure Chemical Etch	Reactive Ion Etch (RIE)		Pure Physical Etch
	Blocking mechanism	Damaging mechanism	
No ion bombardment	Light ion bombardment	Heavy ion bombardment	Only ion bombardment
Photoresist strip	Single crystal Si etch		
Ti strip	Polysilicon etch	Oxide etch	Sputtering etch
Nitride strip	Metal etch	Nitride etch	

polymerization tendency. This effect is caused by the enhanced energy of the ions striking the surface, resulting in polymer sputtering. In an etch of an oxide film, the oxygen byproducts react with C to release more F, and the F/C ratio increases shifting to a more aggressive etching regime. When etching Si or a silicide, no oxygen is released, and F is consumed, lowering the F/C ratio and thus favoring polymerization. For selectivity, etching in the etch-dominant region but near the polymerization region is preferred.

Table 3.11 lists frequently used reactive gases and typical applications.

The Environmental Protection Agency mandate to eliminate the use of CFCs changed the type of gases used for dry etching purposes. This evolution in gas mixtures is captured in Table 3.12.[39]

The primary new etchant gases are SF_6, NF_3, and SiF_4 (fluorine-based etch); Cl_2, BCl_3, and $SiCl_4$ (chlorine-based etch); and Br_2 (bromine-based etch). For very aggressive etch chemistry at low pressures (e.g., to etch deep trenches in Si), HBr has also gained popularity. Other new etchant gases include $SiBr_4$, HI, I_2, and even nonhalogenated gases such as CH_4 and H_2. Table 3.13 lists etchants appropriate for etching some common electronic materials.

Table 3.14 presents an overview of the etch rates or etch ratios for a variety of important microfabrication materials in some popular reactive gases and gas mixtures.[40] Examples include etching of photoresist on Si, SiO_2, aluminum, and so on in O_2 plasma and etching of silicon with a metal mask (e.g., Al) in CF_4 plasma. Metal masks influence the etch rate of silicon or even SiO_2 by catalytic action of fluorinated metal surfaces, leading to excess production of free radicals.

Finally, Table 3.15 reviews mask materials listing their suitability for a number of gases.[25] A plus sign in this table corresponds with a low etch rate and high selectivity; a minus sign stands for high etch rate and low selectivity.

For further reading on gas composition for dry etching and stripping, see Flamm,[42–44] Jansen et al.,[2] Peters,[39] and Moreau.[40]

Simplifying Rules

Some simple rules help interpret Tables 3.11–3.15 when specifying a choice of dry etchant and mask for a specific application. These rules are a set of semiempirical observations; they should not be looked on independently, and some state the same phenomenon in slightly different ways.[41]

Table 3.11 Frequently Used Reactive Plasma Gases

Etchant Purpose	Composition (Additive-Etchant): Application
Oxide etchant: etches through oxide to initiate etching	C_2F_6–Cl_2:SiO_2 BCl_3–Cl_2:Al_2O_3 CCl_4–Cl_2:Al_2O_3
Oxidant: increases etchant concentration and suppresses polymerization	O_2–CF4:Si N_2O–CHF_3:SiO_2 O_2–CCl_4:GaAs, InP
Inert gas: stabilizes plasma, dilutes etchant, improves heat transfer	Ar–O_2: organic material removal He–CF_3Br:Ti, Nb
Inhibitor former: improves selectivity, induces anisotropy	C_2F_6–Cl_2:Si BCl_3–Cl_2:GaAs, Al H_2–CF_4:SiO_2 CHF_3–SF_6:Si O_2 (50%)–SF_6:Si
Water and oxygen scavenger: prevents inhibition, improves selectivity	BCl_3–Cl_2:Al H_2–CF_4:SiO_2
Radical scavenger: increases film formation and improves selectivity	H_2–CF_4:SiO_2 CHF_3–SF_6:Si CF_3Br–SF_6:Si CF_2Cl_2–SF_6:Si

Source: Most data from Flamm, D. L. 1993. Feed gas purity and environmental concerns in plasma etching: Part I. *Solid State Technol* October:49–54. With permission.[75]

Table 3.12 Evolution in Gas Mixtures for Dry Etching

Material Being Etched	Conventional Chemistry	New Chemistry	Benefits
Poly-Si	Cl_2 or BCl_3/CCl_4 + sidewall passivating gases Cl_2 or BCl_3/CF_4 + sidewall passivating gases Cl_2 or $BCl_3/CHCl_3$ + sidewall passivating gases Cl_2 or BCl_3/CHF_3 + sidewall passivating gases	$SiCl_4/Cl_2$ BCl_3/Cl_2 $HBr/Cl_2/O_2$ HBr/O_2 Br_2/SF_6 SF_6 CF_4	No carbon contamination Increased selectivity to SiO_2 and resist No carbon contamination Higher etch rate
Al	Cl_2 BCl_3 + sidewall passivating gases $SiCl_4$	$SiCl_4/Cl_2$ BCl_3/Cl_2 HBr/Cl_2	Better profile control No carbon contamination
Al-Si (1%)-Cu (0.5%)	Same as Al	$BCl_3/Cl_2 + N_2$	N_2 accelerates Cu etch rate
Al-Cu (2%)	$BCl_3/Cl_2/CHF_3$	$BCl_3/Cl_2 + N_2 + Al$	Additional aluminum helps etch copper
W	$SF_6/Cl_2/CCl_4$	SF_6 only NF_3/Cl_2	No carbon contamination Etch stop over TiW and TiN No carbon contamination
TiW	$SF_6/Cl_2/O_2$	SF_6 only	
WSi_2, $TiSi_2$, $CoSi_2$	CCl_2F_2	CCl_2F_2/NF_3 CF_4/Cl_2	Controlled etch profile No carbon contamination
Single crystal Si	Cl_2 or BCl_3 + sidewall passivating gases	CF_3Br HBr/NF_3	Higher selectivity trench etch
SiO_2 (BPSG)	CCl_2F_2 CF_4 C_2F_6 C_3F_8	CCl_2F_2 CHF_3/CF_4 CHF_3/O_2 CH_3CHF_2	CFC alternatives
Si_3N_4	CCl_2F_2 CHF_3	CF_4/O_2 CF_4/H_2 CHF_3 CH_3CHF_2	CFC alternatives
GaAs	CCl_2F_2	$SiCl_4/SF_6$ $SiCl_4/NF_3$ $SiCl_4/CF_4$	Fluorine provides etch stop on AlGaAs
InP	None	CH_4/H_2 HI	Clean etch Higher etch rate than with CH_4/H_2

Source: Peters, L. 1992. Plasma etch chemistry: the untold story. *Semicond Int* May:660–70. With permission.[39]

1. **F/C ratio**: During etching, polymerization occurs simultaneously. Etching stems from the fluorine, and polymerization from the hydrocarbons. The dominant process will depend on the gas stoichiometry, reactive-gas additions, the amount of material to be etched, and the electrode potential. Adding hydrogen causes HF to form and the F/C ratio to decrease, leading to more polymerization and less etching (see example in Figure 3.33). A decrease in the fluorine concentration can also occur by overloading the reactor, leading to overconsumption of fluorine and favoring of polymerization. Adding oxygen to a fluorine carbon mixture leads to formation of CO and CO_2 reaction products, increasing F/C, and thus the etch rate, while decreasing the polymerization tendency. In other words, the addition of oxygen to a gas mixture reduces the tendency of Freons™ to polymerize and increases the concentration

Table 3.13 Plasma Etchants for Microelectronic Materials

Material	Common Etch Gases*	Dominant Reactive Species	Product	Comment	Vapor Pressure in Torr at 25°C
Aluminum	Chlorine-based	Cl, Cl_2	$AlCl_3$	Toxic gas and corrosive gases	7×10^{-5}
Copper	Chlorine forms low pressure compounds	Cl, Cl_2	$CuCl_2$	Toxic gas and corrosive gases	5×10^{-2}
Molybdenum	Fluorine-based	F	MoF_6		530
Polymers of carbon, photoresists (PMMA and polystyrene)	Oxygen	O	H_2O, CO, CO_2	Explosive hazard	H_2O = 26 CO, CO_2 > 1 atm
III-V and II-VI compounds	Alkanes			Flammable gas	
Silicon	Fluorine- or chlorine-based	F, Cl, Cl_2	SiF_4, $SiCl_4$	Toxic gas	SiF_4 > 1 atm $SiCl_4$ = 240
SiO_2	CF_4, CHF_3, C_2F_6, C_3F_6	CF_X	SiF_4, CO, CO_2		SiF_4 > 1 atm CO, CO_2 > 1 atm
Tantalum	Fluorine-based	F	TaF_3		3
Titanium	Fluorine- or chlorine-based	F, Cl, Cl_2	TiF_4 TiF_3 $TiCl_4$		$TiF_4 = 2.10^4$ $TiCl_4$ = 16
Tungsten	Fluorine-containing	F	WF_6		1000

*Common chlorine containing gases: BCl_3, CCl_4, Cl_2, and $SiCl_4$. Common fluorine containing gases: CF_4, SF_4, and SF_6.

Source: From Flamm, D. L. 1993. Feed gas purity and environmental concerns in plasma etching: Part I. *Solid State Technol* October:49–54; Elwenspoek, M., H. Gardeniers, M. de Boer, and A. Prak. 1994. Report No. 122830, University of Twente, Twente, the Netherlands.; Moreau, W. M. 1988. *Semiconductor lithography.* New York: Plenum Press; and Hess, D. W., and D. B. Graves. 1989. *Microelectronics processing: chemical engineering aspects.* Eds. D. W. Hess, and K. F. Jensen. Washington, DC: American Chemical Society.[75,25,40,41]

of halogen etchants ensuing from these gases. Adding oxygen to improve the F/C ratio leads to aggressive resist etching. Gases such as NF_3 and ClF_3 allow high fluorine concentrations (for aggressive Si etching) without the addition of oxygen, thus avoiding resist attack.

2. **Selective versus unselective dry etching**: The polymerizing point of the gas (i.e., the composition of the gas where polymerization takes over) primarily determines selectivity. The closer one works to the polymerization point, the better the selectivity. Factors that

Table 3.14 Frequently Used Reactive Plasma Gases, Reported Etch Rates, and Etch Ratios

Film/Underlayer (F/U)	Etch Ratios
Si_3N_4 over AZ-2400 resist	CF_4-10
SiO_2 over AZ-1350 resist	CF_4/H_2-10
Poly-Si over PBS resist	CF_4-15
PSG over AZ-1350	SF_6-10
SiO_2 over Si underlayer	CF_3Cl-30
Si_3N_4 over SiO_2 underlayer	NF_3-50
Poly-Si over SiO_2 underlayer	CCl_4-10
Si over SiO_2 underlayer	SF_6-30
Etchant	Material and Etch Rate (Å/min)
Ar	Si-124, Al-166, Resist-185, Quartz-159
CCl_2F_2	Si-2200, Al-1624, Resist-410, Quartz-533
CF_4	Si-900, PSG-200, Thermal SiO_2-50, CVD SiO_2-75
C_2F6-Cl_2	Undoped Si-600, Thermal SiO_2-100
CCl_3F	Si-1670

Source: Moreau, W. M. 1988. *Semiconductor lithography.* New York: Plenum Press. With permission.[40]

Table 3.15 Mask Materials in Dry Etching

Mask Material	SF_6	CHF_3	CF_4	O_2	N_2
Si	–	–\+	–	+	+
SiO_2	+\–	–	+\–	+	+
Si_3N_4	+\–	–	+\–	+	+
Al/Al_2O_3	+	+	+	+	+
W	–	–	–	+	+
Au	+	+	+	+	+
Ti	–	–	–	+	+
Resist	+\–	+\–	+\–	–	+
CFs	+	+	+	+	–

Source: Elwenspoek, M., H. Gardeniers, M. de Boer, and A. Prak. 1994. *Micromechanics*. Report No. 122830, Twente, the Netherlands: University of Twente.[25]

have a tendency to increase the polymerization rate increase selectivity. Decreased temperature, high hydrogen concentration, low power, high pressure, and high monomer concentration all increase polymerization and thus selectivity. Typical unselective etchants for Si and polysilicon, used for noncritical etching, are CF_4 and CF_4-O_2. Less toxic gases such as CF_4 and SF_6 have preference over fluorine and are more selective, leading to polymerization. The most commonly used gases for selective Si etching are Cl_2, CCl_4, CF_2Cl_2, CF_3Cl, Br_2, and CF_3Br, along with mixtures such as Cl_2-C_2F_6. Small additions of halogens significantly increase the selectivity of fluorine-based etchants, especially the selectivity of silicon over silicon dioxide. For pure chlorine, the etch rate of SiO_2 is so low that even a native oxide can prevent the Si from etching. The high selectivity of SiO_2/Si is a major objective of many plasma processes trying to mimic the wet HF etch efficiency in this regard.

3. **Substrate bias**: A negative bias on a surface exposed to the plasma increases, at a constant F/C ratio, the etching behavior over polymerization tendency. This effect is caused by the enhanced energy of the ions striking the surface, resulting in polymer sputtering.
4. and 5. **Dry etching of III–V compounds**: Group III halides, particularly fluorides, tend to be involatile. As a result, F plasmas are usually substituted for chlorine-containing plasmas, and elevated substrate temperatures are used to further help volatilize the chlorides (fourth rule). The chemical composition (Ga/As ratio) of the different atomic planes in GaAs varies, and crystallographic etch patterns are observed under etch conditions in which chemical processes dominate (fifth rule). The latter establishes a major difference with dry Si etching, whereas crystallography does not play a role at all.
6. **Metal etching**: Chlorocarbons and fluorocarbons are typically used to etch metal films because they can reduce native metal oxides chemically. Oxygen and water vapor must be rigorously excluded during etching because of the high stability of the metal-oxide bond. Also, because of the stability of such bond, ion bombardment is essential. Because AlF_3 is not volatile, chlorine-containing gases are preferred for Al etching, whereas tungsten can be etched in fluorine (SF_6 or NF_3/chlorine).
7. **Organic films**: To retain an organic mask while etching away materials such as SiO_2 and Si_3N_4, one must work in etching conditions close to the polymerization point so that some loss of resist is compensated by condensation of reaction product, for example, CF_4-C_2H_4 and CF_4-H_2. We have already discussed that CF_4-O_2 plasmas severely degrade resist materials.
8. **Carbon-containing additives**: Carbon-containing additives generally degrade the selectivity of Cl- and Br-based inorganic chemistries for polysilicon etching over SiO_2. The Br atom attack on SiO_2 is thermodynamically unfavorable, whereas reactions between carbon halogen-bonded species and SiO_2 are exothermic. With HBr, a selectivity of 300:1 of polysilicon

over silicon dioxide can be achieved when completely avoiding carbon-containing gases or photoresists; the presence of resists or carbon traces in gases severely diminishes this ratio.

Leading suppliers of plasma etching equipment are Applied Materials, Inc. (Santa Clara, CA; http://www.appliedmaterials.com) and Lam Research Corporation (Fremont, CA; http://www.lamrc.com).

Deep Reactive Ion Etching

Introduction—Need for High-Density Plasmas

For building high-aspect-ratio micromachines and nanomachines with dry etching, until the mid-1990s, one had to battle three major problems. The first was the low etch rates with which to contend; typical Si dry etch rates of 1 µm/min required more than 5 hours to etch 300 µm deep! A second impediment was the inability to maintain high-aspect ratios over etch depths of a few tens of microns. Third, masking layers performed poorly and could not protect during a long etch.[45] New, high-density plasma sources provide an answer to each of these problems. As a result, for many applications, deep reactive ion etching (DRIE) presents an attractive MEMS solution today. An important example is the fabrication of high-aspect-ratio micromolds, which may be made much less expensive with deep DRIE than fabricating them with LIGA—an alternate high-aspect-ratio technique (see Chapter 10).

Ideally, dry etching stations generate high plasma densities (>10^{11}/cm^3) to achieve a high etch rate while operating at low pressure (1–20 mTorr). Low pressure increases ion directionality and discourages microloading. Normally low pressure also means less ionization and low plasma density. However, as we saw earlier, the use of magnetic fields forces electrons to spin with a very small gyroradius. This makes for a longer travel distance and thus more chance for electrons to collide with atoms, making a higher plasma density at low pressure possible. The source also needs to produce ions uniformly in energy and distribution while keeping the energy of the ions low to prevent damaging active electronics. The magnetic field reduces the charge difference in the sheath region, lowering the DC bias and therefore reducing the ion energy.

The two high-density plasma sources discussed here are inductively coupled plasma (ICP) stations [also called transformer coupled plasma (TCP)] and electron cyclotron resonance (ECR) machines. Both systems work at low pressures (a few mTorr) and can independently control ion flux and ion energy. In Figure 3.36 we show a schematic of an ICP and an ECR station.

Inductively Coupled Plasma

In inductively coupled plasmas, a coil, operating at the standard 13.56 MHz, drives the

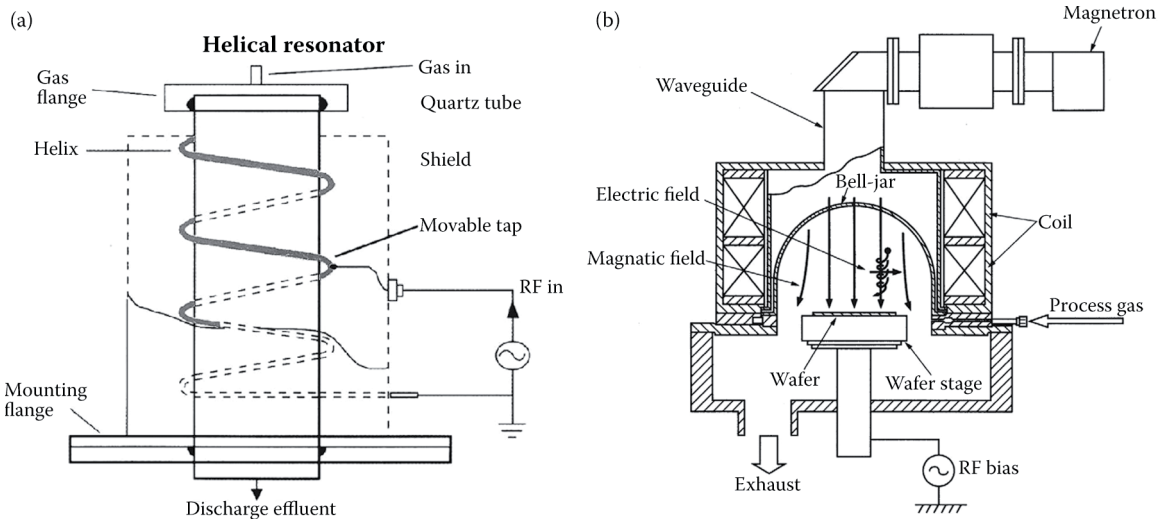


FIGURE 3.36 (a) ICP station. (b) ECR station.

plasma inductively (Figure 3.36a). Such an ICP source creates high-density, low-pressure, low-energy plasmas by coupling ion-producing electrons to the magnetic field arising from the rf voltage. Electrons travel a long way without collisions with the chamber walls or electrodes. The upper part of the ICP chamber, where the plasma is generated, is either ceramic or quartz. The RF source in the upper chamber generates the plasma and controls the ion density, whereas the RF bias in the lower chamber controls the ion bombardment energy. Thus, the ion energy and ion density are effectively controlled independently. By cooling the wafer chuck to liquid nitrogen temperatures (77 K) and using a helium gas flow under the wafer for efficient heat transfer, wafer temperature can be maintained at cryogenic temperatures during etching.[46] The cryogenic cooling of the chuck results in condensation of reactant gases and protects the sidewalls from etching, which renders a more anisotropic process.

The Alcatel RF ICP source is one of the systems advanced as an ideal source for deep anisotropic Si etching. The electron density in this plasma source reaches $>10^{12}/cm^3$, and using fluorine-based gases, silicon etches at rates of up to 6 µm/min with etch uniformity better than ±5% while maintaining a Si/SiO_2 selectivity of more than 150:1. Etch depths greater than 250 µm and profile angles of ±1° were demonstrated while maintaining aspect ratios of 9:1.

Using an ICP etching station from Surface Technology Systems (STS), Klaassen et al.[45] were able to etch 300 µm deep in Si at 5 µm/min using a photoresist layer as mask of a 6-µm thickness only. Polymerization of the photoresist mask on the sidewalls of the etched trenches slows the lateral etch rate and results in high anisotropy. The Si to photoresist selectivity was determined at about 50:1. However, cryogenic cooling (e.g., –100°C) may introduce undesirable thermal stresses, and an important, much-touted alternative is the "Bosch process" patented by Robert Bosch GmbH (Stuttgart, Germany).[47]

The Bosch advanced silicon etch (ASE) typically achieves high-aspect ratios of 20–30:1, and etch rates of ~3 µm/min are standard.[47] In this process carried out, for example, in an STS or PlasmaTherm high-density plasma system, one switches (pulses) between sequential passivation (85 sccm C_4F_8, 8 s, 0 W platen, 600 W coil, 16 mTorr) and etching (80 sccm SF_6, 6 s, 8 W platen, 600 W coil, 33 mTorr) (see Table 3.16). In the passivation step, a ~10-nm-thick fluorocarbon deposits and energetic SF_x^+ ions remove this layer at the bottom of a trench, as sketched in Figure 3.37.

TABLE 3.16 Characteristics of the DRIE Process Used in the STS and PlasmaTherm Systems

SF_6 flow	30–150 sccm
C_4F_8 flow	20–100 sccm
Etch cycle	5–15 s
Deposition cycle	5–12 s
Pressure	0.25–10 Pa
Temperature	20–80°C
Etch rate	1.5–4 µm/min
Sidewall angle	90° ± 2°
Selectivity to photoresist	~100 to 1
Selectivity to SiO_2	~200 to 1

Source: Maluf, N. 2000. *An introduction to microelectromechanical systems engineering.* Boston: Artech House.[48]

The repetitive alternation of etch and passivation steps results in a very directional etch. Some scalloping near the top of the trench is observed, but in general the sidewalls exhibit good surface planarity with roughness less than 50 nm.[48] The scalloping effect is illustrated in Figure 3.38. The experiment depicted in this figure also demonstrates that changing the length of the sequential passivation and etching cycles can control scalloping.

Using the Bosch process, Chow et al.[49] fabricated through-wafer electrical interconnects by dry etching 20-µm holes through a 400-µm wafer (see Example 3.1). By etching from both sides, the required aspect ratio is reduced by two, from 40:1 to 20:1. Each etch takes 180 minutes.

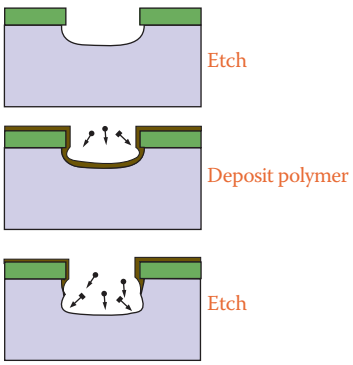


FIGURE 3.37 Illustration of the Bosch process.

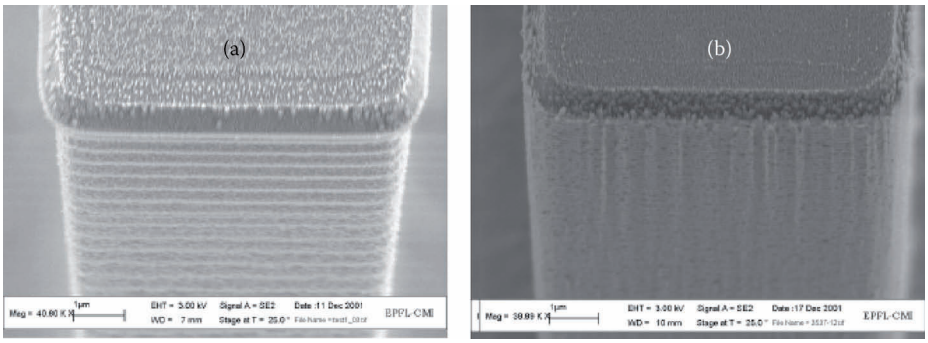


FIGURE 3.38 Sidewall roughness at the top of a deep anisotropic etching of Si (Bosch process) as a function of pulse duration: SF_6/C_4F_8 = 7s/2s (a) and SF_6/C_4F_8 = 3s/1s (b).

Ayon et al.[26] tackled the issue of microloading in the Bosch process by increasing the ion flux through control of the applied source coil power. Ions reaching the wafer surface during the etch cycle promote more efficient removal of the protective polymer coating at the bottom of the trench, obviating the microloading effect. To etch ultradeep anisotropic silicon trenches in the range of 300–500 µm, Ayon et al.[26] compensate the reduction of neutral species flux reaching the feature bottom by increasing the SF_6 flow rates during the etching cycle (thus replenishing the etching species while the higher pressure also facilitates the removal of etching byproducts). This method results in a net improvement in passivation and etch rates and is known as the MIT 59 method (Figure 3.39). The method has already been applied in processes like the fabrication of accelerometers, gyros, microdisk-drive armatures, SCALPEL mask membranes, microgas turbines, tethered motors, power MEMS, and, in general, for etching through wafers.[26,50] A deep silicon back-side Bosch etch also allows fabrication of high-aspect-ratio and MEMS devices integrated with complementary metal oxide semiconductor (CMOS) circuitry (see Example 3.3).[51]

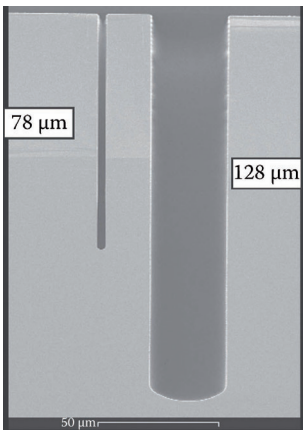


FIGURE 3.39 MIT 59 method. See text.

An extensive database for the Bosch process conditions required to achieve prescribed profiles, as well as targeted silicon etch rates and uniformity, was presented in the *Journal of the Electrochemical Society* by Ayon et al.[52,53]

Electron Cyclotron Resonance

New, high-density plasma sources may also be based on microwave discharges (Figure 3.36b). An RF source controls the ion energy, and a microwave power source controls the ion flux. Microwaves are electromagnetic waves in the 1–100-GHz (corresponding to wavelengths from 30 to 0.3 cm) range. Waves in this frequency range are transmitted by the use of waveguides, which are typically made of metal tubes of rectangular or circular cross-section. Cavities with metallic conducting walls resonate at specific frequencies and harmonics that are easy to calculate. The energy dissipated within the cavity dictates the frequency range above and below the resonant frequency that the cavity will accept and is characterized by the quality factor Q:

$$Q = \frac{f_0}{2\Delta f} \tag{3.21}$$

where $\pm\Delta f$ are the frequencies at which the amplitude is one-half that at the resonant frequency, f_0, for the same input power. In the case of an empty waveguide, the dissipation of energy inside the cavity is determined by the skin depth, δ, of the electrical

field into the metal walls, which is given by (see also Volume I, Chapter 5, Equation 5.215):

$$\delta = \sqrt{\frac{\rho}{\pi \mu f}} \tag{3.22}$$

where ρ is the resistivity in Ωm, μ the permittivity of free space ($4\pi \times 10^{-7}$ Henry/m), and *f* the frequency in hertz. For the generation of microwave plasmas, a cavity in the form of a cylinder with radius *a* and length *l* is generally used operating in $TM_{0,1}$ mode, that is, with the electric field parallel to the axis of the cylinder. In this case:

$$\lambda_0 = 2.61a \tag{3.23}$$

where λ_0 is the microwave wavelength. The resonant frequency of the cavity changes as a result of the change in the dielectric constant of the ionized gas, and *Q* decreases because of the increased power loss in generating the plasma.

In a quartz chamber filled with a gas at low pressure (0.1–10 Torr), the microwave power will cause breakdown when free electrons within the quartz chamber gain sufficient energy to ionize the gas. A static magnetic field generated by solenoid coils is also applied to the quartz chamber, and electrons move in a circular motion at the ECR frequency for that magnetic field (see Equation 3.18). In an ECR etching station, the ECR discharge creates a region, or surface layer, within the quartz chamber where the electron cyclotron frequency and the microwave frequency are equal and the electric field component is perpendicular to the static magnetic field. Under these conditions, the electrons within the ECR region are accelerated until they lose energy by an ionization or excitation collision with a neutral gas molecule. Varying the input microwave power and the gas pressure controls the plasma density. The maximum plasma density is reached if the microwave input power is such that the plasma frequency is equal to the microwave frequency, which is itself equal to the ECR frequency. The interaction of the magnetic field and the microwaves results in an intense, high-density plasma maintainable at low pressure. In operation, a 2.45-GHz microwave frequency is generated by a magnetron and injected into an etching chamber, enclosed by a quartz bell jar, through a waveguide. The maximum number of electrons (and ions) that can be formed within the quartz chamber at 2.45 GHz is about 7.5×10^{10} electrons/cm^3.

In many ECR systems, the wafers are placed downstream of the plasma to further limit their exposure to this intense discharge. Wafers can also be biased to control ion bombardment energy. Like MERIE, ECR and ICP produce higher plasma densities with low-energy density, reducing charge-up damage.

Juan et al.[54] used ECR to etch polyimide molds for the fabrication of electroplated microstructures in one of several reported "pseudo-LIGA" processes (see also Chapter 10). A fast polyimide etch rate of 0.91 μm/min and a high selectivity over a Ti etch mask of 3150:1 were reached in an oxygen plasma. Polyimide (Dupont Pyralin® PI-2611) was spun on in multiple coatings and cured at 380°C for 1 hour. According to the authors, etching with an ECR source is 10 times faster and six times more selective than conventional RIE. The combination of magnetic confinement and microwave power, leading to an order of magnitude greater dissociation efficiency at much lower pressures than RIE, reduces scattering of reactive species and promotes vertical profiles and smooth surface morphology. Cooling down to –130°C helps maintain a vertical etch profile by suppressing spontaneous chemical reactions. This way, lateral resonator elements 32-μm thick with a 1-μm gap have been fabricated in polyimide, and electroplated Ni accelerometers with an aspect ratio of 11:1 were formed from these polyimide molds. Compared with LIGA, where the average wall roughness is <50 nm, the walls produced by this process still appear rough, but for many applications they might be adequate.

For a review on etching high-aspect-ratio trenches in silicon by DRIE, we refer to Jansen et al.[2] This article features an excellent review of possible trench shapes. Equipment for DRIE is available from STS (Abercarn, Wales, UK; http://www.stsystems.com), Plasma Therm Inc. (St. Petersburg, FL; http://www.plasmatherm.com/index.html), and Alcatel, S.A. (Paris, France; http://www.alcatel.com).

Vapor-Phase Etching without Plasma (XeF_2)

As far back as 1983, it was demonstrated that etching of Si with XeF_2 in a CAIBE reactor enabled

sidewall tailoring by changing the partial pressure of XeF_2, the ion energy, and the current density.[55] At the higher partial pressures of XeF_2, isotropic etching and undercutting became predominant. Isotropic etching of Si with xenon difluoride (XeF_2) actually does not require a plasma to generate the etching species. XeF_2 is a white solid with a room temperature vapor pressure of about 4 Torr, which reacts readily with Si. The Si etch occurs in the vapor phase at room temperature and at pressures between 1 and 4 Torr, established by a vacuum pump throttled to the right pressure. The etch rate is very dependent on the feature size in Si to be etched (see microloading). Hoffman et al.[56] observed silicon etch rates as high as 10 µm/min but worked at more typical rates of 1–3 µm. The extreme selectivity of XeF_2 to silicon over silicon dioxide and silicon nitride are well documented, and Hoffman et al.[56] have shown that XeF_2 also displays extreme etching selectivity over aluminum and photoresist. One important problem is that XeF_2 reacts with water to form water and HF. The latter may unintentionally react with Si to form SiO_2. The exothermic etch reaction, according to Chang et al., is given by:

$$2XeF_2 + Si \rightarrow 2Xe + SiF_4 \qquad \text{Reaction 3.5}$$

with only Si in the solid phase.

A 50-nm Al mask or a single layer of hard-baked photoresist suffices as deep etch masks. The simplicity of the process, the fast etch rate, as well as the resistance of even very thin layers of oxide, nitride, and Al metal make this etching process a possible choice for etching micromachined structures in the presence of CMOS electronics (in a "post-CMOS procedure"). The etched surfaces have a granular structure (10 µm and smaller feature size), and this etchant is not suited when polished mirror surfaces are required. In Figure 3.40a, an Al hinge 5 µm wide and 1.1 µm thick, holding an oxide

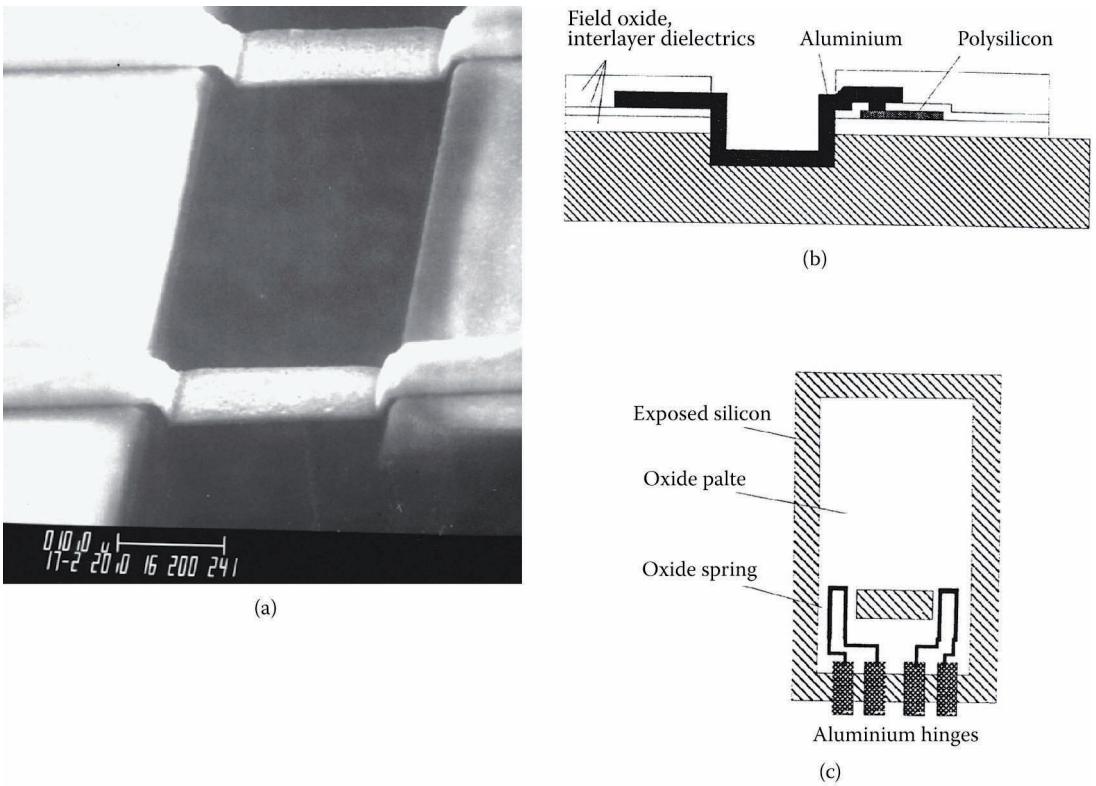


FIGURE 3.40 (a) Al hinge (5 µm wide and 1.1 µm thick) holding a suspended oxide plate (plate is on the right side of the photograph). The Si is etched from underneath the Al hinge and oxide plate by XeF_2. The suspended oxide plate is a 200-µm square. (Courtesy of Dr. Kris Pister.) (b) Al hinge contacts a poly-Si piezoresistor embedded in the oxide plate to protect it from the XeF_2 etchant. (c) The piezoresistive accelerometer design shown enables rotation of the oxide plate out of the plane to detect orthogonal acceleration. (From Hoffman, E., B. Warneke, E. Kruglick, J. Weigold, and K. S. J. Pister. 1995. *Proceedings: IEEE micro electro mechanical systems (MEMS '95).* Amsterdam, the Netherlands: IEEE. With permission.[56])

plate 200-μm square suspended over a Si etch pit, is shown. The etch pit underneath the suspended plate is nearly isotropic and exhibits a surface roughness of several microns. The conductivity of the Al metal was unaffected by the XeF_2 etch. The hinge contacts a polysilicon strain gauge, sandwiched between oxide layers, to protect it from the XeF_2 etchant (see Figure 3.40b). The flexible hinge enables rotation of the oxide plate out of the plane of the wafer. Hoffman et al.[56] suggest that this contraption might function as a piezoresistive accelerometer, with hinges providing mechanical support and electrical connectivity between strain gauges embedded in the oxide plate proof mass and the wafer (see Figure 3.40c).

The rough Si surfaces of a XeF_2 etch in principle may be avoided by using interhalogen gases such as BrF_3 and ClF_3 with xenon gas as diluent.[57] The interhalogens are formed from single-element feed gases (e.g., bromine and fluorine) and result in perfectly isotropic etches and smooth silicon surfaces. Wang et al.[58] measured etch rate ratios for thin films with respect to Si in BrF_3 for several materials:

1. Si:LPCVD silicon dioxide = 3000:1
2. Si:silicon nitride = 400–800:1 (depending on the Si concentration)
3. Si:hard-baked AZ4400 and AZ1518 photoresist = 1000:1
4. Si:Al, Cu, Au, and Ni ≥ 1000:1

A first commercial instrument developed for XeF_2 etching is the Xetch™ from Xactix (http://www.xac tix.com).

Dry Etching Models—In Situ Monitoring

A serious impediment to more effective use of dry etching is the large number of parameters that affect the process.[59] Just consider pressure and temperature. Pressure influences the ion-to-neutral ratio and fluxes of these particles, the sheath potentials and energy of the ions bombarding surfaces, electron energy, and the relative rate of chemical kinetics. Temperatures, both of gas and surface, have a profound influence on discharge chemistry as well. Only the surface temperature is controllable and influences selectivity, etch rates, degradation of resist masks, and surface roughness (higher temperature leads to a rougher surface).

Also, electrode area ratio, RF frequency, and power will alter plasma and electrode potentials, thereby changing ion energies. At lower frequencies, where ions can respond directly to the oscillating field, ions can attain the maximum energy corresponding to the maximum field across the sheath. As a result, for a constant sheath potential, ion bombardment energies are higher at frequencies less than 50 kHz. Nonuniformity can be caused by gradients in etchant concentration. Such gradients are responsible for the dreaded bull's eye etching pattern, where the etch rate decreases monotonically from the wafer periphery to its center (see above). Pressure, power, and gas flow must be adapted until the specified uniformity is obtained.

Predicting the exact relation between the flux and the energy of particles striking a wafer surface and the etch rate still is not state of the art, largely because no equilibrium is attained in these dry etching reactors, and for most reactions the kinetics are not known. Consequently, in situ monitoring of the plasma to improve etching uniformity is an important area for further research.[60,61] Besides measuring these controlling parameters, techniques such as optical emission spectroscopy (OES) to map in situ etchant concentration are coming to the foreground as sensitive and effective endpoint detectors. It is well known that the color of a plasma changes when etching different materials. Emissions can emanate from etchants, etch products, or their fragments. An optical spectrometer is set to the line of a reaction product of interest, and one follows its intensity during an etch cycle. Some etch endpoint wavelengths are summarized in Table 3.17.

Mass spectrometry allows one to monitor the concentration of a wide variety of species; it is limited, however, in that the results are obtained downstream from the real process in the reactor.[62] Laser-induced fluorescence (LIF) has been used to measure ground–state ion and radical concentrations. LIF is more powerful than OES, but it is expensive and requires a lot of experimental expertise; consequently, it is used more in the research lab than on a production line.[63] In imaging

Table 3.17 Etch Endpoint Wavelength

Film	Etchant	Wavelength (Å)	Emitter
Al	Cl_2, BCl_3	2614 3962	AlCl Al
Poly-Si	Cl_2	2882 6156	Si O
Si_3N_4	CF_4/O_2	3370 3862 7037 6740	N_2 CN F N
SiO_2	CF_4 and CHF_3	7037 4835 6156	F CO O
PSG, BPSG	CF_4 and CHF_3	2535	P
W	SF_6	7037	F

of radicals interacting with a surface (IRIS), light from a dye laser optically excites specific radicals in the incident and reflected molecular beam striking a surface. The light emission of radicals in incident and reflected beam is measured and used to extract information on their density. By taking the difference of the emission data, it is possible to determine the intensity of scattered radicals and the reactivity of specific radicals with particular surfaces.[3] Laser interferometry (for thin transparent films such as polysilicon, silicon dioxide, and silicon nitride) and laser spot reflectance (for opaque, highly reflecting metal films) are used to monitor film thickness and film disappearance, respectively. Also, the temperature of the wafer during reactive ion etching affects process quality in terms of photoresist integrity, selectivity, etch rate, and etch residues. An in situ temperature monitor enables one to improve the performance during the plasma etch.[64] With Langmuir probes, one can, in principle, determine the complete electron energy distribution of a plasma. In such an experiment, a small wire probe is inserted into the plasma and biased with respect to some other electrode. By recording the current-voltage plot of the probe, charged particle densities and energies can be deduced. In practice, Langmuir and other electrostatic probes are difficult to apply, invasive, and perturb the plasma.

Bosch-Charpenay et al.[50] integrated an FTIR sensor in an STS Multiplex ICP deep reactive ion etcher. This enabled in situ, real-time measurements of DRIE etch depth and eroding photoresist thickness measurements of MEMS devices using infrared reflectance analysis. Typical capabilities of the system are summarized in Table 3.18.

Examples

Several examples of dry-etched micromachines are presented below. They include deep dry etching to fabricate interconnects through single crystal Si, an example of single crystal reactive etching and metallization (SCREAM), which relies on a combination of anisotropic and isotropic dry etching, a variant on the SCREAM process from the Carnegie Mellon University, dry etching of polymeric materials, and a cantilever micromachining process combining wet and dry etching.

Table 3.18 Summary of FTIR Sensor Capabilities

MEMS structures that can be measured	Pillars, holes, trenches, trench grids, honeycombs of etched width less than ~100 μm
Etch depth precision (reproducibility)	0.05–0.5 μm depending on the structure
Etch depth accuracy (as compared with SEM)	~1–2 μm (estimated)
Maximum measurable etch depth	>200 μm
Current spot size*	~800 μm diameter
Measurement time	~1 s
Measurement spot range†	±5 mm in both *x*- and *y*-direction
Probing IR wavelength	2–10 μm (1000–5000 cm^{-1})

*Optical diffraction limits preclude spot sizes smaller than ~400 μm. MEMS devices smaller than this minimum spot size can, however, still be measured, provided they are geometrically isolated from other DRIE etched features.
†Translation range of infrared spot over sample is strictly related to equipment geometry at the etch chamber—in this case a small overhead view port into the STS Multiplex ICP system.
Source: Bosch-Charpenay, S., J. Xu, J. Haigis, P. A. Rosenthal, P. R. Solomon, and J. M. Bustillo. 2000. *Technical Digest Solid-state sensor and actuator workshop.* Hilton Head Island, SC: Transducers Research Foundation, Inc.[50]

Example 3.1: Via Etching in Si

Fabrication of high-density electrical, through-wafer interconnects on silicon substrates remains a challenging proposition. In MEMS, through-wafer interconnects greatly facilitate their integration with active electronics (see Making Vias through Silicon in Volume III, Chapter 4). Different approaches to fabricating those 3D contact lines are compared in Volume III, Chapter 4 on packaging. Chow et al.[49] use the Bosch process to etch a high-aspect-ratio through-wafer via and subsequently coat an insulator and then fill in a signal-conducting layer.[49] Etching the via is the first and most critical step. A 10-μm-thick photoresist (Shipley SPR220-7) is coated on both sides of the Si wafer, and a double-sided aligner is used to expose 20-μm circles on both sides (±1 μm accuracy). The authors let the wafers then sit for 8 hours to absorb water. The moisture, they claim, helps ensure a shorter development time (90 s in Shipley LDD26W), resulting in straighter resist walls. Using a high-density plasma etcher from STS, a Bosch etch procedure (such as shown in Table 3.16) is performed for 120 minutes from the front side of the wafer. Resist is then coated and baked again, and another 180-minute etch from the back is used to punch through the vias (Figure 3.41). The diameter of the resulting vias varies from 24 μm (top and bottom) to 18 μm in the middle. By etching from both sides of the wafer, the required aspect ratio is reduced from 40:1 to 20:1, and it is possible to etch a 20-μm circle through a 400-μm-thick wafer. For electrical isolation of the via, thermal silicon dioxide (2 μm) is grown at 1100°C. For signal conduction, low-pressure chemical vapor deposition (LPCVD) polysilicon is deposited into the via and doped by phosphorous diffusion. The resistance of the resulting via is 14 Ω, and its low capacitance (<1 pF) makes it attractive for high-speed applications.

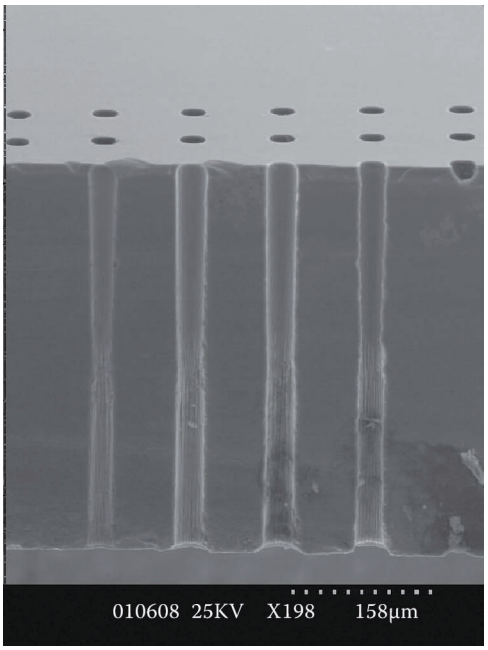


FIGURE 3.41 SEM cross-section after plasma etching vias through the wafer. (From Arnold, J. C., and H. H. Sawin. 1991. Charging of pattern features during plasma etching. *J Appl Phys* 70:5314.[24])

Example 3.2: Single Crystal Reactive Etching and Metallization

As a second example of dry etching in micromachining, we follow some of the work by the Cornell Nanofabrication Laboratory making laterally driven micromachines.[65,66] In laterally driven micromachines, as illustrated in Figure 3.42a, beams and structures of various shapes should preferably be of a narrow and thick construction and able to move freely and in parallel to the supporting substrate while being held rigidly above it. Increased thickness of beams permits greater driving power for a given voltage and increases stiffness of the device perpendicular to the plane of motion. Thin, suspended beams (e.g., 2 μm wide) are fabricated with a combination of dry and wet etching in polysilicon (surface micromachining, Chapter 7). However, thick single crystal beams greater than 20 μm in materials such as quartz, Si, and GaAs can best be made with dry etching, especially because thick lateral moving microstructures of arbitrary shape, given crystallographic limitations, are very difficult to realize using wet bulk micromachining (Chapter 4). The monitoring of the resonant frequency variation of small resonator devices made this way (see Figure 3.42a) can be used to measure monolayer mass changes, as well as pressure, temperature, acceleration, and so on. In the measurement of mass, for example, the sensitivity of the device is a function of resonant frequency and mass of the resonating structure:

$$\text{Sensitivity} \sim \frac{\text{Resonant frequency}}{\text{Mass}} \qquad (3.24)$$

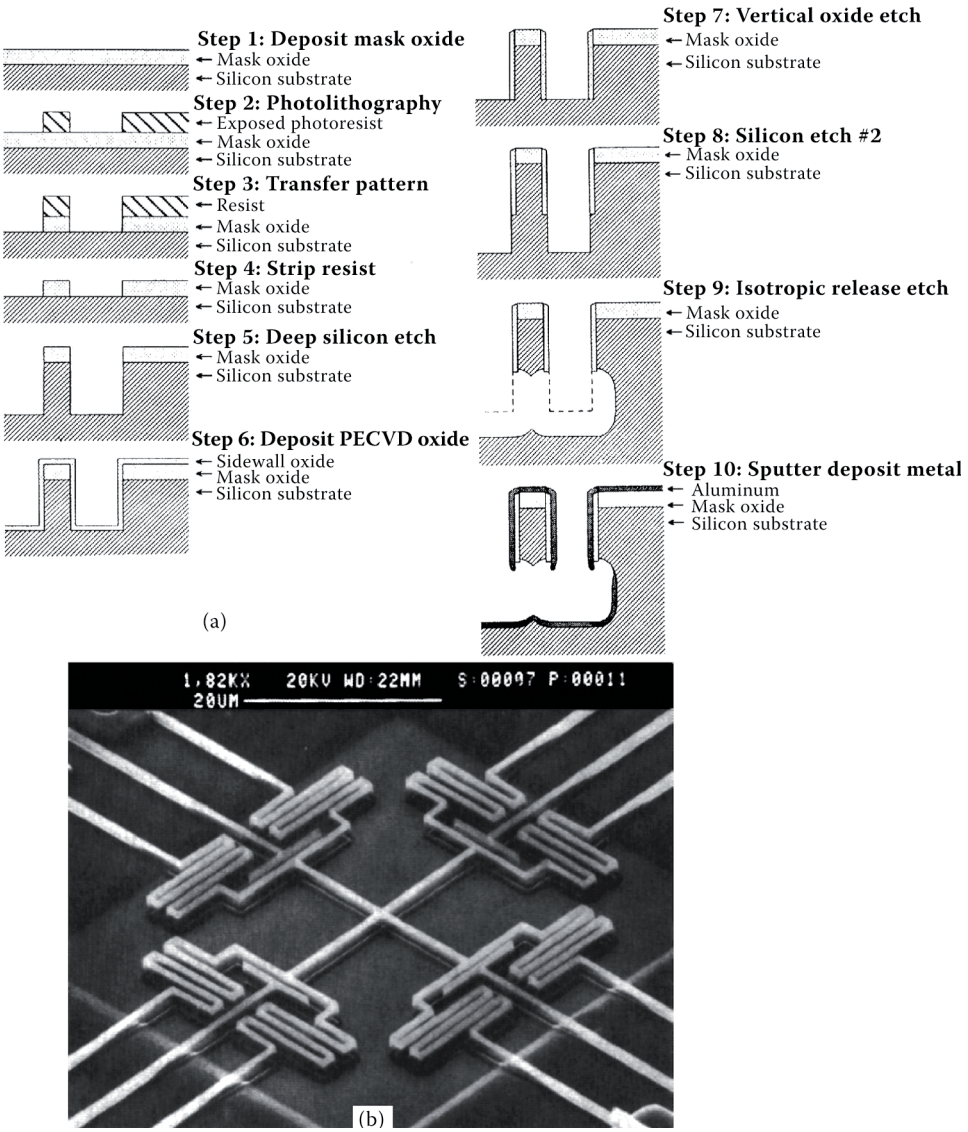


FIGURE 3.42 (a) Si-SCREAM process (see text). The cross-section of a typical beam made using the SCREAM process is illustrated in step 10. This drawing shows a suspended beam and its associated parallel-plate capacitor. The released beam is free to move, whereas the plate on the right is static and can be used to measure the motion of the beam. (From K. A. Shaw et al., *Sensors Actuators A*, 40:63–70 (1994). With permission.) (b) A representative SCREAM structure. SEM micrograph of a Si *x,y* stage with parallel-plate capacitive drives and springs. (From Yao, J. J., S. C. Arney, and N. C. MacDonald. 1992. Fabrication of high-frequency two-dimensional nanoactuators for scanned probe devices. *J Microelectromech Syst* 1:14–22. With permission.[67])

To build the most sensitive detector, the mass must be minimized and the resonant frequency maximized. The mechanical resonant frequency of an ideal resonator is given by:

$$f_0 = \left(\frac{k}{m}\right)^{\frac{1}{2}} \tag{3.25}$$

with k the spring constant and m its mass; in other words, the resonant frequency is proportional to (stiffness/mass)$^{1/2}$. A large spring stiffness must be designed and m minimized to make f_0 as high as possible. For some of the very stiff, low mass structures made at Cornell, f_0 was 5 MHz, with k = 55 N/m and m = 2.0×10^{-13} kg. The above simple considerations illustrate the power of micromachining in enabling more sensitive detectors. For more details on building MEMS resonant beams, see Volume III, Chapter 8 on actuators.

Lateral resonant structures as shown in Figure 3.42a are used to make sets of comb electrodes,

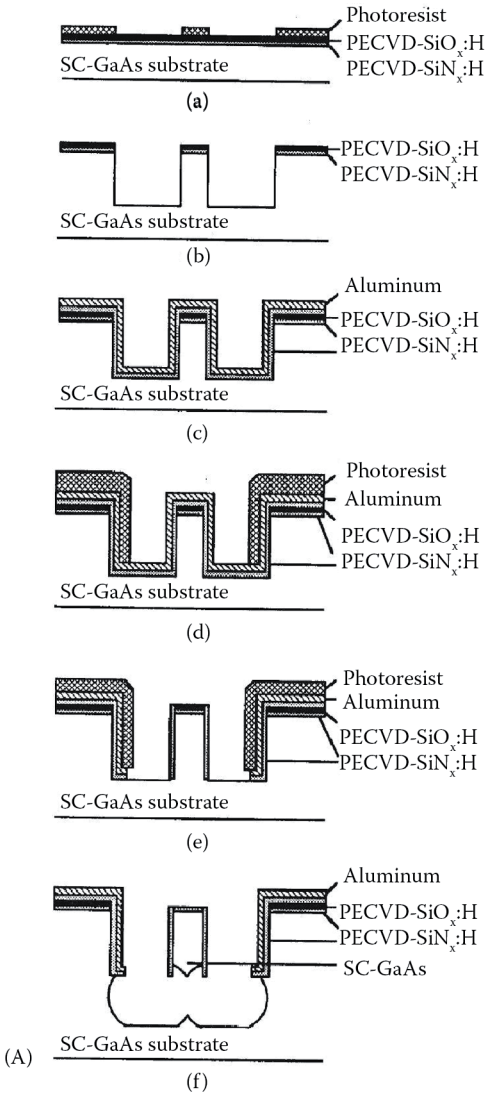
for example, to drive precision *x-y* microstages for accurate positioning or as comb drives in accelerometers. Scanning tunneling microscopes (STMs), 40 × 40 μm in size, have been manufactured using SCREAM technology.[67] In the latter case, field emitter tips were integrated on an *x-y* stage with moving beams with nominal cross-sectional dimensions of 250 nm in width and 1000 nm in height. Displacements in the *x-y* directions through parallel-plate capacitive drives measured ±200 nm for an applied voltage in the 50-V range.

An elegant Cornell dry etching process that obtains such laterally driven, single crystal micromachines is called *SCREAM*. This technology uses reactive ion etching to both define and release structures. SCREAM portrays an important technique from several points of view. It is a self-aligned, single-mask process, run at temperatures lower than 300°C, completed in less than 8 hours, and can be carried out in the presence of active circuitry integrated on the same chip. The example given here is the fabrication of single crystal silicon-released beam structures (see Figure 3.42).[67]

The steps of the SCREAM process are illustrated in Figure 3.42b. As a starting substrate, an arsenic-doped, 0.005-Ωcm, n-type (100) silicon wafer is used. A masking layer of PECVD silicon dioxide, 1–2 μm thick, is coated with resist (step 1). Photolithography on the photoresist layer creates the desired pattern (step 2) to be transferred onto the silicon dioxide using CHF_3-based, magneton reactive ion etching (MERIE) (step 3). The photoresist is then stripped in an oxygen barrel etcher (step 4). The silicon dioxide mask pattern is subsequently transferred to the silicon substrate using a Cl_2/BCl_3 anisotropic etch. The etch depth, depending on the intended micromachine, ranges between 4 and 20 μm (step 5). Following the silicon etch, silicon dioxide is conformably deposited using PECVD. This oxide, at a thickness of 0.3 μm, will serve as a protection for the sidewalls during release (step 6). The next step is interesting and crucial to the process. The objective is to remove the oxide from the bottom and the top of the structure without removing the oxide from the sidewalls. This selective clearing of oxide is carried out in an anisotropic CF_4/O_2 etch at 10 mTorr. It removes 0.3 μm of oxide from the mesas and the trench bottoms while leaving the sidewall oxide intact (step 7). In the next step, a second deep Si etch (BCl_3/Cl_2) deepens the trench further (3–5 μm deeper) to expose some Si sidewall underneath the oxide-covered part (step 8). The release is accomplished with an SF_6 isotropic etch at 90 mTorr. The RIE step etches the Si out from underneath the beams, releasing them in the process (step 9). An etchant such as SF_6 hardly etches the oxide at all and gives good selectivity to the process. A layer of aluminum is then sputter-deposited using DC magnetron sputter deposition (see Chapter 9) (step 10). The Al can be contacted and used as a drive electrode in micromechanical resonator structures. Resonator elements with beam elements 0.5–5 μm wide and aspect ratios >10 have been achieved using this Si SCREAM process.

A variation of the CMOS-compatible Si SCREAM process is used at EG&G's IC Sensors (Milpitas, CA; http://www.meas-spec.com/myMeas/sensors) for the manufacture of accelerometers.

The SCREAM process can also be used to fabricate GaAs micro- and nanomachines. Figure 3.43a outlines the SCREAM process steps in the case of a GaAs-based device.[66] Specifically, the process sequence for the fabrication of a cantilever beam structure identical to the Si one in Figure 3.42a is shown.[68] Considering two identical structures enables us to better compare the degree of complexity of micromachining in those two materials. Single crystal GaAs, in some respects, is an attractive material for micromechanics because of its optoelectronic properties. With GaAs, one can envision the possibility of vertically etched mirrors, lenses, detectors, waveguides, and semiconductor lasers, all adding up to the possibility of a totally integrated optical bench. On the other hand, compared with silicon, GaAs possesses a low fracture strength (its yield strength is considerably lower: 2 vs. 7 GPa) and is more difficult to passivate than Si, as its own oxides are either leaky or chemically unstable. The process example illustrates the difficulty in passivating GaAs versus passivating Si (a dual-layer dielectric is required to passivate GaAs). A more complete comparison of silicon versus GaAs in micromachining is presented in Volume III, Chapter 5 on selected materials for MEMS and NEMS.



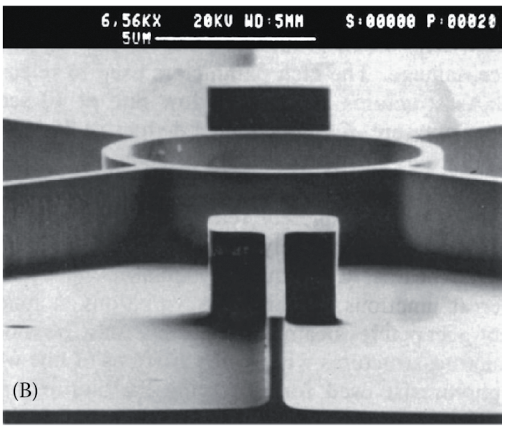


FIGURE 3.43 (A) Formation of an SC-GaAs cantilever beam with aluminum electrodes adjacent to each side of the beam using PECVD-SIN$_x$:H for insulation and for the top and sidewall etch mask. (From Zhang, Z. L., G. A. Porkolab, and N. C. MacDonald. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92).* Travemunde, Germany: IEEE. With permission.[68]) (B) SEM micrograph showing SC-GaAs circular and angled straight-line features after the CAIBE step. (From Zhang, Z. L., and N. C. MacDonald. 1993. Fabrication of submicron high-aspect-ratio GaAs actuators. *J Microelectromech Syst* 2:66–73. With permission.[66])

For creating a simple movable beam, CAIBE produces deep vertical trenches, and RIE laterally undercuts and releases the vertically etched structures from the GaAs substrate.[68] This process produces aspect ratios of 25:1 with trenches of 20 μm vertical depth to 400 nm lateral width. PECVD-SiN$_X$:H (200 nm) and a second layer of SiO$_X$:H (150 nm) are used to mask the GaAs substrate (Figure 3.43Aa). PECVD-SiN$_X$:H on GaAs easily results in structural buckling and layer cracking, but nevertheless is preferred as a conformable layer because the surface of PECVD-SiO$_X$:H on GaAs is too rough as a result of the diffusion of gallium into PECVD-SiO$_X$:H. The PECVD-SiN$_X$:H deposition temperature of 100°C is selected to obtain simultaneously good dielectric properties and low residual stress in the film. To further protect the GaAs during etching, an additional sacrificial PECVD-SiO$_X$:H is deposited. The thermal expansion mismatch between the two dielectrics is less than between either of the dielectrics and the GaAs. The pattern-producing freestanding beams in GaAs are then printed in a photoresist layer on top of the bilayer dielectric. Plasma etching with CHF_3/O_2 transfers this image into the dielectric layer. The pattern in the dielectric stack subsequently is transferred to

the single crystal GaAs by CAIBE (Figure 3.43Ab) using a 500-eV, 0.1-mA/cm^2 Ar ion beam and a Cl_2 flow of 10 ml/min. After the CAIBE process, the remaining photoresist is stripped in oxygen plasma, and a 300-nm PECVD layer of SiN_X:H is conformably deposited. Subsequently, using DC magnetron sputtering, a 400-nm layer of aluminum is deposited (Figure 3.43Ac). Photoresist is spun on the Al layer, and the aluminum side-electrode pattern is printed in the resist, developed (Figure 3.43Ad), and then transferred to the Al by a Cl_2/BCl_3 etch. The latter step removes the aluminum from the areas that have been cleared of resist. Before the Al deposition, contact windows are formed to allow electrical contact with both the GaAs substrate and the movable GaAs beams. To clear the PECVD-SiN_X:H (from the second deposition of this compound) from the bottom of the trenches and to retain the dielectric elsewhere, a highly anisotropic CHF_3/O_2 is applied (Figure 3.43Ae). Finally, the GaAs structures are released from the substrate using a chlorine-based RIE (Figure 3.43Af; the top and sidewalls of the GaAs structure remain protected by the PECVD-SiN_X:H and PECVD-SiO_X:H mask. At the end of the beam-releasing etch, the PECVD-SiO_X:H is consumed, that is, it acts as a sacrificial mask. Finally, the photoresist for the Al patterning is stripped in oxygen. In a simplification of the process described in Figure 3.43A, Zhang et al.[66] later discovered an improved PECVD-SiN_xH_y conformal coating enabling a single dielectric mask for the GaAs SCREAM process. Figure 3.43B demonstrates circular and angled straight-line features in single crystal-GaAs produced with the simplified process.

Example 3.3: Post-CMOS Processing for High-Aspect-Ratio Integrated Silicon Microstructures

Fedder et al.[69] demonstrated the use of a maskless SCREAM-like approach of combined anisotropic and isotropic dry etching to fabricate thin laminated aluminum/silicon dioxide (or dielectrics in general) mechanical structures using the top-level aluminum layer in standard CMOS as an etch mask for the underlying Si.[69] In this process, an anisotropic silicon dioxide etch (CHF_3/O_2) is followed by an anisotropic Si etch (SF_6/O_2), and finally an undercutting isotropic Si etch is used to release the movable mechanical elements. This post-CMOS process is illustrated in Figure 3.44. The etching masks are the interconnect metals from preceding CMOS processing, so no post-CMOS lithography is required, hence the name "maskless post-CMOS micromachining." An aspect ratio of 4:1 can be achieved. The microstructures consist of up to three metal layers that are electrically isolated by dielectrics. Both lateral and vertical CMOS accelerometers with fully differential full-bridge capacitive interface circuits have been fabricated using this post-CMOS micromachining process, and a triaxial microstage has also been realized.[51] A major drawback with this type of CMOS-MEMS is that the residual stress and thermal expansion coefficient in the thin metal/dielectric layers cause curling or bending.

More recently, with the improvements in deep RIE, the same Carnegie Mellon University-based research group has revisited their SCREAM-like

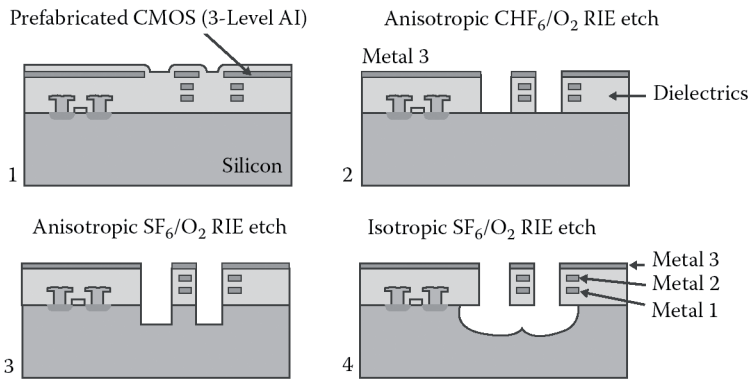


FIGURE 3.44 Cross-sectioned view of thin-film post-CMOS micromachining. Illustration of the process flow showing use of the uppermost metallization level as mask for silicon etching, first with an anisotropic etch and then an isotropic etch to free movable structures.

approach by incorporating a front and a back-side anisotropic Bosch etch. The process flow is illustrated in Figure 3.45A. In step 1, a deep anisotropic etch leaves a 10–100-μm thick single-crystal Si membrane underneath the CMOS microstructural layer. The central idea is to leave a thick layer of single crystal backing the thin CMOS multilayer such that the mechanical properties are dominated by the underlying single-crystal Si, and electrical connections continue to be provided by the CMOS microstructure layer. In step 2, a front-side anisotropic etch is carried out to etch through the thin CMOS dielectric layers. In step 3, the microstructure is released by an anisotropic Si etch instead of the isotropic etch typically used in the SCREAM process. The resulting thick Si, which remains underneath the CMOS layer, results in a totally flat released microstructure (no curling). Moreover, the proof mass and comb-finger capacitance for CMOS-based inertial sensors are significantly increased because of the higher *z*-dimensions. A close-up of one corner of a comb-drive actuator fabricated using the new process is shown in Figure 3.45B.

Example 3.4: Dry Etching of Polymeric Materials

As we learned so far, to obtain the optimum anisotropy in high-aspect-ratio devices, one cools the substrate, uses high RF power density, works at the lowest possible pressure, uses high gas flow rate, relies on sidewall-passivating additives, or, as in the Bosch process, one switches back and forth between etching and passivation. We give examples here of the anisotropic power of MERIE by Furuya et al.[70]

Furuya et al.,[70] at NTT, made a fluorinated polyimide microgrid (Figure 3.46) by using MERIE in oxygen plasma. The grid elements consist of 18 pole-shaped, gold-covered electrodes connected at the base by a common electrode. Individual poles measure 15 μm in diameter and are 100 μm high. The microgrid is formed by pairs of grid elements facing each other. A voltage can be applied between opposing electrodes of the grid-element pairs. A microgrid like this, the authors suggest, can be used in an electrostatic ion drag micropump and in microelectrophoresis devices. The fluorinated polyimide used in this study etches rapidly (3–5 μm/min) and with high selectivity (2600 over a Ti mask) because MERIE produces a high-density, low-energy plasma and because the fluorine in the polyimide produces a high concentration of oxygen radicals in the plasma. From the F/C rule presented above, we predict that a fluorocarbon/oxygen atmosphere leads to a more aggressive and less selective etch. In the current case, by adding the fluorine to the polyimide instead of to the etchant gas, a faster etch rate occurs without losing the high selectivity.

The morphology of the fluorinated polyimide surface, as revealed by atomic force microscopy

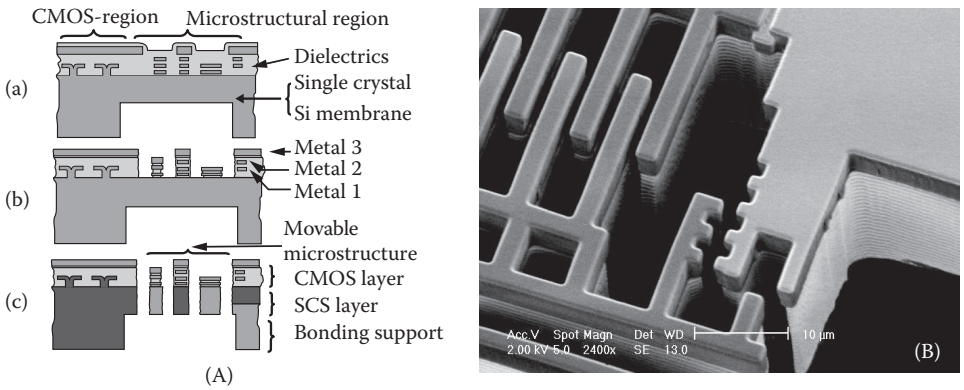


FIGURE 3.45 (A) The process flow for the modified CMOS-MEMS micromachining process; in step (a) the CMOS chip with back-side etch is shown; in step (b) the anisotropic dielectric etch is illustrated; and in step (c) the anisotropic silicon etch for the release is demonstrated. (B) SEM of a corner of a comb-drive actuator fabricated in Hewlett-Packard 0.5-μm CMOS followed by post-CMOS Si DRIE micromachining. The upper 5-μm-thick composite layer is the CMOS three-metal interconnect stack. The underlying Si is 25-μm thick and illustrates the sidewall scalloping caused by the Si DRIE. (Courtesy of Dr. G. Fedder, CMU.)

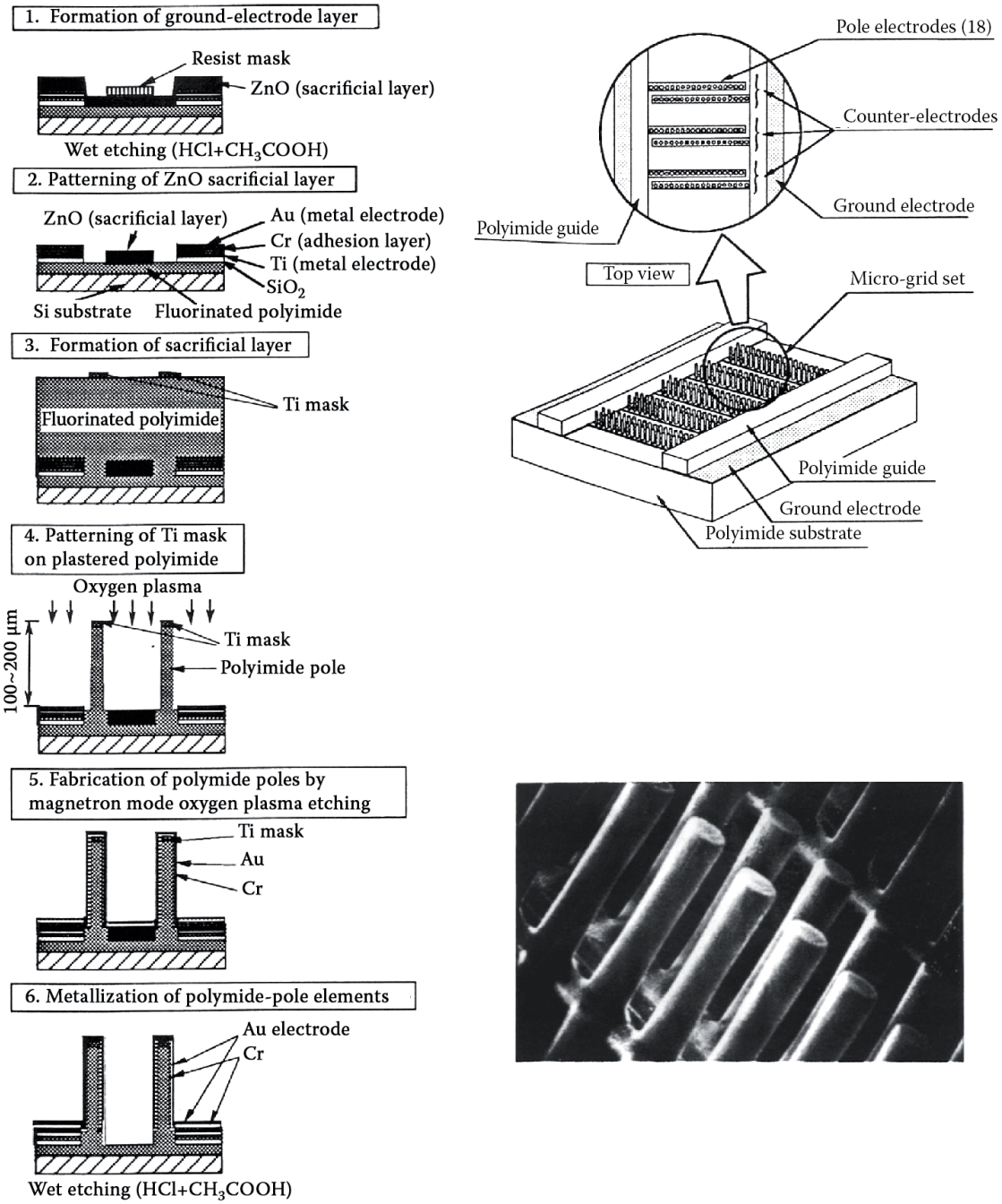


FIGURE 3.46 An array of polyimide posts made by cryogenic MERIE. (Furuya, A., F. Shimokawa, T. Matsuura, and R. Sawada. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93).* Fort Lauderdale, FL: IEEE. With permission.[70])

is smooth at a high etching rate, that is, at high magnetic fields. The NTT fluorinated polyimide is chemically stable, and the fluorine content makes the material extremely transparent, so it can be used to make optical components such as optical waveguides and lenses. The optical quality (and the high-aspect ratio of the produced polyimide structures) could make this fabrication technology competitive for many x-ray lithography-based LIGA structures (see Chapter 10).

Example 3.5: Combination Wet and Dry Etching

Brugger et al.[71] developed a micromachining process, combining wet and dry etching, to fabricate a meander-type cantilever with an

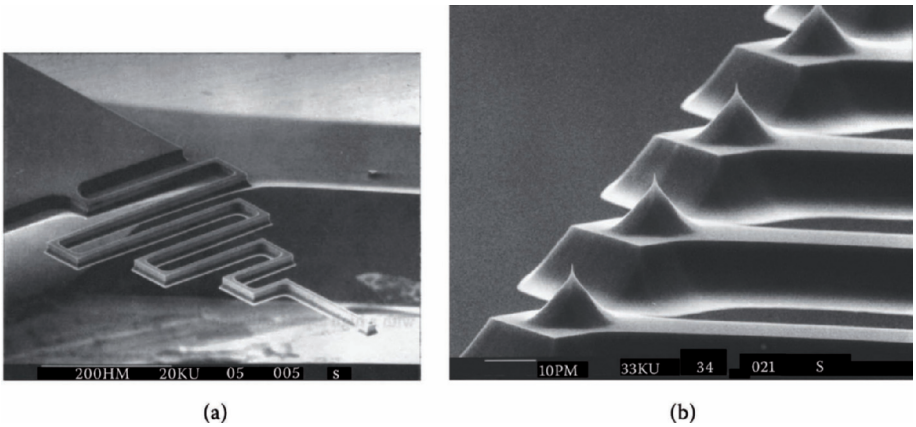


FIGURE 3.47 Si cantilever with square cross-section, designed as a meander shape. (a) Meander cantilever with tip. (b) Close-up view of tips. (From Brugger, J., R. A. Buser, and N. F. de Rooij. 1992. Silicon cantilevers and tips for scanning force microscopy. *Sensors Actuators A* A34:193–200. With permission.[71])

integrated high-aspect-ratio tip for microprobe applications. The crucial part of an atomic force microscope (AFM) or scanning tunneling microscope (STM) microprobe is the tip, which for some applications (e.g., IC trench probing) should have a curvature radius ideally in the range of 50 Å or less, an angle of aperture of about 5°, and a height of 10 µm. The meander structure and integrated tip are presented in Figure 3.47, and the process sequence is in Figure 3.48.

First, a 30-µm thin Si membrane is anisotropically etched in a double-sided polished 280-µm-thick Si wafer (Figure 3.48a). The etching occurs at the back with 40% KOH at 60°C through an opening etched with buffered HF in a 1.5-µm thermal oxide. During the KOH etch, the front side is protected by a mechanical chuck. The front-side oxide is patterned into a two-step profile using two lithographic and two buffered HF etching steps. The 0.75 µm-thick oxide in the two-step profile acts as the mask for the cantilever, and a 1.5-µm-thick oxide serves as the mask for the tip (Figure 3.48a). The topside is etched with 15-µm- deep RIE using chlorine/fluorine gas mixtures (C_2ClF_5/SF_6) to preshape the cantilevers (Figure 3.48b). By adjusting the RIE parameters, vertical sidewalls of the cantilever can be obtained. The 0.75-µm-thick oxide that covered the cantilever is completely removed afterward, and the remaining oxide cap, formerly 1.5 µm and now 0.75 µm thick, serves as a mask for the tip etching. Then a C_2ClF_5/SF_6 gas mixture is applied again to RIE the tip (Figure 3.48c). In the final step (Figure 3.48d), an isotropic

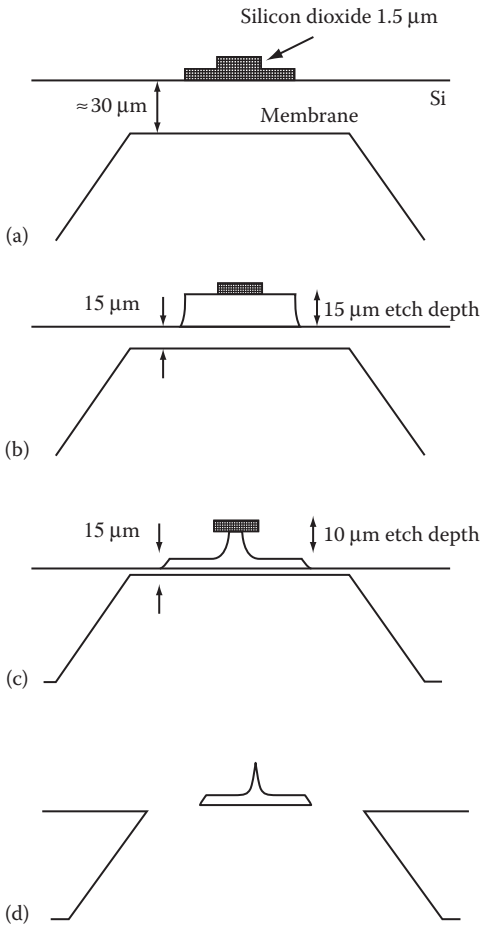


FIGURE 3.48 Fabrication process for a silicon meander beam with integrated tip. (a) Patterning of oxide mask onto preprocessed 30-µm thin membrane. (b) Dry etching of rough cantilever shape. (c) Dry etching of the tip. (d) Final wet etching step releasing the cantilever, forming the tip and smoothing the surface. (From Brugger, J., R. A. Buser, and N. F. de Rooij. 1992. Silicon cantilevers and tips for scanning force microscopy. *Sensors Actuators A* A34:193–200. With permission.[71])

HNO_3:HF:CH_3COOH etch lasting 2 minutes forms the tips (using the oxide cap as mask), releases the cantilever by etching through the remaining Si membrane, smoothes the dry-etched rough surface, and cleans the remaining organic photoresist. Mixtures of HNO_3, HF, with CH_3COOH and/or water etch the silicon isotropically and exhibit high etch rates (10–300 μm/min). The SiO_2 mask withstands the three successive etch steps: cantilever RIE etching, tip RIE etching, and 1–2-min wet etching with HF:HNO_3:CH_3COOH sharpening the tip. Tip heights up to 20 μm with opening angles of approximately 5–10° and tip radii estimated to be 40 nm can be obtained this way.

A postoxidation process yielding Si tips with curvature radii of less than 10 nm by several consecutive oxidation and HF etching steps, exploiting an anomaly of the oxidation behavior at regions with high geometric curvature, has been demonstrated by Marcus et al.[72,73] and by our own group[74] (see also Volume I, Chapter 4, on the dependence of oxidation rate of Si on geometry). By applying this oxidation-sharpening technique to the tips produced in the current example, further sharpening would result. A rigid cantilever structure rather than a meander structure as shown in Figure 3.48a would also be of more use in a practical STM.[74]

Acknowledgments

Special thanks to Genis Turon, Robert Gorkin, and Xavier Casadevall.

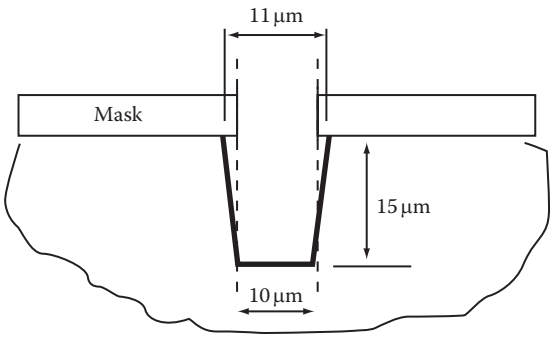
Questions

3.1: How is a DC plasma created, and how does an RF plasma differ? Why is a plasma always positive with respect to the reactor vessel walls? In which etching setup would you prefer to etch an insulator?

3.2: Detail the different dry etching profiles available and explain how they are obtained.

3.3: Consider a single electron in an electric field between two parallel plates located 10 centimeters apart (d = 10 cm). Assume the potential varies sinusoidally between 1000 V and –1000 V at a frequency of 13.5 MHz. Calculate the maximum kinetic energy of the electrons.

3.4: Using the figure below as a model, assume that the silicon wafer has a bias of –150 V. A reactive ion etch is used to etch 10 μm wide grooves.



After 1 hour, the shape of the grooves is measured and found to have the following topology:

(a) If the bias were increased to –200 V and the depth of the groove was maintained at 15 μm, would the width at the top of the groove be less than or greater than 11 μm?
Explain.

(b) If the same etch procedure was used to etch SiO_2, qualitatively describe the shape of the grooves in the SiO_2 if a –150 V bias is applied.

3.5: Explain the DC breakdown voltage vs. electrode distance curve (Paschen's law) and how it is relevant to dry etching. How is miniaturization of an electrode set equivalent to creating a local vacuum?

3.6: Use SCREAM to fabricate a capacitive accelerometer. You may rely on the newest deep dry etching technology.

3.7: Why is reactive ion etching (RIE) somewhat of a misleading name? In physical-chemical etching, there are at least three types of ion-surface interactions which may promote dry etching. Detail each of them.

3.8: Develop a process sequence to fabricate a capacitive pressure sensor using both wet and dry etching techniques.

3.9: Explain loading effects in wet and dry etching. How do you avoid the loading effect? What is the bull's-eye effect? What is the difference between loading and micro-loading?

3.10: Discuss the etch profiles in physical etching (ion sputtering). Also draw profiles exhibiting faceting, ditching, and redepositioning.

3.11: Explain the F/C ratio effect in dry etching and present a couple of means to influence it. How can cryogenic cooling lead to similar etching behavior as inhibitor-driven chemistry?

3.12: Explain the ion energy versus pressure curve for a plasma and how different etching mechanisms are enabled by it. Describe the differences between high-pressure plasmas and reactive ion etching systems. Explain for what applications each is the preferred process.

3.13: You have only a high-pressure plasma etch and an ion mill at your disposal. Select which one to use in the following applications. Justify your answer.

(a) Etching a 5000 Å polysilicon layer that serves as the upper electrode of a large square capacitor. The capacitor dielectric is 50 Å of SiO_2.

(b) Anisotropic patterning of a thin layer of $YBa_2Cu_3O_7$ on a thick insulating film.

(c) Recessing the channel of a GaAs FET. For this application the residual etch damage must be minimized.

3.15: Sketch the voltage distribution for an RF plasma discharge in equilibrium and contrast it with the voltage distribution in a DC plasma.

3.16: How would you monitor a dry etching process? Why does endpoint detection of a reactant species require a loading effect?

3.17: Why is the grounded electrode (anode) in an AC sputtering station made as large as possible? Use equations to illustrate your answer.

3.18: Explain how magnetic coils enable higher ion density plasmas. Use equations to explain.

3.19: Compare dry versus wet etching techniques.

3.20: How would you make the structure below with the least expensive equipment, the highest selectivity, and the fastest process?

Further Reading

Boxman, R. L., D. M. Sanders, and P. J. Martin. 1995. *Handbook of vacuum arc science and technology*. Park Ridge, NJ: Noyes Publications, Chapter 1.

Brown, S. D. 1967. *Basic data of plasma physics, 1966.* Cambridge, MA: MIT Press.

Cobine, J. D. 1958. *Gaseous conductors*. Mineola, NY: Dover Publications.

Nasser, E. 1971. *Fundamentals of gaseous ionization and plasma electronics.* Hoboken, NJ: Wiley-Interscience.

Roth, J. R. 1995. *Industrial plasma engineering.* Bristol, UK: Institute of Physics Publication, Chapters 8–13.

References

1. Lehmann, H. W. 1991. *Thin film processes II*. Eds. J. L. Vossen, and W. Kern. Boston: Academic Press.
2. Jansen, H., H. Gardeniers, M. de Boer, M. Elwenspoek, and J. Fluitman. 1996. A survey on the reactive ion etching of silicon in microtechnology. *J Micromech Microeng* 6:14–28 and Jansen, H. V., M. de Boer, R. Legtenberg, and M. Elwenspoek. 1994. The black silicon method: a universal method for determining the parameter setting of a fluorine-based RIE in deep Si trench etching with profile control. *Proceedings of micromechanics Europe (MME'94)*, 60–64.
3. Oehrlein, G. S., M. F. Doemling, B. E. E. Kastenmeier, P. J. Matsuo, N. R. Rueger, M. Schaepkens, and T. E. F. M. Standaert. 1999. Surface science issues in plasma etching. *IBM J Res Dev* 43:181.
4. Rossnagel, S. M. 1991. *Thin film processes II*. Eds. J. L. Vossen, and W. Kern. Boston: Academic Press.
5. Konuma, M. 1991. *Film deposition by plasma techniques.* Berlin: Springer-Verlag.
6. Sugawara, M., B. L. Stansfield, S. Handa, K. Fujita, S. Watanabe, and T. Tsukamoto. 1998. *Plasma etching fundamentals and applications.* Oxford, UK: Oxford University Press.
7. Coe, S. E., J. A. Stocks, and A. J. Tambini. 1993. An investigation of the cathode region of a fluorescent lamp. *J Phys D Appl Phys* 26:1203–10.
8. Coe, S. E., and G. G. Lister. 1992. Modeling of the negative glow and Faraday dark space of a low-pressure noble gas discharge. *J Appl Phys* 71:4781–87.
9. Raizer, Y. P., and M. N. Shneider. 1997. The nonmonotonicity of transition from Faraday dark space to positive column and the emergence of standing strata behind the cathode region of glow discharge. *High Temperature* 35: 15–21.
10. Bogaerts, A., and R. Gijbels. 1996. Mathematical description of a direct current glow discharge in argon. *Fresenius J Anal Chem* 355:853–57.
11. Bogaerts, A., and R. Gijbels. 1996. Two-dimensional model of a direct current glow discharge: description of the electrons, argon ions, and fast argon atoms. *Anal Chem* 68:2296–303.
12. Bogaerts, A., and R. Gijbels. 1998. Fundamental aspects and applications of glow discharge spectrometric techniques. *Spectrochim Acta (Part B)* 53:1–42.
13. Paschen, F. 1889. Ueber die zum Funkenübergang in Luft, Wasserstoff und Kohlensäure bei verschiedenen Drucken erforderliche Potentialdifferenz. *Wied Ann* 37:69–96.

14. Naidu, M. S., and V. Kamaraju. 1995. *High voltage engineering.* New York: McGraw-Hill.
15. Meek, J. M., and J. D. Craggs. 1953. *Electrical breakdown of gases.* Oxford, UK: Clarendon Press.
16. Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press.
17. van Roosmalen, A. J. 1984. Review: dry etching of silicon oxide. *Vacuum* 34:429–36.
18. Wolf, S., and R. N. Tauber. 1987. *Silicon processing for the VLSI era.* Sunset Beach, CA: Lattice Press.
19. Comello, V. 1993. ICP etchers start to challenge ECR. *R & D* April:79–80.
20. Singer, P. 1991. ECR: is the magic gone? *Semicond Int* July:46–48.
21. van Roosmalen, A. J., J. A. G. Baggerman, and S. J. H. Brader. 1991. *Dry etching for VLSI.* New York: Plenum Press.
22. Taniguchi, N. 1989. *Energy-beam processing of materials: advanced manufacturing using various energy sources.* Oxford, UK: Clarendon Press.
23. Schutz, R. J. 1988. *VLSI technology.* Ed. S. M. Sze. New York: McGraw-Hill.
24. Arnold, J. C., and H. H. Sawin. 1991. Charging of pattern features during plasma etching. *J Appl Phys* 70:5314.
25. Elwenspoek, M., H. Gardeniers, M. de Boer, and A. Prak. 1994. *Micromechanics.* Report No. 122830. Twente, the Netherlands: University of Twente.
26. Ayon, A. A., X. Zhang, and R. Khanna. 2000. *Solid-state sensor and actuator workshop.* Hilton Head Island, SC: Transducers Research Foundation, Inc.
27. Jansen, H., H. Gardeniers, M. de Boer, M. Elwenspoek, and J. Fluitman. 1996. A survey on the reactive ion etching of silicon in microtechnology. *J. Micromech Microeng* 6:14–28.
28. Editorial. 1996. *Introducing plasma jet etching.* San Jose, CA: IPEC Precision, Inc.
29. Jeong, J. Y., S. E. Babayan, V. J. Tu, J. Park, I. Henins, R. F. Hicks, and G. S. Selwyn. 1998. Etching materials with an atmospheric-pressure plasma jet. *Plasma Sources Sci Technol* 7:282–85.
30. Kneissl, M., D. Hostetter, D. P. Bour, R. Donaldson, J. Walker, and N. M. Johnson. 1998. Dry-etching and characterization of mirrors on III-nitride laser diodes from chemically assisted ion beam etching. *J Crystal Growth* 189/190:846–49.
31. Carlstrum, C. F., S. Anand, and G. Landgren. 1999. Extremely smooth surface morphologies in $N_2/H_2/CH_4$ based low energy chemically assisted ion beam etching of InP/GaInASP. *Thin Solid Films* 343–344:374–77.
32. Li, Y. X., P. J. French, P. M. Sarro, and R. F. Wolffenbuttel. 1995. *Proceedings: IEEE electro mechanical systems conference.* Amsterdam, the Netherlands: IEEE.
33. Schwartz, G. C., and P. M. Schaible. 1979. Reactive ion etching of silicon. *J Vac Sci Technol* 16:410–13.
34. Oehrlein, G. S. 1990. *Reactive ion etching: handbook of plasma processing technology.* Ed. S. M. Rossnagel. Park Ridge, NJ: Noyes.
35. Juan, W. H., J. W. Weigold, and S. W. Pang. 1996. Dry etching and boron diffusion of heavily doped, high aspect ration SI trenches. *Proceedings: SPIE*. Austin, TX: SPIE.
36. Flamm, D. L. 1992. Dry plasma resist stripping part I: overview of equipment. *Solid State Technol* 35:37–39.
37. Mogab, C. J., A. C. Adams, and D. L. Flamm. 1979. Plasma etching of Si and SiO_2: the effect of oxygen additions to CF_4 plasmas. *J Appl Phys* 49:3796–803.
38. Heinecke, R. H. 1975. Control of relative etch rates of SiO_2 and Si in plasma etching. *Solid State Electron* 18: 1146–47.
39. Peters, L. 1992. Plasma etch chemistry: the untold story. *Semicond Int* May:660–70.
40. Moreau, W. M. 1988. *Semiconductor lithography.* New York: Plenum Press.
41. Hess, D. W., and D. B. Graves. 1989. *Microelectronics processing: chemical engineering aspects.* Eds. D. W. Hess, and K. F. Jensen. Washington, DC: American Chemical Society.
42. Flamm, D. L. 1992. Dry plasma resist stripping Part II: physical processes. *Solid State Technol* 35:40–42.
43. Flamm, D. L. 1993. Feed gas purity and environmental concerns in plasma etching: part II. *Solid State Technol* November:43–50.
44. Flamm, D. L. 1992. Dry plasma resist stripping part III: production economics. *Solid State Technol* 35:43–48.
45. Klaassen, E. H., K. Petersen, J. M. Noworolski, J. Logan, N. I. Maluf, J. Brown, C. Storment, W. McCulley, and G. T. A. Kovacs. 1995. Proceedings: Eighth international conference on solid-state sensors and actuators, and Eurosensors IX. Stockholm, Sweden: IEEE.
46. Kovacs, G. T. A., N. I. Maluf, and K. E. Petersen. 1998. Bulk micromachining of silicon. *Proc IEEE* 8:1536–51.
47. Laermer, F., and A. Schilp. 1996. Method of anisotropically etching silicon. US Patent 5,501,893.
48. Maluf, N. 2000. *An introduction to microelectromechanical systems engineering.* Boston: Artech House.
49. Chow, E. M., A. Partridge, C. F. Quate, and T. W. Kenny. 2000. *Solid-state sensor and actuator workshop.* Hilton Head Island, SC: Transducers Research Foundation, Inc.
50. Bosch-Charpenay, S., J. Xu, J. Haigis, P. A. Rosenthal, P. R. Solomon, and J. M. Bustillo. 2000. *Technical digest: solid-state sensor and actuator workshop.* Hilton Head Island, SC: Transducers Research Foundation, Inc.
51. Xie, H., L. Erdmann, X. Zhu, K. J. Gabriel, and G. K. Fedder. 2000. *Technical digest: solid-state sensor and actuator workshop.* Hilton Head Island, SC: Transducers Research Foundation, Inc.
52. Ayon, A. A., R. Braff, C. C. Lin, H. Sawin, and M. A. Schmidt. 1999. Characterization of a time multiplexed inductively coupled plasma etcher. *J Electrochem Soc* 146: 339–49.
53. Ayon, A. A., R. Braff, R. Bayt, H. H. Sawin, and M. A. Schmidt. 1999. Influence of coil power on the etching characteristics in a high density plasma etcher. *J Electrochem Soc* 146:2730–36.
54. Juan, W. H., S. W. Pang, A. Selvakumar, M. W. Putty, and K. Najafi. 1994. *Technical digest: 1994 solid-state sensor and actuator workshop.* Hilton Head Island, SC: Transducers Research Foundation, Inc.
55. Chinn, J. D., I. Adesida, and E. D. Wolf. 1983. Chemically assisted ion beam etching for submicron structures. *J Vac Sci Technol* B1:1028–32.
56. Hoffman, E., B. Warneke, E. Kruglick, J. Weigold, and K. S. J. Pister. 1995. *Proceedings: IEEE micro electro mechanical systems (MEMS '95).* Amsterdam, the Netherlands.
57. Köhler, U., A. E. Guber, W. Bier, M. Heckele, and T. Schaller. 1996. *Proceedings: SPIE miniaturized systems with micro-optics and micromechanics.* San Jose, CA: SPIE, 18–22.
58. Wang, X.-Q., X. Xang, K. Walsh, and Y.-C. Tai. 1997. *Proceedings: Transducers '97*. Chicago.
59. Mucha, J. A., and D. W. Hess. 1983. *Introduction to microlithography.* Eds. L. F. Thompson, C. G. Willson, and M. J. Bowden. Washington, DC: ACS Symposium Series.
60. Economou, D., E. S. Aydil, and G. Barna. 1991. In situ monitoring of etching uniformity in plasma reactors. *Solid State Technol* 34:107–11.

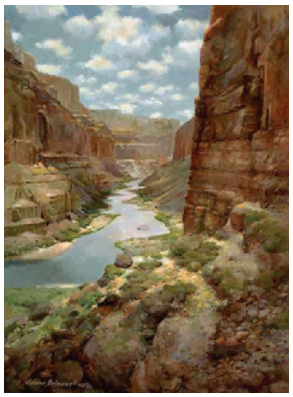
61. Elta, M. 1993. Developing "smart" controllers for semiconductor processes. *Res Dev* February:69–70.
62. Oehrlein, G. S. 1986. Reactive-ion etching. *Phys Today* 39:26–33.
63. Deshmukh, V. G. I., and T. I. Cox. 1988. Physical characterization of dry etching plasmas used in semiconductor fabrication. *Plasma Physics Controlled Fusion* 30: 21–33.
64. Duek, R., N. Vofsi, M. Haemek, S. Mangan, and M. Adel. 1993. Improving plasma etch with wafer temperature readings. *Semicond Int* July:208–10.
65. Zhang, Z. L., and N. C. MacDonald. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco, CA.
66. Zhang, Z. L., and N. C. MacDonald. 1993. Fabrication of submicron high-aspect-ratio GaAs actuators. *J Microelectromech Syst* 2:66–73.
67. Yao, J. J., S. C. Arney, and N. C. MacDonald. 1992. Fabrication of high frequency two-dimensional nanoactuators for scanned probe devices. *J Microelectromech Syst* 1:14–22.
68. Zhang, Z. L., G. A. Porkolab, and N. C. MacDonald. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92)*. Travemunde, Germany.
69. Fedder, G. K., S. Santhanam, M. L. Reed, S. C. Eagle, D. F. Guillou, M. S.-C. Lu, and L. R. Carley. 1997. Laminated high-aspect-ratio microstructures in a conventional CMOS process. *Sensors Actuators* A57:103–10.
70. Furuya, A., F. Shimokawa, T. Matsuura, and R. Sawada. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL: IEEE.
71. Brugger, J., R. A. Buser, and N. F. de Rooij. 1992. Silicon cantilevers and tips for scanning force microscopy. *Sensors Actuators A* A34:193–200.
72. Marcus, R. B., and T. S. Ravi. 1993. Method for making tapered microminiature silicon structures, Bell Communications Research. US Patent 5,201,992.
73. Marcus, R. B., T. S. Ravi, T. Gmitter, K. Chin, D. Liu, W. J. Orvis, D. R. Ciarlo, C. E. Hunt, and J. Trujillo. 1990. Formation of silicon tips with <1 nm radius. *Appl Phys Lett* 56:236–38.
74. Editorial. 1991. Tips for STM. TSDC (Teknekron). Menlo Park, CA.
75. Flamm, D. L. 1993. Feed gas purity and environmental concerns in plasma etching: Part I. *Solid State Technol* October:49–54.
76. Manos, D. M., and D. L. Flamm, eds. 1989. *Plasma etching: an introduction*. Boston: Academic Press.

4

Wet Chemical Etching and Wet Bulk Micromachining—Pools as Tools

Fear of things invisible is the natural seed of that which every one in himself calleth religion.

Thomas Hobbes
Leviathan (1651)



Pools as Tools, Nankoweap, by Jerome Grimmer

Outline

Introduction
Historical Note
Chemical Etching
Wet Bulk Micromachining
Comparing Wet and Dry Etching of Si
Questions
References

Ignorance more frequently begets confidence than does knowledge: it is those who know little, and not those who know much, who so positively assert that this or that problem will never be solved by science.

Charles Darwin
The Descent of Man (1871)

The human mind treats a new idea the way the body treats a strange protein; it rejects it.

P.B. Medawar

Introduction

Chapter 4 is split in two parts: one part deals with *wet chemical etching* and the other with *wet bulk micromachining*. Both these subtractive manufacturing techniques cover purely chemical etching, photochemical etching, electrochemical etching, and combinations thereof. Wet bulk micromachining is derived from chemical etching, but wet etching and bulk micromachining are typically used in very different industrial settings. In chemical etching, one carves isotropic features, large (milling) and small (blanking), out of metals, glasses, polymers, and ceramics. The technique is mostly used in more traditional industrial manufacturing settings such as aerospace, electronics, automotive, medical, computer, instrument manufacturing, and decorating. Wet bulk micromachining, on the other hand, is used to sculpt isotropic and anisotropic microstructures principally in semiconductors substrates. It is a technique championed in the integrated circuit (IC) industry and in micro- and nanomachining.

We start with a historical note covering both chemical etching and wet bulk micromachining.

Historical Note

The ancient Egyptians etched copper jewelry with citric acid as long ago as 2500 BC. The Hohokam people, of what is now Arizona, etched snail-shell jewelry with fermented cactus juice around 1000 BC. The earliest use of wet etching of a substrate, using a mask (wax) and etchants (acid-base), appears in the late fifteenth or early sixteenth century for decorating armor (Figure 4.1).[1,2] Engraving hand tools were not hard enough to work the armor, and more powerful acid-based processes took over. By the early seventeenth century, chemical etching to decorate arms and armor was a well-established process. Some pieces stemming from that period have been found with designs as small as 0.5 mm. This is remarkable as masking involved cutting the maskant (linseed oil paint) with a scribing tool and peeling the maskant off where etching (with vinegar) was desired. This purely chemical-driven etching process is today also called chemical milling (larger pieces) or chemical blanking (smaller pieces).

FIGURE 4.1 Armor pieces have been found where the chemical milling led to designs as small as 0.5 mm. A maskant (wax) is cut with a scribing tool and peeled off where etching is desired. Elements of a Light-Cavalry Armor, ca. 1510 Italian (Milan). Steel, etched, and gilded; weight 19 lb., 13 oz. (9 kg). Gift of William H. Riggs, 1913. The Metropolitan Museum of Art. http://www.metmuseum.org/TOAH/hd/rarm/ho_14.25.716.htm.

Harris[1] describes in detail the improvements that, by the mid-1960s, made chemical milling a valuable and reliable method of manufacturing, especially popular in the aerospace industry (M. Sanz of North American Aviation patented chemical milling in 1953 for use on aerospace parts, and in 1956, Turco Inc. purchased the patent rights). The method enabled many parts to be produced more easily and cheaply than by other means and, in many cases, provided design and production configurations not previously possible.

Through the introduction of photosensitive masks by Niépce in 1822 (see Chapter 1), chemical milling in combination with lithography became a reality, and yet smaller feature sizes were within reach. This type of chemical etching is called photochemical milling (PCM). In PCM, photomasks are used to mask parts typically smaller and thinner than those used in chemical milling. The inventor of the printed circuit board (PCB) is generally assumed to be the Austrian engineer Paul Eisler (1907–1995), who, working in England, made a printed circuit board in 1936 as part of a radio set (Eisler is sometimes referred to as the "father of the PCB"). In about 1943 the United States began to use the

technology on a large scale to make rugged radios for use in World War II. After the war, in 1948, the United States released the invention for commercial use. Printed circuits did not become commonplace in consumer electronics until the mid-1950s. A PCB basically consists of a planar (flat) substrate that has electronic components mounted on it that are interconnected by conductive tracks. The high volumes required for this product provided strong impetus for the development of more sophisticated PCM methods, particularly in the design of better etching equipment. Today the two most important applications of lithography-based chemical milling remain the manufacture of PCBs and, by 1961, the fabrication of Si-integrated circuitry. The latter application of chemical milling became better known as *wet bulk micromachining*. In this case, the substrates are typically smaller yet (Si wafer size) than in PCM, and the etch profile can be isotropic or anisotropic. The method used to produce a "cut" when etching silicon integrated circuits (ICs) is the same as that used for photochemical milling of metals in the aerospace industry, but the IC and microelectromechanical system (MEMS)/nanoelectromechanical system (NEMS) industries have refined the technique to reach minimum feature sizes orders of magnitudes smaller and to obtain much higher aspect ratios (height-to-width ratio of the features as high as 400 in anisotropic etching of Si). Whereas the device sizes in microfabrication are often much larger than in IC processing, the aspect ratios in MEMS/NEMS are generally much greater than in IC fabrication (see Figure 4.2).

Isotropic etching in silicon semiconductor processing has been used since its introduction in the early 1950s. One finds representative work from that period in an impressive series of papers by Robbins and Schwartz[3–6] on chemical isotropic etching and in Uhlir's early work on electrochemical isotropic etching.[7] The usual chemical isotropic etchant used for silicon was HF in combination with HNO_3, with or without acetic acid or water as diluent. The early work on isotropic etching in an electrochemical cell (i.e., *electropolishing*) was carried out mostly in nonaqueous solutions to avoid black or red deposits that formed on the silicon surface in aqueous solutions.[8] Turner[9] showed that, if a critical current density is exceeded, silicon can be electropolished in aqueous HF solutions without the formation of those deposits.

In the mid-1960s, Bell Telephone Laboratories started work on anisotropic Si etching in mixtures of, at first, KOH, water, and alcohol and later in KOH and water. The need for high-aspect-ratio cuts in silicon arose in the fabrication of dielectrically isolated structures in integrated circuits, such as those for beam leads. Both chemical and electrochemical anisotropic etching methods were pursued.[10–20] In the mid-1970s, a new surge of activity in anisotropic etching was associated with the work on V-groove and U-groove transistors.[21–23]

In Volume I, Chapter 1, we mentioned that the first use of Si as a micromechanical element can be traced back to a discovery and an idea from the mid-1950s and early 1960s, respectively. The discovery was the large piezoresistance in Si and Ge by Smith in 1954.[24] The idea stemmed from Pfann et al.,[25] who in 1961 proposed a diffusion technique for the fabrication of Si piezoresistive sensors for stress, strain, and pressure. This led eventually to

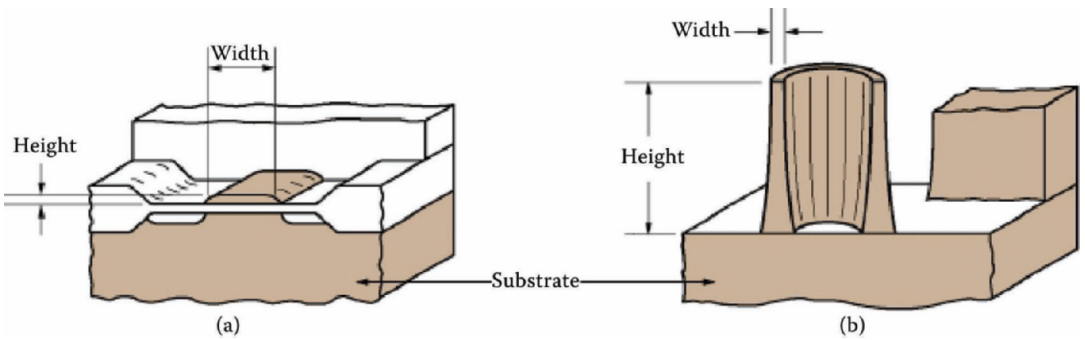


FIGURE 4.2 Aspect ratio (height-to-width ratio) typical in fabrication of integrated circuits (a) and microfabricated components (b).

the development of commercial Si micromechanical pressure sensors, accelerometers, and gyros made by wet bulk micromachining.

Chemical Etching

Introduction

Under chemical etching we review a number of different processes where material removal through etching of preferentially exposed (masked) surfaces produces stress-free parts in many types of substrates. The energy source at the work piece in this subtractive process is purely chemical (chemical milling or CM), photochemical (photochemical milling or PCM and photofabrication), or electrochemical (all types of electrochemical removal processes).

In the simplest embodiment of chemical etching, i.e., chemical milling or chemical machining, the method relies on nonlithographic masking, in which a masking layer is physically carved out to open up a pattern. A scribing tool, of the surgical knife type, is used to cut through the maskant along the line defined by the edge of a template.

In PCM, photomasks are used to mask parts typically much smaller and thinner than those used in chemical milling. The process uses photolithography just as in the etching of Si and other semiconductors, in which case one refers to wet bulk micromachining (see below). Photofabrication is a type of photochemical milling where ultraviolet exposure through a mask results in a strong modification of the solubility in an optically clear solid layer. This permits the direct photochemical production of thick three-dimensional (3D) glass, plastic, or ceramic structures without the need for a separate pattern transfer to an underlying substrate.

Electrochemical removal techniques use an electrical field in an electrolyte that destroys the atomic bonds of the material.[26] Under electrochemical removal techniques, we will review electrochemical machining (ECM), including microelectrochemical machining (μ-ECM), electrochemical jet etching, laser-assisted electrochemical jet micromachining, scanning electrochemical microscope (SECM), and through-mask electrochemical micromachining and electrochemical grinding (ECG).

Chemical Milling

Chemical milling or chemical machining involves material removal through chemical etching of preferentially exposed (masked) surfaces, producing stress-free parts in many metals. It is the oldest among nontraditional machining processes and has been used to engrave metals and hard stones. The method is also referred to as immersion chemical blanking/milling, and it is a purely chemical process that dissolves material from the unmasked (unprotected) areas of metallic parts immersed in a tank filled with heated and agitated chemical reagents. The term "blanking" denotes small, thin work pieces, and "milling" indicates relatively large work pieces. The method produces clean, sharp, well-defined edges with no burrs (often a source of fatigue failure in highly stressed parts). The work piece does not require finishing of the surface, and no electric current to remove material is involved. Although most common with aluminum and titanium, several other metals have been successfully chemically milled, with different metals etched in different etch solutions. The method relies on nonlithographic masking, in which a masking layer is physically carved out to open up a pattern as illustrated in Figure 4.3.[1] The mask, an organic layer—usually a tape or chemically resistant paint—covers a large area substrate that may be up to 2 cm thick. The etch factor, A/R, or eat-back ratio, is identified in Figure 4.3a, with R the etch depth and A the overhang of the maskant. A scribing tool, of the surgical knife type, is used to cut through the maskant along the line defined by the edge of a template (Figure 4.3b). Lateral tolerances of ±0.25–0.5 mm are typical for chemical milling, an accuracy mainly determined by the manual scribing process. For the IC industry and micromachining, this lateral tolerance would be very poor performance, but for many industrial applications, this tolerance is adequate. Chemical machining of metals is isotropic in nature, and aspect ratios are low. Creating high-aspect-ratio structures (>1), e.g., to make deep channels in stainless steel sheets, would necessitate laminating several of these sheets. Obviously, this type of technology is better suited for etching shallow features in large, single-substrate devices. Chemical milling has at least two key

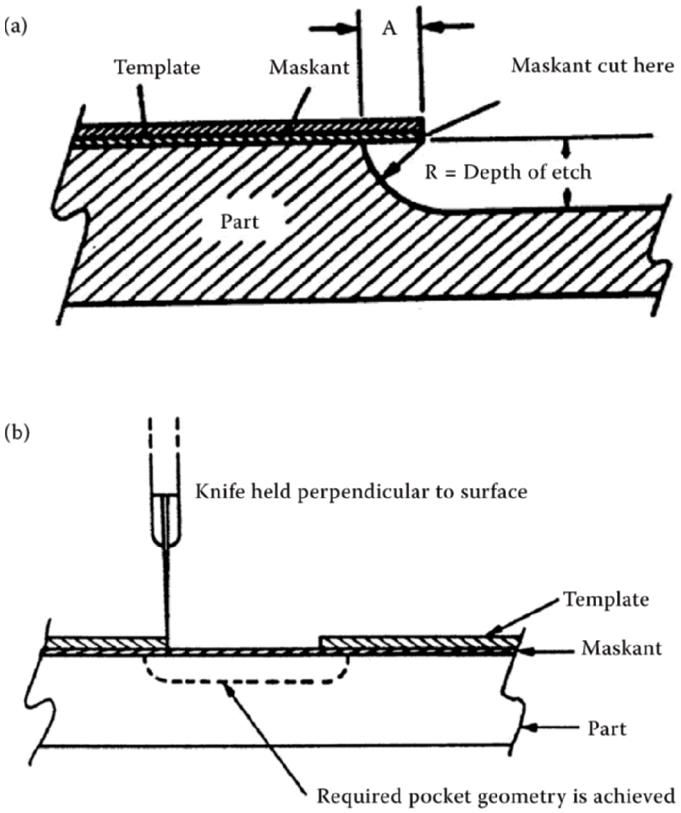


FIGURE 4.3 Chemical milling showing the effect on the work piece. (a) Undercut caused by chemical milling. (b) Scribing of the maskant. (From Harris, T. W. 1976. *Chemical milling.* Oxford: Clarendon Press. With permission.[1])

advantages over traditional mechanical milling. First, mechanical milling is difficult on thin metal sheets because the tool tends to pull the thin metal foil, severely damaging the work piece. Second, chemical milling can produce complexly shaped parts, such as the inside leading edge of an airplane wing, which a milling machine could not reach.

Chemical milling of metals is widely applicable, simple, and requires little operator skill. Practical cutting (etching) depths are generally limited to less than 2 cm because the etchant process is relatively

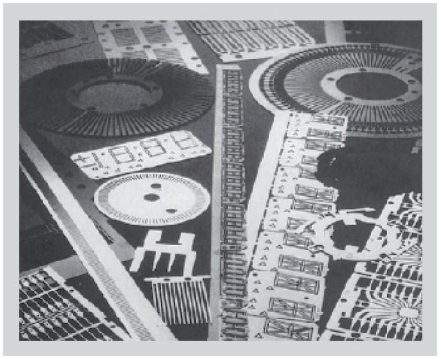


FIGURE 4.5 Various parts made by chemical blanking. Note the fine detail. (Courtesy of Buckabee-Mears St. Paul.)

slow, and undercutting occurs behind the edges of the protective mask. Industrial applications are mainly in the shaping of thin metal foils and large shallow areas. Most modern aerospace designs use chemical milling to remove about half the weight of their products. It is used extensively on commercial and military airplanes, satellites, and the space shuttle. Large parts can be accommodated; aluminum airplane wing parts, for example, are chemically machined in 50-ft.-long tanks. In Figure 4.4 we show a chemical milling tank and a piece that has been thinned using three consecutive masking steps. Example products made by chemical blanking are shown in Figure 4.5.

A good further review with many relevant details on wet etching as it relates to the IC and electronics industry is presented by Romankiw.[27]

Photochemical Milling

Photochemical milling (PCM) is an engineering production technique for the manufacture of burr-free and stress-free flat metal components by selective chemical etching through a photographically

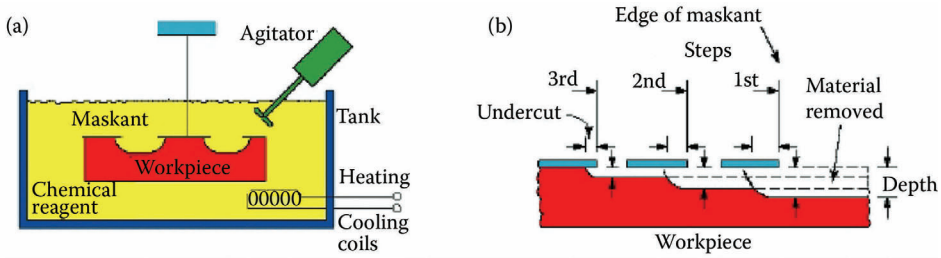


FIGURE 4.4 (a) Schematic illustration of the chemical machining process. Note that no forces or machine tools are involved in this process. (b) Stages in producing a profiled cavity by machining; note the undercut.

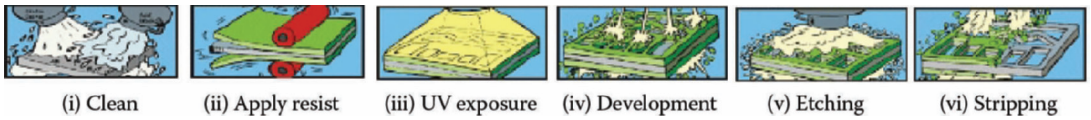


FIGURE 4.6 Photochemical milling (PCM) of metal parts.

produced mask. In PCM of metals, illustrated in Figure 4.6, photomasks are used to mask parts typically much smaller (a maximum of about 60 × 60 cm) and thinner (<0.5 mm) than those used in chemical etching (see above). The process uses photolithography just as in the etching of Si and other semiconductors as discussed in the section below on wet bulk micromachining. To ensure high accuracy, oversized drawings (100×) of a design are reduced photographically to actual size. The photomask is then used to contact print the component image onto a metal sheet covered with a photosensitive coating. The photosensitive layer may be a dry resist layer such as Etertec® 5600 from Eternal (http://www.eternal-group.com) or Riston CM206 from DuPont (http://www2.dupont.com/Imaging_Materials/en_US/tech_info/product_selector.html) (see also Chapter 1). Careful masking is crucial to achieve precise parts in both chemical and photochemical cutting. However, because of the absence of scribing errors in photochemical milling, the photosensitive maskant is extremely accurate and leads to far superior tolerances for small intricate parts. Photochemical milling was designed specifically to alleviate the lateral-accuracy problem of chemical milling, achieving it extraordinarily well; lateral accuracies (e.g., for printed circuit board production) are ±0.013 mm. Another reason why the tolerances in PCM are so much better than in chemical milling is related to the attempted etch depths. For example, the geometry of the cut produced with photoetching printed circuit boards or integrated circuits might be similar to the chemical-milling cut of the aerospace industry, but one must remember the many orders in magnitude difference in the depth of the cut.[1] As the dimensions of the components decrease (i.e., from precision products to printed circuit boards and to integrated circuits), the necessity for chemical cleanliness of the material surfaces increases and requires clean-room manufacture in the case of IC fabrication.

Photoetching of metals is less generic than chemical milling but can be used, for example, with the following metals (in order of increasing difficulty): copper, nickel, carbon steel, stainless steel, aluminum, titanium, and molybdenum.[28,29] Etching rates—with sodium hydroxide for aluminum, hydrochloric and nitric acids for steel, and iron chloride with nitric acid for copper—vary from $1.3–7.6 \times 10^{-3}$ cm/min. Some PCM etchants are listed in Table 4.1.

Photochemical etching of metals, like chemical etching, is isotropic, and aspect ratios are consequently low. Normally, the diameter of a hole or width of a slot needs to be equal to or exceed the metal thickness. A rule of thumb for PCM tolerance is ±10% of the metal substrate thickness. This guideline, however, does not express the ultimate capabilities of the PCM process because variable etch factors and reduced sheet size can enhance dimensional control.

Photoetching of metals requires moderate investments and competence in two fields: metal

TABLE 4.1 Example PCM Etchants and Materials That Are Etched by Them

Aqueous Etchants	Material Etched in the PCM Process
Acidified ferric chloride	Aluminum, Alloy 42, copper and copper alloys, HyMu, Inconels, Invar, Kovar, Permalloy, Monel, Mumetal, nickel, Nimonics, phosphor bronze, stainless steels and other steels, tin
Acidified cupric chloride	Beryllium copper, copper and copper alloys including brass and bronze, lead
Alkaline potassium ferricyanide	Aluminium, molybdenum, and tungsten
Sodium hydroxide	Aluminium, anodized aluminium
Hydrofluoric acid	Beryllium, columbium (niobium), titanium, zirconium, glasses, and ceramics

Source: Allen, D. M. 2004. Photochemical machining: from 'manufacturing's best kept secret' to a $6 billion per annum, rapid manufacturing process. *Annals of the CIRP*, 53: 559–572.

Table 4.2 PCM and Chemical Milling Compared

Fabrication Method or End Product	Typical Material Etched	Thickness of Material Etched	Underlying Material or Support
Photochemical machining (PCM)	Metals, glasses, or ceramics	2 mm (typical maximum)	Not applicable
Printed circuit boards (PCBs)	Electroformed foil (a conductor)	17.5 μm (typical)	Rigid epoxy fiber glass or other insulator such as flexible polyimide
Integrated circuits (ICs)	Silicon dioxide or silicon nitride	0.01–0.1 μm (typical)	Silicon
Chemical milling	Aluminum, titanium, and aerospace alloys	0.1–10 mm (typical)	Not applicable

etching and lithography. PCM is a fast, efficient, cost-competitive, burr- and stress-free method of producing metal components. Applying photochemical etching to large surface areas is a very expensive proposition,[30] and, in that case, chemical milling may be a better approach. Common industries served include aerospace, electronics, automotive, computer, and instrument manufacturing. Major applications are lead frames, PCBs, shadow masks for color televisions, heat sinks, shields, covers, lids, gaskets, springs, encoders, disks, grids, screens connectors, panels, shims, bus bars, and combs.

In Table 4.2 we compare PCM on metals, glasses, and ceramics with PCM on printed circuit boards and ICs and chemical milling in Al, Ti, and aerospace alloys.

A typical foundry for PCM in the United States is PhotoMachining, Inc. (http://www.photomachining.com/apps.html). Good information on Photochemical Milling is available at http://www.pcmi.org. Standard tool production time in these foundries is 2–3 days but can be several hours if necessary. Tooling is produced and stored in an environmentally controlled facility to ensure accuracy and quality. Modern digital photography uses customer-supplied IBM/Mac-based programs, as well as all AutoCAD versions and DXF and GbR files. In Figure 4.7 some typical PCM products are shown.

Photofabrication

Introduction

Photofabrication is a subtractive process in which ultraviolet exposure through a mask results in a strong modification of the solubility in an optically clear solid layer. This permits the direct photochemical production of thick 3D structures without the need for a separate pattern transfer to an underlying substrate. In photofabrication, photosensitive glasses, plastics, and ceramics are patterned into intricate 3D shapes. As opposed to most photolithography processes, where photoresist is a sacrificial layer only, in photofabrication the photosensitive material itself becomes a permanent part of the microdevice under construction. The photosensitive materials in photofabrication are permanent, and in the case of polymers this typically involves negative photoresists (see also Chapter 1). The method offers simple, inexpensive, high-aspect-ratio 3D microstructures for small to medium batch sizes without the need to resort to more expensive methods such as Si wet bulk micromachining or LIGA (see Chapter 10). The industrial application of photofabrication, creating sharp images in ceramics, plastics, and glasses, preceded the Si microfabrication era. Given the current interest in high-aspect-ratio 3D structures, this technique is worth revisiting.

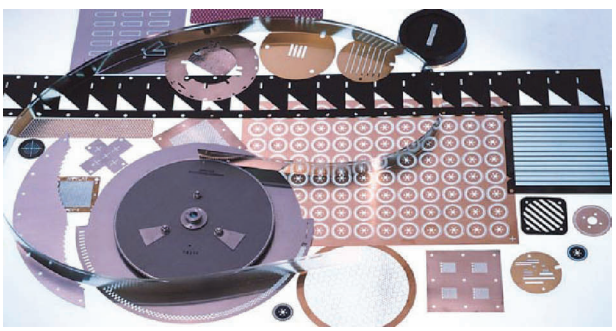


FIGURE 4.7 Mechanical metal parts produced by photochemical milling: leaf springs, connector arrays, masks, free standing grids, laser and IR target apertures, precision pin holes, and encoders.

For most photoplastics, aspect ratios do not exceed 3:1, but, with photosensitive glasses and ceramics, it is possible to make a 0.13-mm-diameter hole through a 2.7-mm-thick plate, with an aspect ratio of ~21:1; this is no small achievement. We will mostly detail photosensitive glasses here. A few words on photosensitive plastic sheets suffice because the topic of dry permanent resists was covered already in Chapter 1.

Note stereolithography, as discussed in Chapter 8, is an additive technique in which plastic shapes are generated by laser writing in a molten plastic mass; the technology is sometimes referred to as *microphotoforming* and should not be confused with the subtractive photofabrication discussed here.

Photosensitive Glasses/Ceramics—Photography in Glass

Electrochromic glasses temporarily change color when exposed to strong light. Photosensitive glasses, on the other hand, develop an invisible, permanent "latent image" after UV illumination through a mask. Subsequent heat treatment can develop the latent image into a permanent structural and/or color change with accompanying differences in solubility. In the mid-1940s, Donald Stookey invented both photosensitive and electrochromic glasses at Corning Glass Works (now Corning, Inc.) To read up on Stookey's career, consult *Profiles in Ceramics* (http://www.ceramicbulletin.org/months/Mar00/StookeyProfile.pdf)[31] and *Explorations in Glass: An Autobiography S. Donald Stookey.*[32] A typical photosensitive glass composition, FotoForm® glass, developed by Corning Glass Works, was SiO_2, 81.5; Li_2O, 12.0; K_2O, 3.5; Al_2O_3, 3.0; CeO_2, 0.03; and Ag, 0.02 (all in weight percent). On exposure to UV light, photoelectrons released by cerium, the photosensitizing ingredient, are trapped at specific sites in the silicate-glass network to form a latent image. By heating the exposed glass to around 500°C, the electrons are released from the traps, and silver ions are reduced to silver atoms, which aggregate into minute colloidal islands. The minute silver clumps serve as nuclei on which crystals, such as lithium metasilicate, sodium bromide, or sodium fluoride, precipitate. If sufficient numbers of fairly large crystals form (~4 μm), an opalescent image results, whereas the masked regions remain glassy and transparent. In HF, opal regions etch 15–30 times faster than the unexposed glass. No masking is required during the etching process. In this way, by exposing a photosensitive glass plate to UV and etching in HF after heat treatment, a wide variety of microstructures can be made. Through-holes in photosensitive glass were experimentally found to have a sidewall taper of only 2–3°; this means that substantially parallel holes may be cut, having a very high depth-to-diameter ratio. The sloping effect can be reduced even further if the HF etching occurs from both sides of the glass simultaneously. Etching from both sides produces a profile with two truncated cones and gives a minimum overall increase of hole diameter. Although etching of holes with diameters appreciably less than one-tenth of the thickness of the glass is not recommended, aspect ratios of 20 are possible. In contrast, in photoetching of metal, as we have seen above, holes cannot be made with a diameter less than the thickness of the sheet. If sidewall taper in photosensitive glass is desired, it can be increased to 20–30°.

Further heat treatment of the finished photosensitive glass pieces, Stookey discovered by accident in 1957, converts the glass to a stronger, partially crystalline material, e.g., Fotoceram®. Stookey left a sample of lithium silicate glass containing silver in an oven at too high a temperature. He expected to find a puddle of glass once he realized his mistake, but instead found a piece of white ceramic because the glass had crystallized with a fine grain size.

The maximum sheet thickness of FotoForm glass available was 0.6 cm, but several pieces can be thermally bonded, forming excellent seals and converting the glass to Fotoceram glass. One disadvantage of the bonding process is that it causes the photoglasses to become dark, which makes the approach less attractive for optical applications.

Corning's photosensitive FotoForm glass found an important application in fluidic logic in the mid-1960s (see for example Figure 4.8), but as that application started fading the production of FotoForm glass and Fotoceram unfortunately was discontinued.[31]

Fluidic components, providing functions such as flip-flops, logic elements such as OR-NOR and AND

gates, proportional amplifiers, and fluid resistors, exemplified the capabilities of photofabrication with FotoForm glass in the mid-1960s. The fluidic-control device in Figure 4.8 shows the individual layers of an AND device after through-etching.[1]

Fusion of stacks of photosensitive glass plates provided a permanently sealed and rugged unit, complete with the necessary input and output vents. Through-etching provides the connection between the different planes. The geometrical accuracy of these fluidic devices was maintained within ±0.025 mm, and the tolerance that may generally be expected on this type of etched glass product was ±0.05 mm on etched edges up to 12.7 mm in length. With hole spacings up to 25 mm apart, accuracy to within 0.025 mm center to center was possible.

Few of the Si and plastic microfluidic structures produced today have reached the degree of complexity achieved with photofabrication methods of almost four decades ago when they were applied to fluidic logic (see Figure 4.8). "Fluidic" elements made in the 1960s could achieve many of the functions of an electronic circuit, and they were robust, simple to fabricate, and suited to operating in a hostile environment. Of course, they were much slower (milliseconds instead of nanoseconds) and much larger than their electronic counterparts. The latter spelled the eventual demise of microfluidic "computing" in the late 1960s. In the 1990s, several micromachined fluidic devices were developed for analytical/biomedical applications, without reference to this earlier sophisticated work in photosensitive glasses. Early in the development of these analytical/biomedical applications, many fluidic devices were attempted in Si. However, the relatively large size of fluidic elements, the need for a robust environment, the relatively low numbers of required devices, and the fact that electronics usually do not need to be integrated render the use of silicon for this application hardly justifiable. The photofabrication machining approach, which is inexpensive and simple, coupled with the chemical inertness of glass, represents a more attractive approach. Fluidic oscillators, fabricated using the more expensive and hardly accessible LIGA technique, are even more difficult to justify when compared with photoformed devices.[33,34] The latter misapplication of miniaturization tools was eventually recognized, and photoformed glass was reintroduced as the best solution for this application by one of the inventors of LIGA technology[35] (see Volume III, Figure 1.2). In July 1996 the commercial entity "Mikroglas" was founded out of the Institute of Microtechnology in Mainz, Germany. Mikroglas obtained the exclusive rights for the use of Foturan™ glass from Glaswerk-Schott (http://www.us.schott.com/english). In 1999, Mikroglas started with the development of microfluidic components for chemical and biotechnological applications. Products include chemical microreactors and whole equipment to use these microreactors (http://www.mikroglas.com/index_e.html).

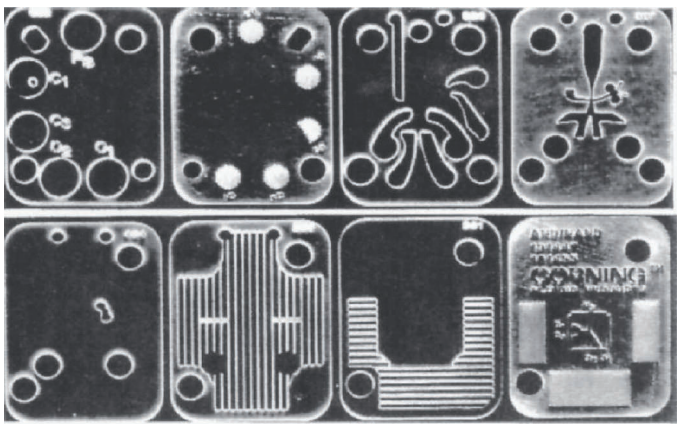


FIGURE 4.8 Fluidic-control device using chemically milled photosensitive glass. The individual layers of an AND device after through-etching with HF. (From Harris, T. W. 1976. *Chemical milling.* Oxford: Clarendon Press. With permission.[1]) See also Volume III, Figure 1.2.

Unlike ceramics, Foturan is pore-free, has higher temperature stability and chemical resistance than plastics, and has a higher breaking strength than silicon. Processing of Foturan consists of exposure to ultraviolet light, crystallization through heat treatment, and etching as illustrated in Figure 4.9a. In Figure 4.9b we show example hole plates made this way. The thermal expansion coefficient of the processed Foturan can be precisely varied between 8 and 16 ppm/°C. Variously shaped recesses can readily be formed on both sides of the "wafer," as well as 40:1 aspect ratio through holes. In addition, the surfaces can be metallized and patterned by conventional techniques. The material can be processed in sheets of up to 250 mm × 200 mm. Besides for making photosensitive glass ornaments,

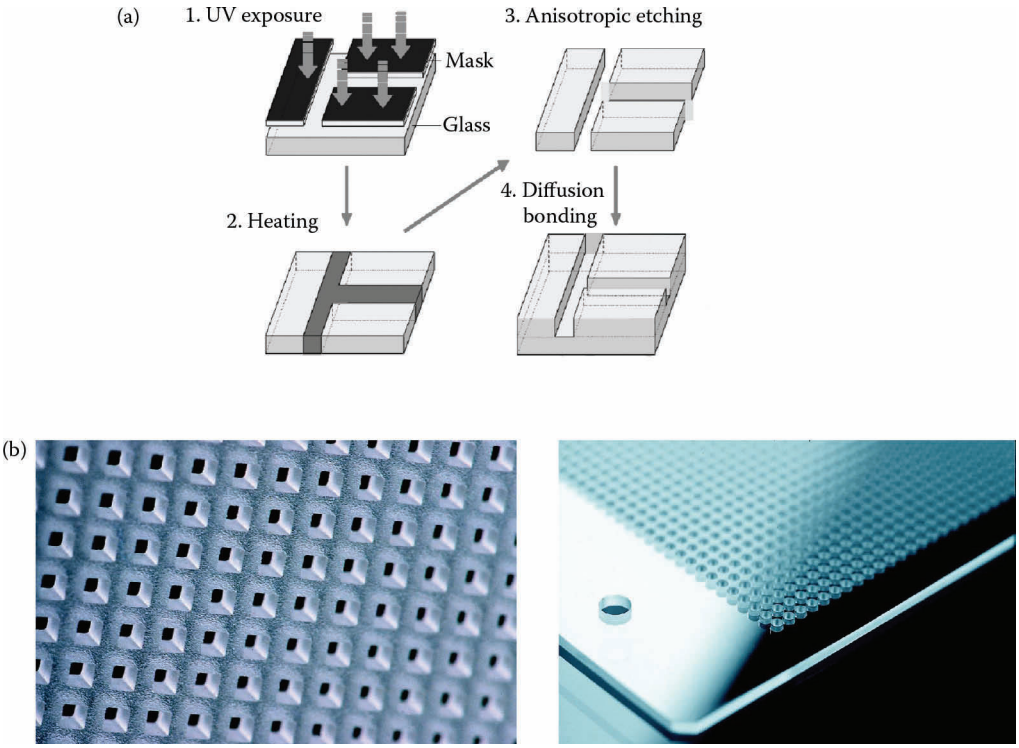


FIGURE 4.9 (a) The photofabrication process in Foturan. (b) Hole plates shapes, round or rectangular; diameter of holes, down to 50 μm; thickness of plates, 0.1 up to 2.5 mm and up to 100 holes/mm^2 (http://www.mikroglas.com).

photosensitive glass has been used to make microlenses (±100 μm in diameter), fluidic elements (see Figure 4.8), spacers in photomultiplier tubes, color filters, integrated optics cell sheets in gas-discharge displays, and charge plates and nozzles in ink-jet printers.

In a very intriguing further development of photosensitive glasses, it was found that the different properties of opal and glass areas on photosensitive glass enable the direct formation of electrical circuitry on glass. Immersing a sample with an opal pattern in a molten salt bath and heating it in a hydrogen atmosphere forms a conducting silver film on the opal regions. The film that forms on the glassy regions remains nonconducting.[31] This discovery could be used in the construction of DC and AC electrokinetic devices with complex patterns of metal electrodes on a glass substrate (see Volume I, Chapter 7).

Suggestion: A most useful new MEMS material would be a Foturan-type photosensitive glass with a thermal expansion coefficient matching that of Si (like Corning 7740 glass used in anodic bonding). Such a product would enable the packaging of all types of Si sensors (see Volume III, Figure 4.9) combining a Si micromachined sensing element with a photofabricated glass package.

Photosensitive Plastics

Although dry photoresists and photosensitive plastics are covered in Chapter 1 and Volume III, Chapter 4, some additional photosensitive plastics are briefly covered here. In the early 1960s, DuPont started developing a photosensitive nylon sold in sheet form, either bonded to a substrate such as carbon steel, aluminum, or a polyester Cronar film, or freestanding. The alcohol-soluble nylon is made photosensitive by adding a cross-linking agent such as methylene bis-acrylamide and an activator such as benzophenone. On exposure to UV, the benzophenone forms free radicals, which react with both of the other compounds, hence insolubilizing the nylon; in other words, the material acts as a negative tone resist. The unexposed areas are then washed away with a dilute aqueous sodium hydroxide spray, forming channels and cavities. By using the benzophenone in a low concentration, about 5% by weight, a layer with a relatively low absorption coefficient results, and thick layers can be insolubilized.

Stencils with 1-mm-high vertical sidewalls with a taper of 2° can be produced this way, and 25-µm-deep, 25-µm-wide channels do not show any undercutting. The aspect ratio did not exceed 3:1, and the operating temperature range is limited to –100 and +200°F. To seal cavities or channels (e.g., for the fabrication of fluidic devices), the first nylon sheet can be laminated with another Cronar film (http://www2.dupont.com/Products/en_RU/Cronar_en.html).[28]

Riston®, Fodel®, and VACREL® photopolyer films are other DuPont dry photoresists (see Figure 1.33). These complex photosensitive systems are designed to polymerize and harden when exposed to ultraviolet light. Fodel can pattern 75-µm vias on a 150-µm pitch, and conductors can be patterned with 50-µm lines on a 100-µm pitch. Fodel extends the density capability of the thick film process to allow densities typically achievable using more costly thin film processes.

Negative-tone photosensitive resins such as SU-8 are particularly suited for high-aspect ratio, high-thickness photofabrication. As we saw in Chapter 1, SU-8 is an epoxy-type, near-UV photoresist based on EPON SU-8 epoxy resin (from Shell Chemical) that was originally developed and patented by IBM. The main advantage of photosensitive resins over glasses and plastics is the fact that they are deposited in liquid form, thus conforming better to existing 2D and 3D structures. Other advantages include exquisite control over the thickness of the layer, higher attainable aspect ratios, higher feature density, and easier alignment to existing features.

Conclusions

Photofabrication is an attractive alternative for microfabrication of all types of high-aspect-ratio, passive elements, especially for intricate shapes that are >1 cm^2 in size and are only produced in modest volumes. The current interest in micromachined fluidic elements for analytical purposes is reviving interest in photosensitive glasses, plastics, and resins. Photosensitive glasses are also being considered for the fabrication of photonic crystals (see Volume I, Chapter 5). It certainly makes more sense to use photofabrication of insulators than to fabricate fluidic elements from Si because the latter, after etching, must be passivated. This passivation may undergo electrical breakdown, especially in an aqueous environment. We expect dry photoresist formulations to continue to improve, with tolerances approaching those of traditional spin-on photoresists (see Chapter 1). Applications such as RF-MEMS and fluidics might also lead to the introduction of novel photoformable ceramics and glasses.

Electrochemical Removal Techniques

Introduction

The term *electrochemical machining* is used to describe electrochemical removing and forming processes with or without masking the substrate. Both processes involve electrochemical energy. Despite performance, speed, and cost advantages, electrochemical milling (ECM) is used little in the IC field when other techniques, especially dry processes, are available. This seems to be connected in part with the overall trend toward dry processing and the disproportionately large number of electrical engineers versus chemical and electrochemical engineers in the IC world. LIGA and other replication techniques described in Chapter 10 have reinvigorated interest in electrodeposition, however. Both electrochemical removing and forming techniques are of extraordinary importance in miniaturization science—without electrochemical removal there would have been, for example, no porous Si (PS) work, and without electroplating of Ni master molds, there would be no plastic micromolding. Here we will review subtractive electrochemical processes only. Additive electrochemical processes are covered in Chapter 8. In electrochemical material removal, an electrical field in an electrolyte destroys the atomic bonds of the material.[26] Under electrochemical removal techniques, we will review electrochemical grinding (ECG), ECM including µ-ECM, electrochemical jet etching, laser-assisted electrochemical jet micromachining, scanning electrochemical microscope (SECM), and through-mask electrochemical micromachining.

Electrochemical Grinding

Introduction Grinding usually constitutes a mechanical machining process that removes small

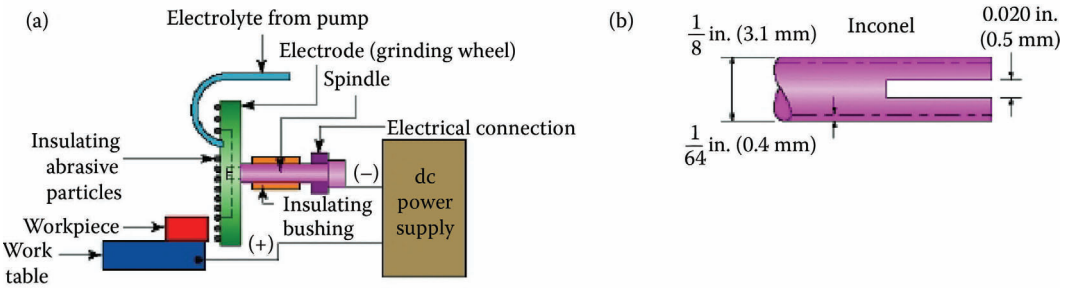


FIGURE 4.10 (a) Schematic illustration of the electrochemical grinding process. (b) Thin slot produced on a round nickel alloy tube by this process.

amounts of material from a metallic work piece in the form of tiny chips through the contact of small, hard, sharp, nonmetallic particles often embedded in a grinding wheel (see Chapter 6). In electrochemical grinding (ECG), the abrasive action of an electrically conductive wheel, the cathode, accounts only for 10% of the metal removal; the remainder is electrochemical. This technique should not be confused with electrolytic in-process dressing or ELID, where the material removal process (any type of material in this case) is purely mechanical but an electrochemical dissolution process of the metal matrix dresses the abrasive wheel continuously (see Chapter 6).

Electrochemical Grinding Operating Principles In ECG, an electrically conductive abrasive wheel, the cathode, and an anode work piece are connected to a low-voltage DC power source (Figure 4.10).[36]

An electrolyte is flushed onto the abrasive wheel, and the abrasive particles in the wheel's surface remove electrochemically dissolved material, always keeping the surface fresh for further electrochemical attack. Very low DC voltages (usually between 3 and 15 V) are used in ECG; therefore, spark discharge is minimized and little or no heat is developed in the cut. All hard conductive materials are candidates for ECG (e.g., sharpening of tungsten-carbide tool bits). The method is also good for fragile parts that would be damaged by conventional mechanical grinding. The method produces parts free of residual stresses and heat damage and surface finishes ranging from 0.1–0.5 μm. The surfaces are also free of scratches and burrs. This process provides not only burr-free machining but also considerably less cutting wheel wear. ECG is similar to electrochemical machining (see below) except that the cathode is a specially constructed grinding wheel instead of a tool shaped like the contour to be machined.

Compositron Corporation is a service company providing an electrochemical grinding homepage (http://www.evercut.com), and Everite Machine Products Co. sells ECG equipment (e.g., Everite Model EG618; http://www.everite.net).

Electrochemical Machining

Electrochemical machining (ECM) historically followed electrochemical grinding (ECG). In ECM one uses a cathode electrode shaped to provide the complementary structure in an anode work piece (Figure 4.11). A highly conductive electrolyte stream separates the cutting tool from the work piece, and metal removal is accomplished by passing a DC current of up to 100 A/cm^2 through the salt solution cell. As the cathode tool approaches the anode work piece it erodes its complementary shape in it. Thus, complex shapes may be made from a material such as

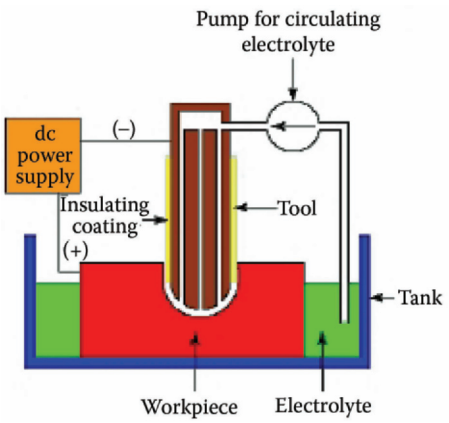


FIGURE 4.11 Schematic illustration of the electrochemical machining process. This process is the reverse of electroplating, described in Chapter 8.

soft copper and used to produce negative duplicates of it. The process is also called electrochemical sinking. Machines having current capacities as high as 40,000 A and as low as 5 A are available. The pressurized electrolyte (concentrated solutions of inorganic salts such as sodium chloride, potassium chloride, and sodium nitrate) passes at high speed (10–60 m/s) through the gap (about 0.1–0.6 mm) between the work piece and the tool to prevent metal ions from plating onto the cathode tool and to remove the heat that is generated as a result of the high current flow. The cathode is advanced into the anode work piece at a rate matching the dissolution rate, which is between 0.5 and 10 mm/min when applying current densities of 10–100 A/cm^2. The supply voltage commonly used in ECM ranges from 5–20 V; the lower values are used for finish machining (creating a final smooth surface), and the higher voltages are used for rough machining. The rate of material removal is the same for hard or soft materials, and surface finishes are between 0.3 and 1 μm. These cutting speeds and surface finishes are comparable with those of electrical discharge machining (EDM; a thermal subtractive technique described in Chapter 5).

The cathode tool must have these four characteristics: be machinable, be rigid (high Young's modulus), be a good conductor, and have good corrosion resistance. The three most common cathode materials used are copper, brass, and stainless steel. The outside of the electrode is insulated to prevent machining actions on the sides. This makes the fabrication of these tools challenging. Because there is no contact between the tool and the work, the tool does not have to be harder than the work, as in traditional machining methods. Hence, this is one of the few ways to machine very hard material; another is spark-discharge machining (see Chapter 5).

The advantages of ECM are that it has relatively low running costs, as there is little scrap, no tool wear, and unskilled labor or automatic machine operation can be used. Because shear is not involved in this process, the surface of the work is, as in chemical milling, free from imperfections and burrs. It is also used in deburring already machined components. A very good surface finish is usually obtained. The disadvantages of ECM are that the machine is initially very expensive—several times the cost of an EDM machine; large systems may cost up to $600,000 (about the same as multiaxis milling machines). The tool is also difficult to make because its sides have to be insulated. The electrolyte has to be well contained, and the most common electrolyte, sodium chloride, corrodes the equipment, tool, and work. Another issue is that the electrochemical erosion reduces sharp surfaces and profiles so the method is not suited for making products with sharp edges. Also, irregular cavities may not be produced to the desired shape with acceptable dimensional accuracy, and designs should make provisions for small taper for holes and cavities to be machined. Overall, the technique is best suited for mass production of high-precision, small, and complex shapes in hard-to-machine materials (accuracy better than 10 μm).[37] To date, ECM has found its widest acceptance in the aircraft, automotive, and business machine industries. In the aircraft industry, ECM is used extensively for shaping turbine blades. In the automotive industry, the technique is used to make connecting rods, pistons, and fuel-injection nozzles. Some example products are shown in Figure 4.12.

The fact that metal removal in ECM is not achieved by mechanical shearing (as in mechanical machining) or by melting and vaporization of the metal (as in EDM or laser machining; see Chapter 5) means that, as with chemical machining, no thermal damage occurs and no residual stresses are produced on the worked surface. For this reason, ECM is often selected for highly stressed or fatigue-sensitive applications in the aerospace industries.

Pulsed electrochemical machining (PECM) is a refinement of ECM, where current pulses are used instead of DC currents. This improves fatigue life and eliminates the recast layer left on die and mold surfaces by EDM (see Chapter 5).

Sometimes, chemical and electrochemical removal machining are in competition. Chemical machining might be preferred when very large work pieces have to be machined because of the large currents involved in ECM, and also for ultrasmall parts, where the electric connection causes problems for ECM.

Typical ECM service providers are Köppern (http://www.koeppern.com/en_ecm_set.htm) and Raycon/AMCHEM (Ann Arbor, MI). A handbook

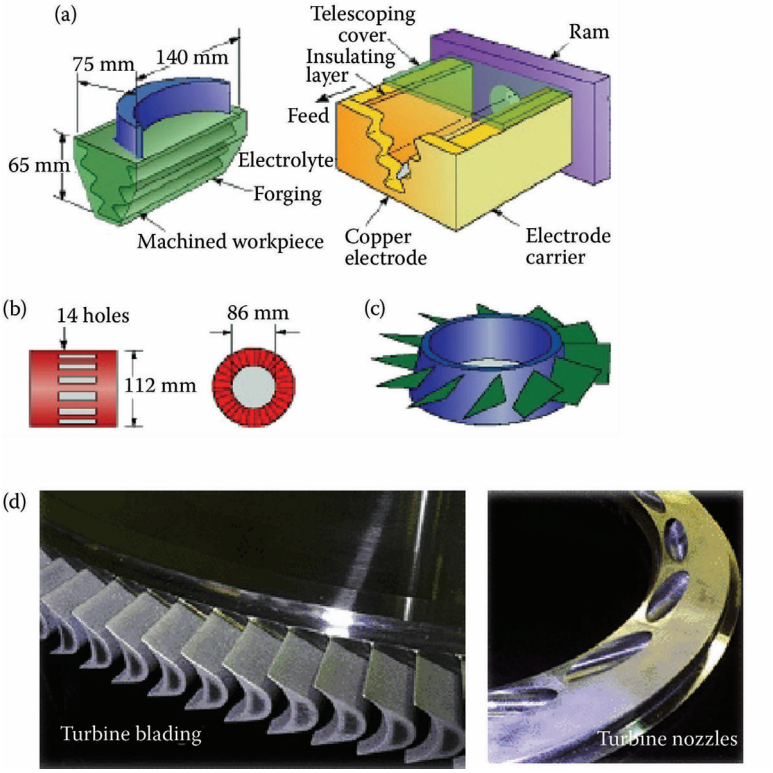


FIGURE 4.12 Typical parts made by electrochemical machining. (a) Turbine blade made of a nickel alloy, 360 HB. (From *Metal handbook,* 9th ed., vol. 3. Materials Park, OH: ASM International, 1980, p. 849). (b) Thin slots on a 4340-steel roller-bearing cage. (c) Integral airfoils on a compressor disk. (d) Because of the low forces on the work piece, ECM can be used to make holes at very large angle to a surface; in the example, we show turbine nozzle holes (http://www.barber-nichols.com).

covering ECM is the *Non-Traditional Machining Handbook* by Carl Sommer.[38]

Electrochemical Micromachining or Microelectrochemical Machining

Introduction The application of ECM in thin film processing and in the fabrication of microstructures is referred to as *electrochemical micromachining* (EMM) or *microelectrochemical machining* (μ-ECM).[39] Different from ECM, the cathode does not necessarily have the shape of the contour desired in the anode work piece. The 3D shaping in EMM may involve maskless or through-mask material removal, and we will examine examples of both approaches. The tool may also be connected to a CNC machine to produce even more complex shapes with a single tool as illustrated in Figure 4.13.

In conventional ECM, the gap between cathode tool and anode work piece is typically about 150 μm; in μ-ECM, the gap is closer to 15–20 μm, and feature sizes change from 150–200 μm to 15–20 μm as we move from the ECM to the μ-ECM domain. The major challenge in moving from the conventional ECM to the μ-ECM domain is to control the size of the reaction region. Methods to accomplish this include:

1. Reduce the size of electrodes—μ-EDM is used.
2. Shield the electrode—for stray currents.
3. Gap control strategies.
4. Use ultrashort-pulsed voltages having time duration in the ranges of nanoseconds.

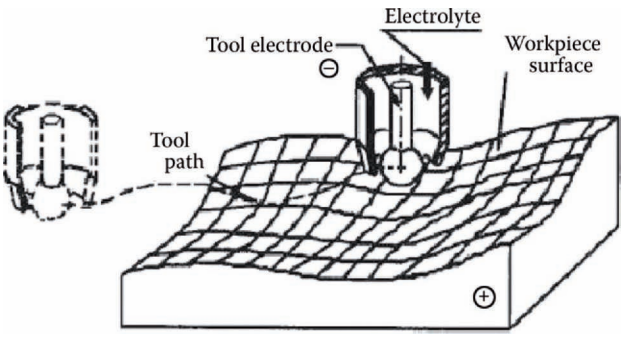


FIGURE 4.13 The ECM microcathode is connected to a CNC machine to produce more complex shapes with a single tool.

With EMM, most metals, alloys, and conducting ceramics of interest in the microelectronics and MEMS/NEMS industry can be anodically dissolved in a variety of neutral salt electrolytes, such as sodium nitrate, sulfate, or chloride. The dissolved metal ions form hydroxide precipitates in these electrolytes and remain suspended in solution so they may easily be filtered out. The latter minimizes safety problems and waste disposal. With hydrogen evolution generally as the main cathodic reaction, metal plating on the cathode is avoided. EMM is currently receiving considerable attention in the electronics and other high-technology industries, particularly as an alternative, "greener" method of processing metallic parts.[39] Compared with chemical etching, the EMM process offers better control, less aggressive chemicals, and thus minimizes safety and environmental concerns.[39]

Electrochemical Jet-Etching and Laser-Assisted Electrochemical Jet-Etching Thin film patterning by maskless EMM may be accomplished by highly localized material removal induced by the impingement of a fine electrolytic jet emanating from a small nozzle.[39] Investigation of jet EMM at IBM demonstrated that neutral salt solutions can be effectively used for high-speed micromachining of many metals and alloys. An example of such an investigation, using a nozzle diameter of 200 µm and a 5 M sodium nitrate solution, is shown in Figure 4.14. Through-slots were micromachined in a 50-µm-thick stainless steel foil to create the IBM logo, which demonstrates the feasibility of using an electrolytic jet for generating complicated patterns in metallic foils and substrates.

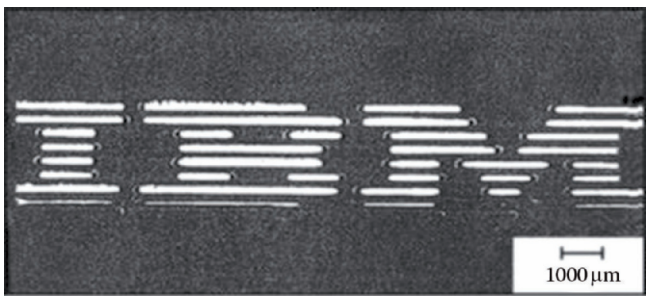


FIGURE 4.14 IBM logo made by jet EMM. Through-slots machined in a 50-µm-thick stainless sheet using a 5 M sodium nitrate electrolyte jet. The jet nozzle diameter is 200 µm. (Courtesy of Datta, M. J., 1995. Fabrication of an array of precision nozzles by through-mask electrochemical micromachining. *Electrochem Soc* 142:3801–05.[44])

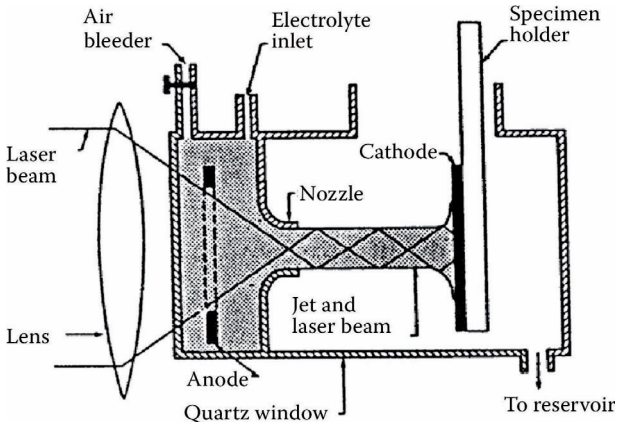


FIGURE 4.15 Schematic representation of laser-enhanced electrochemical jet etching/plating showing the coincident laser beam and solution jet with the jet acting as the waveguide. (From Romankiw, L. T., and T. A. Palumbo. 1987. *Proceedings: symposium on electrodeposition technology, theory and practice*. Pennington, NJ: The Electrochemical Society, Inc. With permission.[27])

An interesting variation on electrochemical jet etching is a combination of a fluid-impinging jet and laser illumination (see Figure 4.15). In laser-enhanced electrochemical jet etching, properly chosen lasers, whose energy is not absorbed by the etching solution but is absorbed by the solid, cause local heating of the substrate (up to 150°C), resulting in highly increased etching. The jet is used as a light pipe for the laser and at the same time as a means for the local high rate of supply of ions. For stainless steel, etch rates of 10 µm/s have been demonstrated using laser-enhanced electrochemical jet machining.

For more background on electrochemical jet etching and laser-enhanced jet etching, refer to Romankiw et al.[27] and references therein and Datta et al.[40] Several of these techniques offer tremendous opportunities for the development of alternative micromachining methods.

Water jet etching is a mechanical process, and as such it is covered in Chapter 6 on mechanical energy-based removing. Water jet-guided laser etching without the electrochemical component is a purely thermal technique and is covered in Chapter 5. In this important method we will see that a fine water-jet again guides the laser beam, provides cooling for the work piece, and expels the molten material.

Scanning Electrochemical Microscope The scanning electrochemical microscope (SECM) is a scanned probe microscope (SPM) related to the familiar scanning tunneling (STM) and atomic force microscopes (AFM) (see Volume III, Chapter 6). All SPMs operate by scanning or "rastering" a small probe tip over the surface to be imaged. In SECM, imaging occurs in an electrolyte solution with an electrochemically active tip. In most cases, the SECM tip is an ultramicroelectrode (UME), and the tip signal is a Faradaic current from an electrochemical reaction at the surface. In some SECM experiments, the tip is an ion-selective electrode (ISE). In Chapter 1, we saw how a scanning tunneling microscope may induce electrochemical anodization reactions (an oxidation) on surfaces such as Si and GaAs. An SECM can also be used for local etching and deposition with high resolution in the x, y, and z dimensions, basically forming a high-resolution ECM setup (Figure 4.16).

Metal and semiconductor surfaces can be etched by the generation of a suitable etchant at the tip, for example Br_2, which will etch holes in GaAs.[41] In many cases, dissolution of semiconductors can be further enhanced by irradiation of the surface, creating holes in the valence band, which react with the lattice of the semiconductor. Besides electrochemical micromachining, SCEM is also used for chemical microscopic imaging and the measurement of physicochemical constants and coefficients.

In 2008, a review on the technology appeared in the *Annual Review of Analytical Chemistry.*[42] For an update on SECM on the Internet, visit http://www.msstate.edu/Dept/Chemistry/dow1/secm/secm.html and http://www2.warwick.ac.uk/fac/sci/chemistry/cim/research/electrochemistry/research/secm.

Through-Mask Electrochemical Micromachining Through-mask electrochemical micromachining (EMM) involves selective metal dissolution from unprotected areas of a one- or two-sided photoresist-patterned work piece that is made the anode in an electrolytic cell.[39,43] The technique is faster than chemical milling and provides better control on a micro- and macroscale of shape and surface texture of anodically dissolved materials. The performance of through-mask EMM depends largely on the successful control of voltage distribution in the cell, hydrodynamics, and surface kinetics.[43] In wet isotropic chemical etching, the material is removed both vertically and laterally at the same rate, and the etch boundary usually recedes at a 45° angle relative to the surface. In EMM, however, the metal-removal rate in the lateral direction may be significantly reduced through proper manipulation of mass transport and current distribution.[39]

Datta[44] described a cost-effective, high-speed process for the fabrication of precision nozzles for ink-jet printer heads using through-mask EMM. The process involves fabrication of a series of rombezoidal nozzles in a metal foil and is applicable to a variety of materials, including high-strength, corrosion-resistant materials such as conducting ceramics. For the

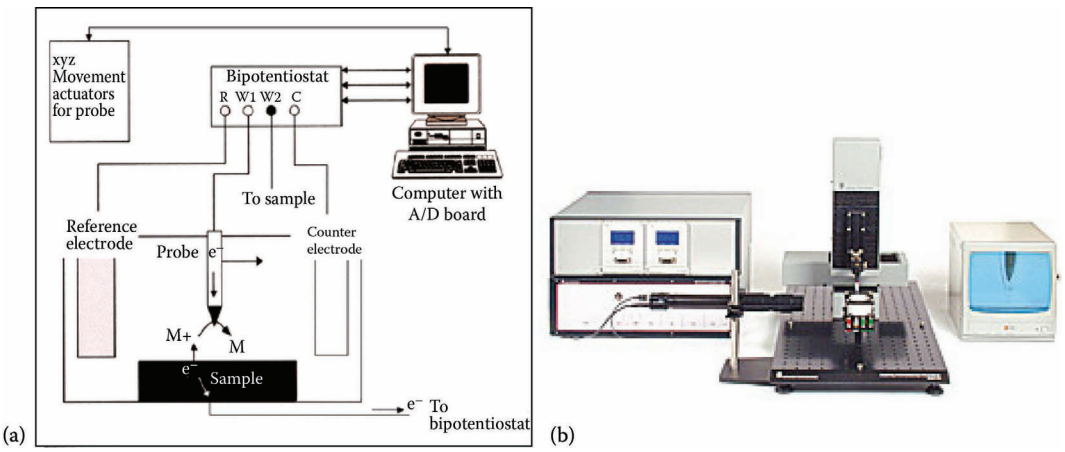


FIGURE 4.16 (a) A schematic setup for SECM (http://www.msstate.edu/Dept/Chemistry/dow1/secm/secm.html) and (b) a typical commercial product (http://www.princetonappliedresearch.com/products/SurfaceScanning.cfm).

fabrication of the example nozzles in Figure 4.17, a cleaned 25-µm-thick stainless steel foil was laminated on both sides with 25-µm-thick dry photoresist. On one side, the photoresist was then exposed and developed, and the photoresist on the back remained as a protective insulating layer. The photoresist pattern was an array of circular openings 55 µm in diameter. Direct- and pulsed-voltage experiments were then performed with a neutral salt solution of sodium chloride and glycerol mixture as the electrolyte. Pulsating-voltage EMM was found to be more effective in providing dimensional etching uniformity over the whole nozzle array. This result is based on the application of extremely high peak currents (voltages), which, in addition to giving directionality, enables breakdown and elimination of inhibiting layers, thus facilitating activation of all the openings at the same time. A desired nozzle angle of 27° was produced with an etch factor (ratio of straight-through etch to undercut) of 2 (see Figure 4.17). This way, Datta demonstrated the feasibility of fabricating arrays of hundreds to thousands of precision nozzles with microsmooth surfaces in copper and stainless steel foils. Alternate ink-jet nozzle fabrication techniques in silicon are shown in Figure 4.45.

Through-mask EMM has also been used to fabricate controlled topographies on titanium to improve its biocompatibility.[45] Arrays of very smooth, high-precision cavities 30 µm in diameter were fabricated. Such structures affect attachment and growth of both soft and hard tissues. Biocompatibility has been linked to surface structures in the micrometer and nanometer range (see also Volume III, Chapter 4).[46]

Summary The driving force for metal dissolution reactions in EMM is derived from an external current; as a consequence, the electrolyte does not have to be as aggressive as in chemical etching. In fact, even neutral salt solutions are applicable. Metal-removal rates in EMM are determined by the machining current, which can be very high if proper hydrodynamic conditions prevail within the electrochemical cell; hence EMM etch rates are generally orders of magnitude higher than those in chemical etching. More importantly, through manipulation of mass transport and current distribution, metal-removal rate in the lateral direction can be significantly influenced, thus providing better wall slope control in EMM. Other EMM benefits include minimized safety and environmental concerns. All the listed features make the EMM process a cost-effective, environmentally friendly processing technology. Newly emerging technologies such as advanced microelectronic packaging, microsensors, and MEMS/NEMS provide

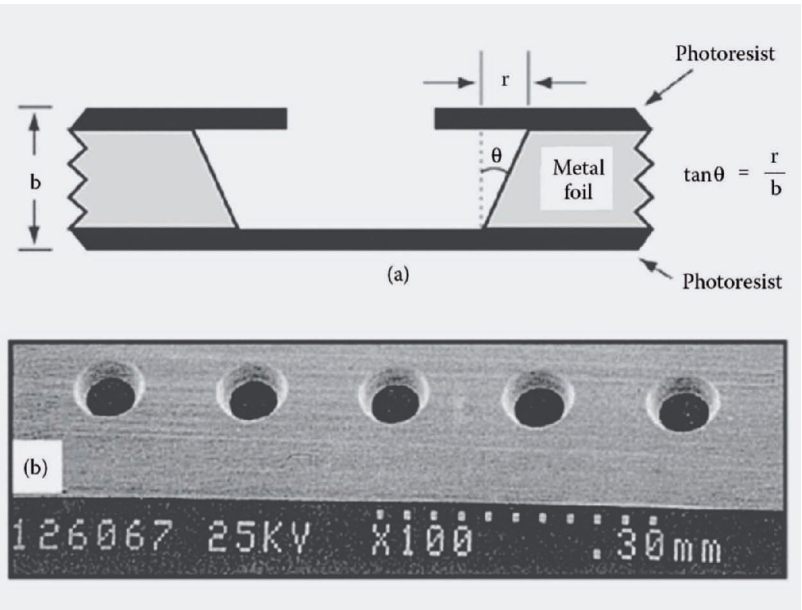


FIGURE 4.17 (a) One-side through-mask EMM with angular sidewall. (b) SEM of nozzles fabricated in 25-µm-thick stainless steel. (Courtesy of Datta, M. J., 1995. Fabrication of an array of precision nozzles by through-mask electrochemical micromachining. *Electrochem Soc* 142:3801–05.[44])

Table 4.3 Comparison of Electrochemical Machining (ECM) with Microelectrochemical Machining (μ-ECM)

Major Machining Characteristics	ECM	μ-ECM
Voltage	10–30 V	<10 V
Current	150–10,000 A	<1 A
Current density	20–200 A/cm^2	75–100 A/cm^2
Power supply, DC	Continuous/pulsed	Pulsed
Frequency	Hz–KHz range	KHz–MHz range
Electrolyte flow	10–60 m/s	<3 m/s
Electrolyte type	Salt solution	Natural salt or dilute acid/ alkaline solution
Electrolyte temperature	24–65 °C	37–50 °C
Electrolyte concentration	>20 g/L	<20 g/L
Size of the tool	Large to medium	Micro
Interelectrode gap	100–600 μm	5–50 μm
Operation	Maskless	Mask/maskless
Machining rate	0.2–10 mm/min	5 μm/min
Side gap	>20 μm	<10 μm
Accuracy	±0.1 mm	±0.02–0.1 mm
Surface finish	Good: 0.1–1.5 μm	Excellent: 0.05–0.4 μm
Problems caused by waste disposal/ toxicity	Low	Low to moderate

ample opportunities for further investigation of EMM. With a better understanding of the principles involved in manipulating high-rate anodic dissolution of metals from narrow and deep cavities, EMM will become an increasingly important processing technology for micro- and nanofabrication.[39]

In Table 4.3 we compare electrochemical machining (ECM) with microelectrochemical machining (μ-ECM).

μ-ECM has some features in common with microelectrical discharge machining (μ-EDM), a thermal technique covered in Chapter 5. In Table 5.2 μ-ECM is compared with μ-EDM.

Wet Bulk Micromachining

Introduction

As mentioned in the introduction to this chapter, wet bulk micromachining is chemical etching used to sculpt isotropic and anisotropic 3D microfeatures in single crystal semiconductor substrates. Removal of material from the surface of a crystalline material (such as Si), immersed in an alkaline or acidic solution, occurs as a result of complicated chemical reactions between the surface atoms and the etchant molecules (see Figure 4.18a). The isotropy or anisotropy of the etching refers to the orientation dependency of the etch rate: for the isotropic case, there is no orientation dependence; for the anisotropic case, there is orientation dependence (see Figure 4.18b).

Wet bulk micromachining is a technique championed in the IC industry and in micro- and nanomachining. In this section, we are mostly concerned with lithography-based chemical etching for the IC industry and micro- and nanofabrication. In this application, lithography is used to create a mask through which an etchant can etch away the underlying crystalline substrate. A major difference between the IC and MEMS fields in this regard is in the aspect ratio (height-to-width ratio) of the features crafted. In the IC industry, one deals mostly with very small, flat structures with aspect ratios of 1–2 (see Figure 4.2). In the microfabrication field, structures typically are somewhat larger, and aspect ratios (height-to-width ratio) may be as high as 400. In wet bulk micromachining, features are sculpted in the bulk of materials such as silicon, quartz, SiC, GaAs, InP, and Ge by orientation-independent (isotropic) or orientation-dependent (anisotropic) wet etchants. The technology uses pools of liquid as tools,[1] instead of the (dry) plasmas studied in

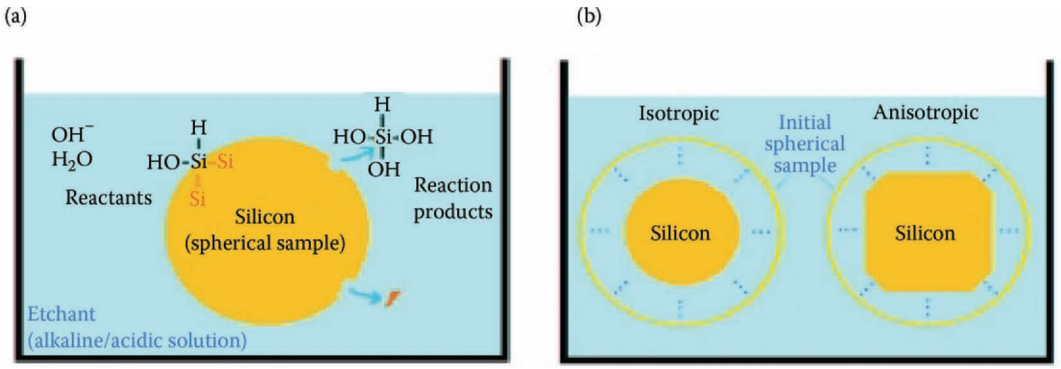


FIGURE 4.18 (a) Schematic illustration of the dissolution process during wet etching. (b) Wet etching can be isotropic or anisotropic. (Gonsalvez, M. 2003. *Atomistic modelling of anisotropic etching of crystalline silicon.* PhD diss., Helsinki University of Technology.[47])

Chapter 3. In Figure 4.19 we show some examples of high-aspect-ratio MEMS structures created using wet anisotropic etching of properly masked <100>- and <110>-oriented silicon wafers.

A vast majority of wet bulk micromachining work is based on single-crystal silicon and glass as a companion material [7740 Jena glass (Pyrex) is often used for packaging; see Volume III, Chapter 4], mostly for mechanical-type sensors such as pressure, acceleration, and gyros. There has been some work on quartz, crystalline Ge, SiC, and GaAs, and a minor amount on GaP and InP.

Wet bulk micromachining, together with surface micromachining (Chapter 7), form the two principal commercial Si micromachining tool sets used today. A third technique, micromolding—often from a lithography-defined master—also is becoming more commercially viable (Chapter 10).

The emphasis in this chapter is on the wet etching process itself. Other machining steps typically used in conjunction with wet bulk micromachining include many of the alternative subtractive techniques described elsewhere in Part II, the additive processes described in Part III, and the bonding processes that are part of the packaging processes dealt with in Volume III, Chapter 4.

A historical note on Si wet bulk micromachining, clarifying the importance of Si as a sensor material,

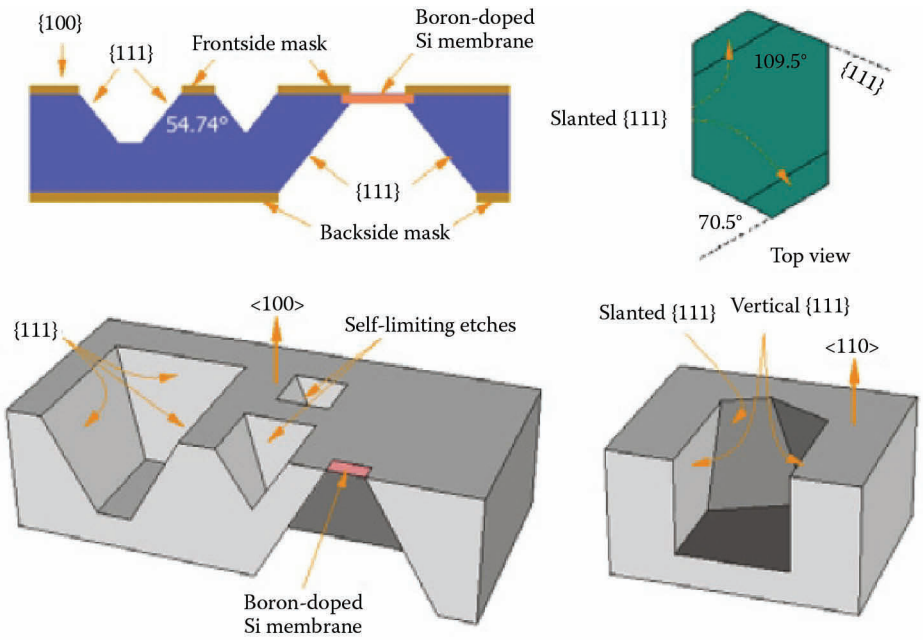


FIGURE 4.19 Examples of structures created using wet anisotropic etching in a Si <100> wafer and in a Si <110> wafer. The significance of boron doping will be clarified further below.

heads this chapter (an expanded treatise on this subject can be found in Volume I, Chapter 1). We start the current section with a review of some empirical data on wet etching of single-crystal Si, discuss mask alignment with the correct crystallographic direction, review the various etchants and masking material approaches, and then launch into describing different etch stop techniques, which catapulted micromachining into an industrially accepted manufacturing method. Subsequently, we review models explaining wet isotropic and anisotropic etching and pick up the discussion anew of anodic polishing, photoetching, and formation of porous Si in HF solutions. Next comes a discussion of the technical issues encountered in bulk micromachining, such as IC incompatibility and extensive real-estate usage. We conclude with some example applications of wet bulk micromachining.

The first MEMS modeling software was developed around wet bulk micromachining processes and mechanical MEMS structures. Such programs are reviewed in Volume III, Chapter 6.

Empirical Data on Wet Etching of Single-Crystal Si

Isotropic and Anisotropic Etched Shapes in Single-Crystal Si

In Figure 4.20, we compare a wet isotropic etch (Figure 4.20a) with examples of anisotropic etches on masked flat pieces of single-crystal Si (Figure 4.20b–e). In the anisotropic etching examples, a square (Figure 4.20b and c) and a rectangular pattern (Figure 4.20d) are defined in an oxide mask with sides aligned along the family of <110> directions on a [100]-oriented silicon surface. For a refresher on the use of < >, [], and () (the Miller indices), see Volume I, Chapter 2 (Table 2.2). The square openings are precisely aligned (within 1° or 2°) with the [110] directions on the (100) wafer surface to obtain pits that conform exactly to the oxide mask rather than undercutting it. Most [100] silicon wafers have a main flat parallel to a [110] direction in the crystal, allowing for a rough alignment of the mask (see Volume I, Figure 4.3). Etching with the square pattern results in a pit with well-defined {111} sidewalls (at angles of 54.74° to the surface) and a (100) bottom. Volume I,

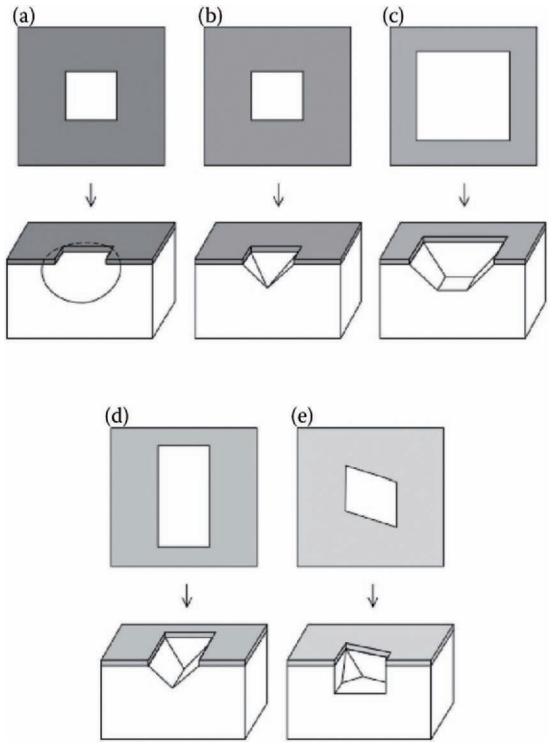


FIGURE 4.20 Isotropic and anisotropic etched features in [100] and [110] wafers. (a) Isotropic etch. (b–e) Anisotropic etch. (a–d) [100]-oriented wafers. (e) [110]-oriented wafer.

Equation 4.2, $W_0 = W_m - 2z\cot 54.74° = W_m - \sqrt{2}z$ allows one to calculate the dimensions of the hole at the bottom of the pit. For a larger square opening in the mask, the point where the {111} sidewalls of the pit intersect lays deeper inside the crystal. If the oxide opening is wide enough, that is, $W_0 > \sqrt{2}\,l$ (with l = 600 µm for a typical 6-in. wafer, this means $W_0 >$ 849 µm), the {111} planes do not intersect within the wafer. The etched pit in this particular case extends all the way through the wafer, creating a small square opening on the bottom surface (see Volume I, Figure 4.8). As shown in Figure 4.20b–d, no underetching of the etch mask is observed as a result of the perfect alignment of the concave oxide mask opening with the [110] direction. In Figure 4.20a, an undercutting isotropic etch (acidic) is shown. In the case of an anisotropic etch, misalignment results in pyramidal pits, but the mask will be severely undercut. A rectangular pattern aligned along the <110> directions on a <100> wafer leads to long V-shaped grooves (see Figures 4.20d and 4.21) or an open slit, depending on the width of the opening in the oxide mask.

Using a properly aligned mask (see Volume I, Chapter 4) on a [110] wafer, grooves with four

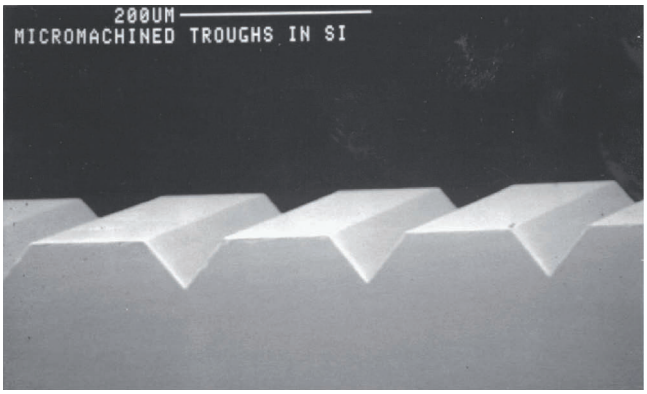


FIGURE 4.21 Long V-shaped grooves in a (100) Si wafer.

vertical and two slanted {111} planes result (see Figures 4.20e and 4.22a). Even a slight mask misorientation leads to all skewed sidewalls instead (Figure 4.22b). Before the emergence of slanted {111} bottom planes, the groove is defined by four vertical (111) planes and a horizontal (110) bottom (Figure 4.19, bottom right), and between there is a competition between {110} and {100} bottom planes. Self-stopping occurs when the tilted end {111} planes intersect at the bottom of the groove (Figure 4.22d). It is easy to etch very long, narrow grooves deeply into a [110] silicon wafer (Figure 4.23). However, it is impossible to etch a short, narrow groove deeply into it[48] because the narrow dimension of the groove

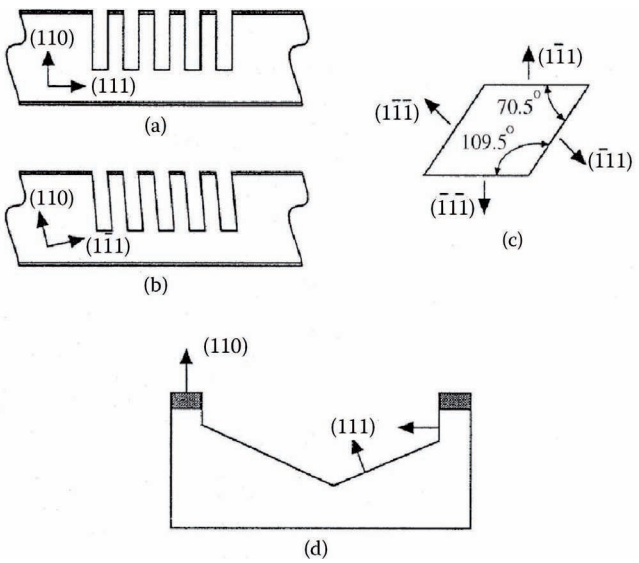


FIGURE 4.22 Anisotropic etching of [110] wafers. (a) Closely spaced grooves on correctly oriented (110) surface. (b) Closely spaced grooves on misoriented [110] wafer. (c) Orientations of the {111} planes looking down on a (110) wafer. (d) Shallow slanted (111) planes eventually form the bottom of the etched cavity.

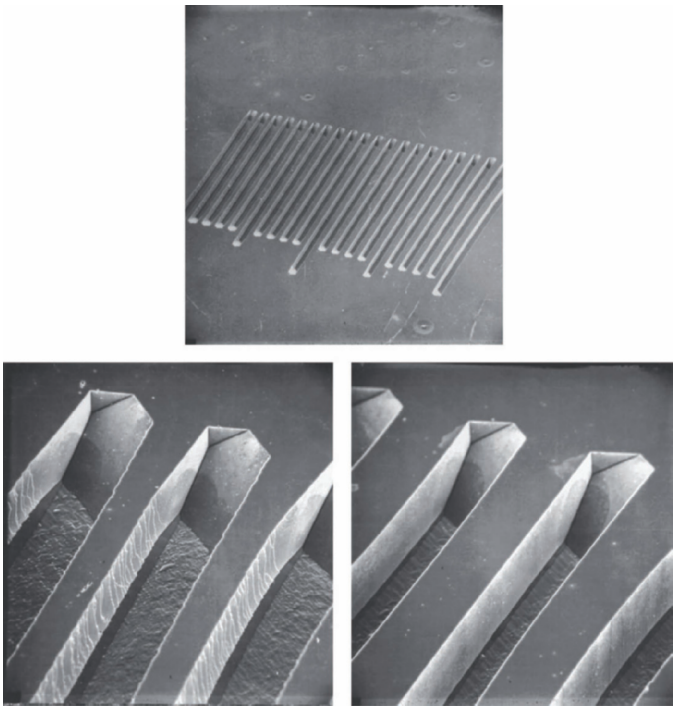


FIGURE 4.23 Test pattern of U-grooves in a <110> wafer to help in the alignment of the mask; final alignment is done with the groove that exhibits the most perfect long perpendicular walls.

is quickly limited by slow-etching {111} planes that subtend an angle of 35.26° to the surface and cause etch termination. At a groove of length L = 1 mm on the top surface, etching will stop when it reaches a depth of 0.289 mm, that is, $D_{max} = L/2\sqrt{3}$. For very long grooves, the tilted end planes are too far apart to intersect in practical cases, making the end effects negligible compared with the remaining trench-shaped part of the groove. Tuckerman and Pease[49] demonstrated the use of such trenches as liquid cooling fins for integrated circuits. The set of long, narrow grooves in a <110> wafer shown in Figure 4.23 can also be used as a test pattern to help in the alignment of the mask; final alignment is done with the groove that exhibits the most perfect long perpendicular walls (see also Figure 4.25).

The orientation of the wafer is of extreme importance, especially when machining surface structures by undercutting. Consider, for example, the formation of a bridge in Figure 4.24a.[50] When using a (100) surface, a suspension bridge cannot form across the etched V-groove; two independent truncated V-grooves flanking a mesa structure result instead.

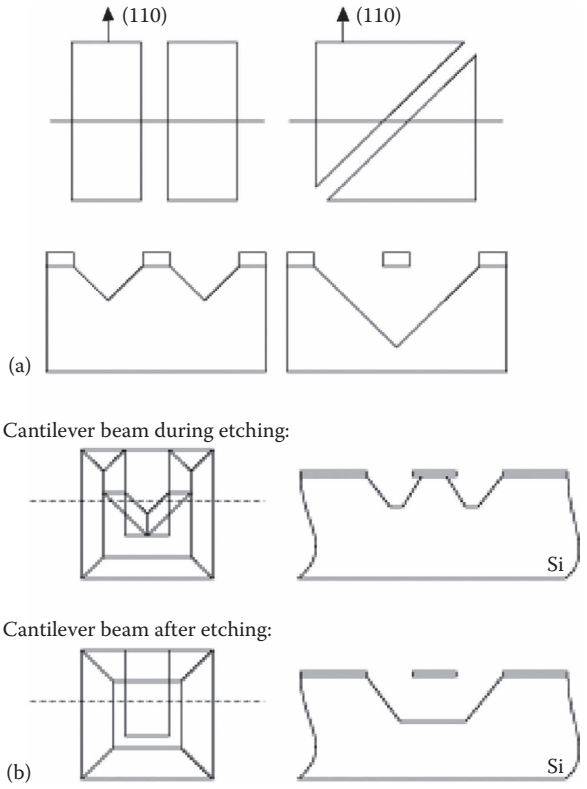


FIGURE 4.24 How to make a suspension bridge from a [100] Si wafer (a) and a diving board from a [110] Si wafer (b). (Barth, P. W., P. J. Shlichta, and J. B. Angel. 1985. *Third international conference on solid-state sensors and actuators.* Philadelphia.[50])

To form a suspended bridge, the V-groove must be oriented away from the [110] direction. Contrast this with a (110) wafer where a microbridge crossing a V-groove with a 90° angle will be undercut. Convex corners will be undercut by etchant, allowing formation of cantilevers as shown in Figure 4.24b. The diving board shown in Figure 4.20b forms by undercutting, starting at the convex corners. For more details, Table 4.4 is in the section on selection of [100]- or [110]-oriented silicon wafers, where we compare the main characteristics of etched features in [100]- and [110]-oriented wafers.

Alignment Patterns—Wagon-Wheel Masking Patterns

We just learned how a mask pattern may be aligned with the wafer flat to obtain a specific desired anisotropic etched feature. When alignment of a pattern along a certain crystallography direction is more critical, pre-etch alignment targets become useful to delineate the planes of interest because the wafer flats often are aligned to ±1° only. To find the proper alignment for the mask, a test pattern of closely spaced lines can be etched (see Figure 4.23). As discussed above, the groove with the best vertical walls determines the proper final mask orientation. Along this line, Ciarlo[51] made a set of lines 3 mm long and 8 μm wide, fanned out like spokes in a wagon wheel at angles 0.1° apart. This target was printed near the perimeter of the wafer and then etched 100 μm into the surface. Again, by evaluating the undercut in this target, the correct crystal direction could be determined. Alignment with better than 0.05° accuracy was accomplished this way.

The wagon-wheel pattern technique has also been used to obtain detailed experimental data on crystal orientation dependence of etch rates. For example, Seidel et al.[52] used a wagon-wheel or star-shaped mask (e.g., made from CVD-Si_3N_4, SiH_4, and NH_3 at 900°C), consisting of radially divergent segments with an angular separation of 1° (Figure 4.25a). By selectively masking the wafer surface material with material that is inert to the etchant (in this case, Si_3N_4) the removal of material is forced to occur vertically at the nonmasked areas and laterally under the masking spokes (Figure 4.25b). Beautiful patterns are formed as a result of the sidewalls being etched at different rates (Figure 4.25c). The etch rate for each sidewall in the wheel (distance r) is obtained from the lateral underetch rate (distance w) from $r = w\sin\theta$ (Figure 4.25b).

Seidel et al.[52] found an etch pattern emerging on a <100>-oriented wafer covered with wagon-wheel mask as shown in Figure 4.26a. The blossom-like figure is the result of the total underetching of the passivation layer in the vicinity of the center of the wagon wheel, leaving an area of bare, exposed Si. The radial extension of the bare Si area depends on the crystal orientation of the individual segments, leading to a different amount of total underetching. The observation of these blossom-like patterns is used for qualitative guidance of etching rates only. To establish quantitative numbers for the lateral etch rates, the width, w, of the overhanging passivation layer is measured with an optical line-width measurement system. Laser beam reflection is used to identify the crystal planes, and ellipsometry is used to monitor the etching rate of the mask itself.

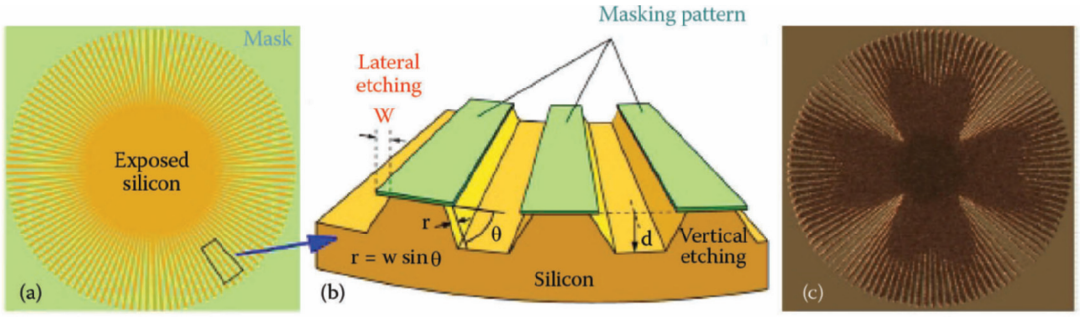


FIGURE 4.25 (a) Wagon-wheel masking pattern on silicon before etching. (b) Vertical etching (distance d) and lateral underetching (distance w). (c) Flower pattern generated during anisotropic etching. (Gonsalvez, M. 2003. *Atomistic modelling of anisotropic etching of crystalline silicon*. PhD diss., Helsinki University of Technology.[47])

Lateral etch rates determined in this way on <100>- and <110>-oriented wafers at 95°C in EDP (470 mL water, 1 L ethylene-diamine, 176 g pyrocatechol) and at 78°C in a 50% KOH solution are shown in Figure 4.27 (these two etchants are detailed further below). Etch rates shown are normal to the crystal surface and are conveniently described in a polar plot in which the distance from the origin to the polar plot surface (or curve in two dimensions) indicates the etch rate for that particular direction. Note the deep minima at the {111} planes. It can also be seen that, in KOH, the peak etch rates are more pronounced. A further difference is that with EDP the minimum at {111} planes is steeper than with KOH. For both EDP and KOH, the etch rate depends linearly on misalignment. All the above observations have important consequences for the interpretation of anisotropy of a given etch (see below). The difference between KOH and EDP etching behavior around the {111} minima has the direct practical consequence that it is more important for etching in EDP to align the crystallographic direction more precisely than in KOH.[53]

When determining etch rates without using underetching masks but using vertical etching of beveled silicon samples, results are different from working with masked silicon.[54] The etch rates on open areas of beveled structures are much larger than in underetching experiments with masked silicon, and different crystal planes develop. Herr et al.[54] conclude that crevice effects may play an important role in anisotropic etching. The model of Elwenspoek et al.,[55,56] analyzed below, is the only model that predicts such a crevice effect. The authors explain why etching rates slow down and increase anisotropy when etchants are in a small restricted crevice area and are not refreshed fast enough.

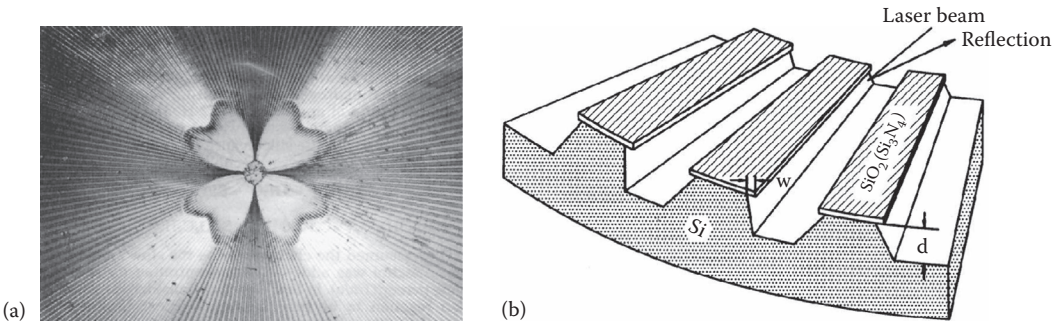


FIGURE 4.26 (a) Etch pattern emerging on a wagon wheel-masked, <100>-oriented Si wafer after etching in an EDP solution. (b) Schematic cross-section of a silicon test chip covered with a wagon wheel-shaped masking pattern after etching. The measurement of *w* is used to construct polar diagrams of lateral underetch rates as shown in Figure 4.27. (From Seidel, H., L. Csepregi, A. Heuberger, and H. Baumgartel. 1990. Anisotropic etching of crystalline silicon in alkaline solutions–part I: orientation dependence and behavior of passivation layers. *J Electrochem Soc* 137:3612–26. With permission.[52])

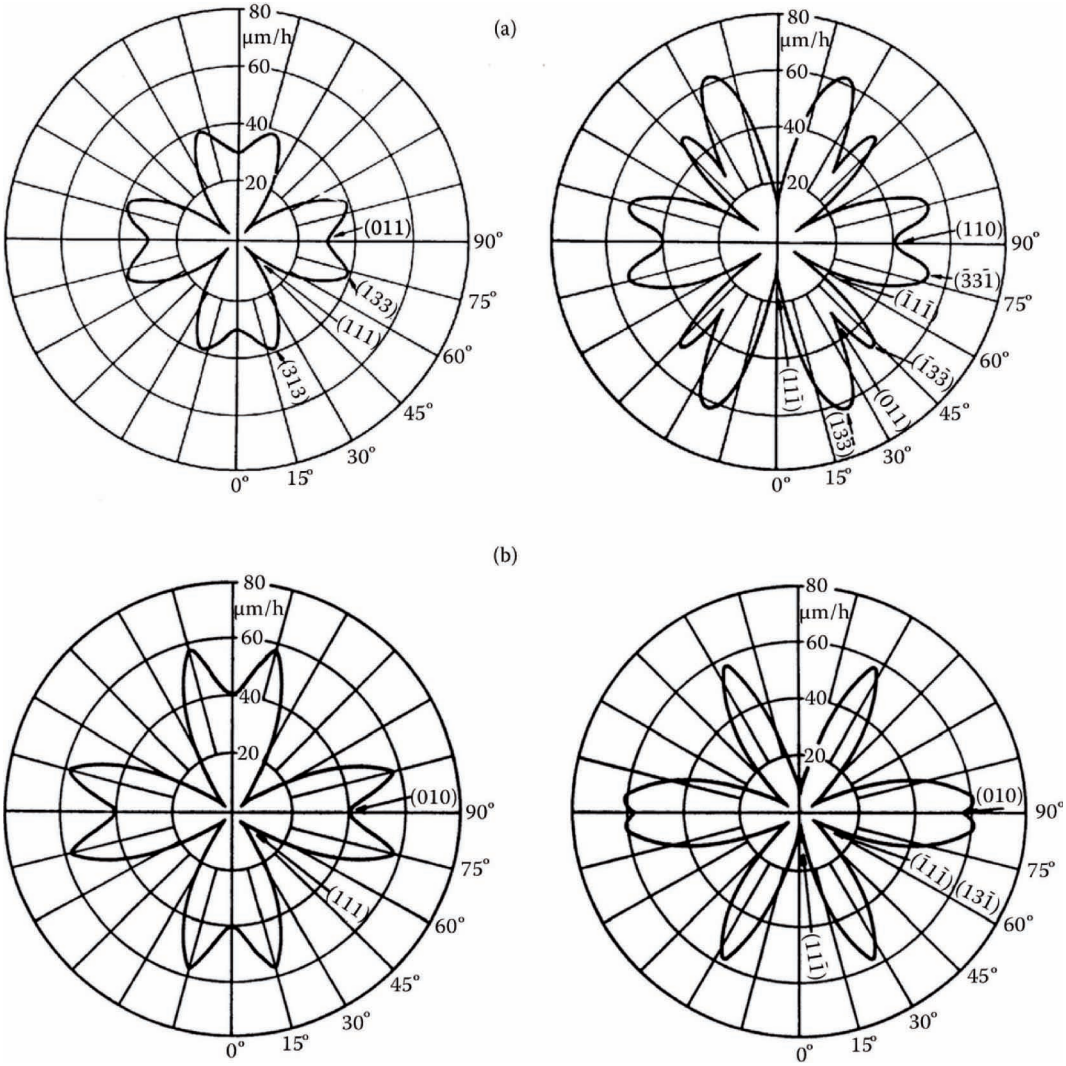


FIGURE 4.27 Lateral underetch rates as a function of orientation for (a) EDP (470 ml water, 1 l ED, 176 g pyrocatechol) at 95°C and (b) KOH (50% solution) at 78°C. Left, <100>- and right, <110>-oriented Si wafers. (From Seidel, H., L. Csepregi, A. Heuberger, and H. Baumgartel. 1990. Anisotropic etching of crystalline silicon in alkaline solutions–part I: orientation dependence and behavior of passivation layers. *J Electrochem Soc* 137:3612–26. With permission.[52])

Selection of [100]- or [110]-Oriented Silicon Wafers

In Table 4.4, we compare the main characteristics of etched features in [100]- and [110]-oriented wafers. This table may help in the decision which orientation to use for a specific microfabrication application at hand.

From this table, it is obvious that for membrane-based sensors, [100] wafers are preferred, but an understanding of the geometric considerations with [110] wafers is important as well to fully appreciate all the possible single-crystal silicon micromachined shapes and to better understand corner compensation schemes (see below). Moreover, processes for providing dielectric isolation require that the silicon be separated into discrete regions. To achieve a high component density with anisotropic etches on [100] wafers, the silicon must be made very thin because of the aspect ratio limitations as a result of the sloping walls (see above). With vertical sidewall etching in a [100] wafer, the etch mask is undercut in all directions to a distance approximately equal to the depth of the etching. Vertical etching in [110] surfaces relaxes the etching requirement dramatically and enables more densely packed structures such as beam leads or image sensors. Kendall[48,57] describes and predicts a wide variety of applications for [110] wafers, such as fabrication of trench

Table 4.4 Matrix for Optimal Selection of Wafer Type

[100] Orientation	[110] Orientation
Inward sloping walls (54.74°)	Vertical {111} walls
The sloping walls cause a lot of real estate loss	Narrow trenches with high-aspect ratio are possible
Flat bottom parallel to surface is ideal for membrane fabrication	Multifaceted cavity bottom ({110} and {100} planes) makes for a poor diaphragm
Bridges perpendicular to a V-groove bound by (111) planes cannot be underetched	Bridges perpendicular to a V-groove bound by (111) planes can be undercut
Shape and orientation of diaphragms convenient and simple to design	Shape and orientation of diaphragms awkward and more difficult to design
Diaphragm size, bounded by nonetching {111} planes, is relatively easy to control	Diaphragm size is difficult to control (the <100> edges are not defined by nonetching planes)

capacitors, vertical multijunction solar cells, diffraction gratings, infrared interference filters, large area cathodes, and filters for bacteria.

Laser Spoiling

A laser can be used to melt or "spoil" the shallow {111} surfaces, making it possible to etch deep vertical-walled holes through a [110]-oriented wafer as shown in Figure 4.28a.[50,58,59] The absorbed energy of a neodymium-doped yttrium aluminium garnet (Nd:YAG) laser beam causes a local melting or evaporation zone, enabling etchants to etch the shallow {111} planes in the line of sight of the laser. Etching proceeds until "unspoiled" (111) planes are encountered. Some interesting resulting possibilities, including partially closed microchannels, are shown in Figure 4.28c.[58,60] Note that with this method it is possible to use [111] wafers for micromachining. The light of the Nd:YAG laser is very well suited for this micromachining technique as a result of the 1.17 eV photon energy, just exceeding the bandgap energy of Si. Details on this laser machining process can be found, for example, in Alavi et al.[60,61]

Wet Isotropic and Anisotropic Etching

Introduction

Besides for the shaping of 3D microstructures, wet etching is used for cleaning, removing surface damage, polishing, and characterizing structural and compositional features.[7] The most important parameters in wet chemical etching are bias (undercut), tolerance, etch rate, anisotropy, selectivity, overetch, feature size control, and loading effects. These controlling parameters are very similar to the parameters controlling dry chemical etching (see Chapter 3). The emphasis in this chapter is on wet etching of single-crystal Si. Wet chemical etching of Si provides a higher degree of selectivity than dry etching techniques. Wet etching is usually also faster: a few microns to tens of microns per minute for isotropic

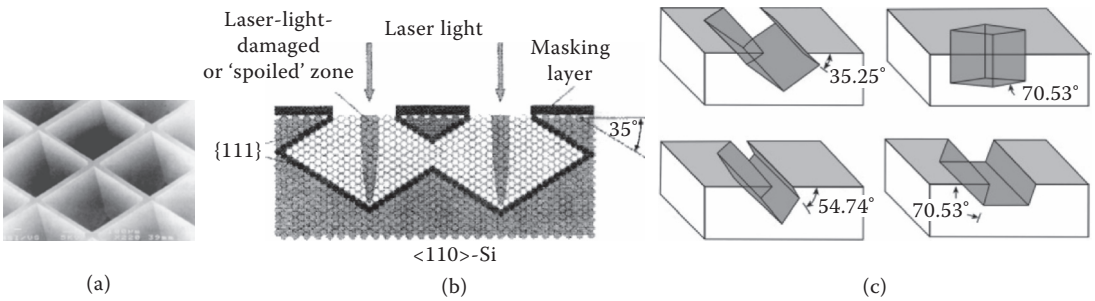
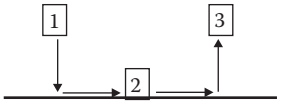


FIGURE 4.28 Laser/KOH machining. (a) Holes through a (110) Si wafer created by laser spoiling and subsequent KOH etching. The two sets of (111) planes making an angle of 70° are the vertical walls of the hole. The (111) planes making a 35.26° angle with the surface tend to limit the depth of the hole, but laser spoiling enables one to etch all the way through the wafer. (b) After laser spoiling, the line-of-sight (111) planes are spoiled, and etching proceeds until unspoiled (111) planes are reached. (c) Some of the possible features rendered by laser spoiling. (From Schumacher, A., H.-J. Wagner, and M. Alavi. 1994. Mit Laser und Kalilauge. *Technische Rundschau* 86:20–23. With permission.[58])



1 Reactant transport to surface

2 Selective and controlled reaction of etchant with the film to be etched

3 Transport of by-products away from surface

FIGURE 4.29 Wet etching processes.

etchants and about 1 μm/min for anisotropic wet etchants versus 0.1 μm/min in typical dry etching. More recently, however, rates of up to 6 μm/min were achieved using high-density plasmas for dry etching of Si (see Chapter 3).

Modification of the nature of the wet etchant and/or operating temperature can also alter the selectivity to the silicon dopant concentration and type and, especially when using alkaline etchants, to crystallographic orientation. Etching proceeds by reactant transport to the surface (1), surface reaction (2), and reaction product transport away from the surface (3) (see Figure 4.29). If (1) or (3) is rate determining, etching is diffusion limited, and etching is controlled by the hydrodynamics of the system; thus, the process may be enhanced by stirring. Mass transport-determined etching, being isotropic, gives rise to rounded features. This also means that the surface reactions (2) are fast and that the etch rate is not sensitive to the nature of the surface, e.g., its crystallographic orientation or morphology. Diffusion-limited processes have lower activation energies (in the order of a few kilocalories per mole) than reaction rate-controlled processes and therefore are relatively insensitive to temperature variations. If a surface reaction (2) is the slowest step and thus constitutes the rate-determining step, etching is reaction rate limited and depends strongly on the temperature, the crystallographic orientation, and/or morphology of the etching material, as well as solution composition. In other words, with surface kinetics important the etch rate is very likely to be sensitive to surface orientation. In general, reaction rate limitation is preferred because it is easier to reproduce a temperature setting than a stirring rate. The best generic etching apparatus has both a good temperature controller and a reliable stirring facility.[62,63]

Preferential or selective etching (also *structural etching*) is a type of etching that is usually isotropic but exhibits some anisotropy.[64] These etchants are used to produce a difference in etch rate between different materials or between compositional or structural variations of the same material on the same crystal plane. These types of etches are often the fastest and simplest techniques to delineate electrical junctions and to evaluate the structural perfection of a single crystal in terms of, for example, slip and stacking faults. The artifacts introduced by the defects etch into small pits of characteristic shape. Most of the etchants used for this purpose are acids with an oxidizing additive.[65–69]

Isotropic Etching

Usage of Isotropic Etchants When etching silicon with aggressive acidic etchants, rounded isotropic patterns form. The method is widely used for:

1. Removal of work-damaged surfaces
2. Rounding of sharp anisotropically etched corners (to avoid stress concentration)
3. Removing of roughness after dry or anisotropic etching
4. Creating structures or planar surfaces in single-crystal slices (thinning)
5. Patterning single-crystal, polycrystalline, or amorphous films
6. Delineation of electrical junctions and defect evaluation (with preferential isotropic etchants)

For isotropic etching of silicon, the most commonly used etchants are mixtures of nitric acid (HNO_3) and hydrofluoric acids* (HF). In the case of HNO_3, water can be used as a diluent, but acetic acid (CH_3COOH) is preferred because it prevents the dissociation of the nitric acid better and thus preserves the oxidizing power of HNO_3, which depends on the undissociated nitric acid species for a wide range of dilution.[4] The etchant is called the HNA system; we will return to this etch system below.

* Be careful when working with HF; it is deceptive because it looks like water, and even when not a strong acid, it penetrates the skin (adsorption) and slowly attacks bones.

Simplified Reaction Scheme In acidic media, the Si etching process involves hole injection into the Si valence band by an oxidant, an electrical field, or photons. Nitric acid in the HNA system acts as an oxidant; other oxidants such as H_2O_2 and Br_2 also work.[70] The holes attack the covalently bonded Si, oxidizing the material, followed by a reaction of the oxidized Si fragments with OH^- and subsequent dissolution of the silicon oxidation products in HF. It was Kooij et al.[71] who, from electrochemical measurements on the Si/HF/HNO_3 system, concluded that the holes, in the absence of photons and an applied field, are introduced by HNO_3:

$$HNO_3 + H_2O + HNO_2 \rightarrow 2HNO_2 + 2OH^- + 2h^+ \qquad \text{Reaction 4.1}$$

The holes (h^+) in Reaction 4.1 are generated in an autocatalytic process; HNO_2 generated in the above reaction re-enters into the further reaction with HNO_3 to produce more holes. With a reaction of this type, there is an induction period before the oxidation reaction takes off until a steady-state concentration of HNO_2 has been reached. This induction period has been observed at low HNO_3 concentrations.[70] After hole injection, OH^- groups attach to the oxidized Si species to form SiO_2, liberating hydrogen in the process:

$$Si^{4+} + 4OH^- \rightarrow SiO_2 + H_2 \qquad \text{Reaction 4.2}$$

HF dissolves the SiO_2 by forming the water-soluble H_2SiF_6. The overall reaction of HNA with Si is then:

$$Si + HNO_3 + 6HF \rightarrow H_2SiF_6 \text{ (water-soluble)} + HNO_2 + H_2O + H_2 \text{ (bubbles)} \qquad \text{Reaction 4.3}$$

The simplification in the above electrochemical component of the reaction scheme is that only holes are taken into account. In the actual Si acidic corrosion reaction, both holes and electrons are involved. The question of hole and/or electron participation in Si corrosion will be considered after the introduction below of the model for the Si/electrolyte interfacial energetics. We will learn from that model that the rate-determining step in acidic etching involves hole injection in the valence band, whereas in alkaline anisotropic etching, it involves electron injection in the conduction band by surface states. The reactivity of a hole injected in the valence band is significantly greater than that of an electron injected in the conduction band. The observation of isotropy in acidic etchants and anisotropy in alkaline etchants centers on this difference in reactivity.

Isotropic etching can be performed chemically, in very aggressive solutions containing HF and an oxidizing agent such as HNO_3, or electrochemically, in much less aggressive fluoride-containing solutions. Both methods offer the possibility of forming porous Si. Porous Si formation is easier to control using the electrochemical method. For electrochemical etching, an electrical contact must be made to the back side of the sample, which limits the possibilities of the process (see below).

Iso-Etch Curves By the early 1960s, the isotropic HNA silicon etch was well characterized. Schwartz and Robbins published a series of four very detailed papers on the topic between 1959 and 1976.[3–6] Most of the material presented below is based on their work.

HNA etching results, represented in the form of iso-etch curves, for various weight percentages of the constituents are shown in Figure 4.30. For this work,

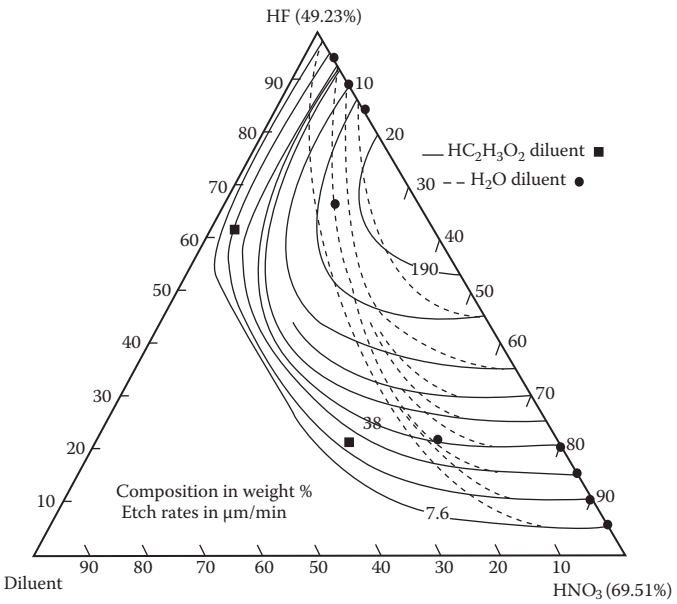


FIGURE 4.30 Iso-etch curves. (Robbins, H., and B. Schwartz. 1960. Chemical etching of silicon-II: the system HF, HNO_3, H_2O, and $HC_2C_3O_2$. *J Electrochem Soc* 107:108–11,[4] recalculated for one-sided Si etching and expressed in microns per minute.)

normally available concentrated acids of 49.2 wt% HF and 69.5 wt% HNO_3 are used. Water as diluent is indicated by dashed-line curves, whereas acetic acid is noted by solid-line curves. Acetic acid is less polar than water and helps prevent the dissociation of HNO_3, thereby allowing the formation of more of the species directly responsible for the oxidation of Si. A typical formulation for HNA is 250 mL of HF, 500 mL of HNO_3, and 800 mL of CH_3COOH. When used at room temperature, this formulation results in an etch rate of about 4–20 μm/min, increasing with agitation.[72] In Figure 4.30 we have recalculated the curves from Robbins and Schwartz[4] to express the etch rate in microns per minute and divided the authors' numbers by 2 as we are considering one-sided etching only. The highest etch rate is observed around a weight ratio HF-HNO_3 of 2:1 and is nearly 100 times faster than anisotropic etch rates. Adding a diluent slows down the etching. From these curves, the following characteristics of the HNA system can be summarized:

1. At high HF and low HNO_3 concentrations, the iso-etch curves describe lines of constant HNO_3 concentrations (parallel to the HF-diluent axis); consequently, the HNO_3 concentration controls the etch rate. Etching at those concentrations tends to be difficult to initiate and exhibits an uncertain induction period (see above). In addition, it results in relatively unstable silicon surfaces proceeding to slowly grow a layer of SiO_2 over time. The etch is limited by the rate of oxidation; therefore, it tends to be orientation-dependent and affected by dopant concentration, defects, and catalysts (sodium nitrate is often used). In this regime, the temperature influence is more pronounced, and activation energies for the etching reaction of 10–20 kcal/mol have been measured.
2. At low HF and high HNO_3 concentrations, iso-etch curves are lines parallel to the nitric-diluent axis; that is, they are at constant HF composition. In this case, the etch rate is controlled by the ability of HF to remove the SiO_2 as it is formed. Etches in this regime are isotropic and truly polishing, producing a bright surface with anisotropies of 1% or less (favoring the <110> direction) when used on <100> wafers.[73] An activation energy of 4 kcal/mol is indicative of the diffusion-limited character of the process; consequently, in this regime, temperature changes are less important.
3. In the region of maximal etch rate, both reagents play an important role. The addition of acetic acid, as opposed to the addition of water, does not reduce the oxidizing power of the nitric acid until a fairly large amount of diluent has been added. Therefore, the rate contours remain parallel with lines of constant nitric acid over a considerable range of added diluent.
4. In the region around the HF vertex, the surface reaction rate-controlled etch leads to rough, pitted Si surfaces and sharply peaked corners and edges. In moving toward the HNO_3 vertex, the diffusion-controlled reaction results in the development of rounded corners and edges, and the rate of attack on (111) planes and (110) planes becomes identical in the polishing regime (anisotropy less than 1%; see point 2).

In Figure 4.31, we summarize how the topology of the Si surfaces depends strongly on the composition of the etch solution. Around the maximum etch rates the surfaces appear flat with rounded edges, and very slow etching solutions lead to rough surfaces.[6]

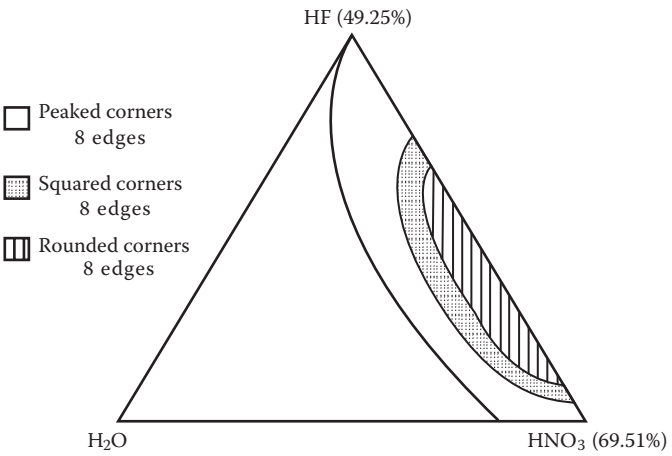


FIGURE 4.31 Topology of etched Si surfaces. (From Schwartz, B., and H. Robbins. 1976. Chemical etching of silicon-IV: etching technology. *J Electrochem Soc* 123:1903–9. With permission.[6])

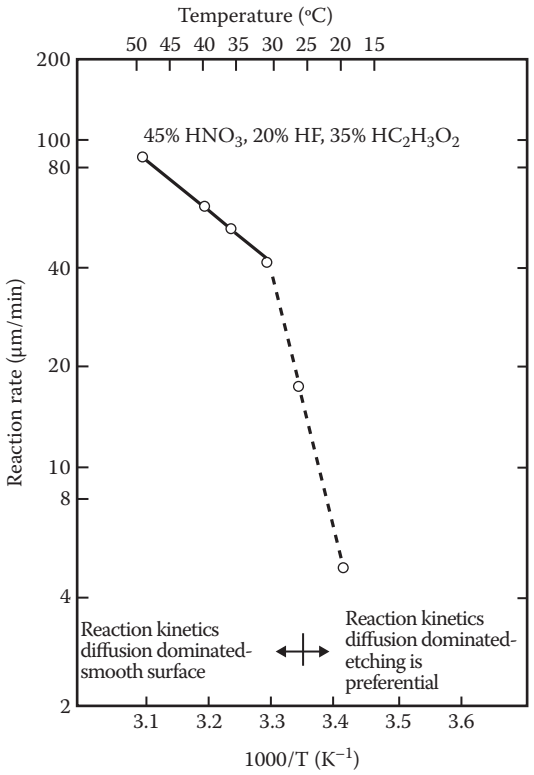


FIGURE 4.32 Etching an Arrhenius plot. Temperature dependence of the etch rate of Si in HF:HNO_3:CH_3:COOH (1:4:3). (From Schwartz, B., and H. Robbins. 1961. Chemical etching of silicon-III: a temperature study in the acid system. *J Electrochem Soc* 108:365–72. With permission.[5])

Arrhenius Plot for Isotropic Etching Schwartz and Robbins[5] studied the effect of temperature on the reaction rate in the HNA system in detail. An Arrhenius plot for etching silicon in 45% HNO_3, 20% HF, and 35% $HC_2H_3O_2$, culled from their work, is shown in Figure 4.32. Increasing the temperature increases the reaction rate. The graph shows two straight-line segments, indicating higher activation energy at less than 30°C and a lower one above this temperature. In the low temperature range, etching is preferential, and the activation energy is associated with the oxidation reaction. At higher temperatures, the etching leads to smooth surfaces, and the activation energy is lower and associated with diffusion-limited dissolution of the oxide.[5]

Loading Effects in Isotropic Etching Two HNA solutions that etch silicon isotropically at a reasonable rate (around 2 μm/min) and leave smooth walls [roughness in the order of 5 nm, as determined from scanning electron microscope (SEM) pictures] are shown in Table 4.5.[74] The etch rates given in this table are averages because the etch rate depends on the mask opening size and structure density on the wafer, resulting from the diffusion-controlled nature of the etching process.

TABLE 4.5 Compositions of Two Typical Isotropic HNA Solutions

	Vol % HF (50%)	Vol % HNO_3 (69%)	Vol % Water	Etch Rate (μm/min)
Sol 1	11	22	67	3
Sol 2	5	15	80	2

Source: Tjerkstra, R. W. 1999. *Isotropic etching of silicon in fluoride containing solutions as a tool for micromachining.* PhD diss., University of Twente, Twente, the Netherlands.[74]

With isotropic etchants, the etchant moves downward and outward from an opening in the mask, undercuts the mask, and enlarges the etched pit while deepening it (Figure 4.33). The resulting isotropically etched features show more symmetry and rounding when agitation accompanies the etching (the process is diffusion limited). This agitation effect is illustrated in Figure 4.33. With agitation, the etched feature approaches a more ideal round cup; without agitation, the etched feature resembles a rounded box.[75] However, the flatness of the bottom of the rounded box generally is poor because the flatness is defined by agitation. Therefore, sometimes solutions are on purpose not stirred during etching because stirring enhances the etch rate in the direction of fluid flow, yielding badly shaped channels.

As we saw above, the etching of silicon in HF/HNO_3 solutions involves two processes: the oxidation of the silicon surface and the subsequent removal of the oxidized silicon by the fluoride in the solution. In the case of isotropic etching, the rate of the latter process is determined by the transport of fluoride ions to the surface (diffusion limited). This can lead to loading effects as illustrated during the isotropic etch of a spiral in Figure 4.34a.

We observe that the channels on the rim of the spiral have a larger diameter than the channels in the middle of the spiral. This is because on the rim of the spiral there is no silicon that etches and consumes fluoride ions so the ions above the mask material next to the outermost channels can diffuse to these channels. This leads to nonuniformities in

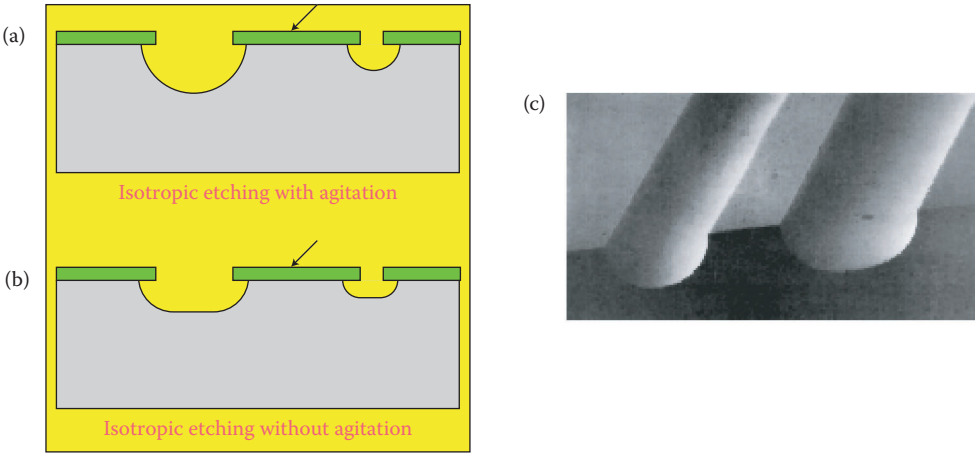


FIGURE 4.33 Isotropic etching of silicon with (a) and without (b) etchant solution agitation; (c) Examples of isotropic etching in Si. Left groove with agitation, right groove no agitation.

the widths of the channels on the rim of the spiral. Stirring the solution during etching does not necessarily help because, as we just saw, that enhances the etching in the direction of fluid flows, yielding badly shaped channels. In an attempt to solve the loading effect problem, the spiral may be surrounded with sacrificial channels with the same width and the same distance from each other as the channels in the spiral ("compensation structures") as illustrated in Figure 4.34b. The etching of these extra structures takes away the extra fluoride that would otherwise cause the broadening of the outer channels. From work by Kuiken et al.,[76] Tjerkstra[74] concluded that for short etch times (etch times in the order of seconds) the bulk concentration is reached at a distance of around $3(Dt)^{1/2}$, in which D is the diffusion coefficient and t the time.

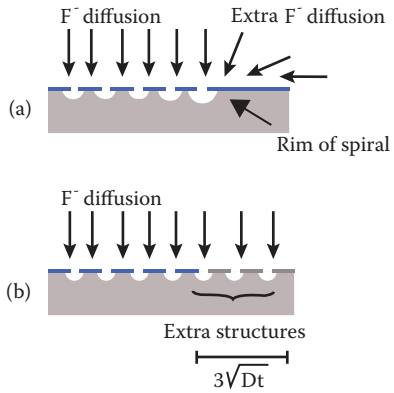


FIGURE 4.34 Side view of anisotropically etched spiral, illustrating the loading effect. (a) The fluoride ions in the area next to the etching spiral can diffuse to the outer channels. These channels etch faster than the other channels because of the extra supply of fluoride. (b) To reduce this effect, extra structures are placed around the spiral that reduce the extra supply of fluoride. This results in a uniform channel width (Tjerkstra, R. W. 1999. *Isotropic etching of silicon in fluoride containing solutions as a tool for micromachining.* PhD diss., University of Twente, Twente, the Netherlands.[74])

The enhanced etching of structures at the rim of the spiral in Figure 4.34 may also be a temperature effect. Etching is the result of an exothermic reaction; thus, the temperature of the solution increases during etching. This causes the appearance of fluid flow, as shown in Figure 4.35. The channels at the rim of the spiral may get an extra supply of fresh solution and thus etch faster.

In Chapter 3 we encountered loading effects also when we discussed dry etching, where they are more prevalent because the etchant concentration is much lower in the gas phase.

Masking for Isotropic Silicon Etchants Acidic etchants are very fast; for example, an etch rate for Si of

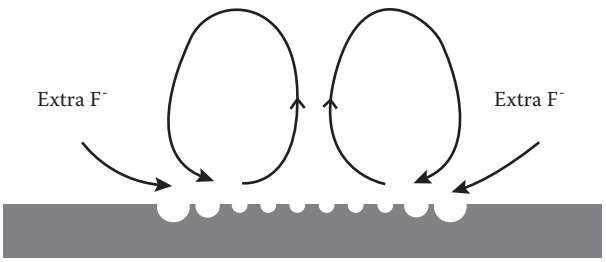


FIGURE 4.35 Enhanced etching caused by fluid flow resulting from the increased temperature of the etching solution. (Tjerkstra, R. W. 1999. *Isotropic etching of silicon in fluoride containing solutions as a tool for micromachining.* PhD diss., University of Twente, Twente, the Netherlands.[74])

Table 4.6 Masking Materials for Acidic Etchants

	Etchants		
Masking	Piranha (4:1, H_2O_2: H_2SO_4)	Buffered HF (5:1NH_4F: conc.HF)	HNA
Thermal SiO_2		0.1 μm/min	300–800 Å/min. Limited etch time, thick layers often are used because of ease of patterning
CVD (450°C) SiO_2		0.48 μm/min	0.44 μm/min
Corning 7740 glass		0.063 μ/min	1.9 μ/min
Photoresist	Attacks most organic films	OK for a short while	Resists do not stand up to strong oxidizing agents like HNO_3 and are not used
Undoped Si polysilicon	Forms 30 Å of SiO_2	0.23–0.45 Å/min	0.7–40 μm/min at RT [at a dopant concentration $<10^{17}$ cm^{-3} (*n* or *p*)]
Black wax			Usable at room temperature
Au/Cr	OK	OK	OK
LPCVD Si_3N_4		1 Å/min	Etch rate is 10–100 Å/min; preferred masking material

The many variables involved necessarily means that the given numbers are approximate only.

up to 50 μmin^{-1} can be obtained with 66% HNO_3 and 34% HF (volumes of reagents in the normal concentrated form).[64,77] Isotropic etchants are so aggressive that the activation barriers associated with etching the different Si planes are not differentiated; all planes etch equally fast, making masking a real challenge.

Although SiO_2 has an appreciable etch rate of 300–800 Å/min in the HF:HNO_3 system, thick layers of SiO_2 are often used as a mask, especially for shallow etching, as the oxide is so easy to form and pattern. A mask of nonetching Au or Si_3N_4 is needed for deeper etching. Photoresists do not stand up to strong oxidizing agents such as HNO_3, and neither does Al.

Silicon itself is soluble to a small extent in pure HF solutions; for a 48% HF, at 25°C a rate of 0.3 Å/min was observed for n-type, 2-Ω-cm (111)-Si. It was established that Si dissolution in HF is not the result of oxidation by dissolved oxygen. Diluted HF etches silicon at a higher rate because the reaction in aqueous solutions proceeds by oxidation of Si by OH^- groups.[78] A typically buffered HF (BHF) solution, also called buffered oxide etch, has been reported to etch Si at radiochemically measured rates of 0.23–0.45 Å/min, depending on doping type and dopant concentration.[79] In BHF, HF is buffered with NH_4F to control the pH value (–log[H^+]) at constant level and to replenish the depletion of fluoride ions to maintain a stable etch performance. BHF is the preferred SiO_2 etch:

$$SiO_2 + 4HF + 2NH_4F \rightarrow (NH_4)_2\,SiF_6 + 2H_2O$$

$$SiO_2 + 3HF_2^- + H^+ \rightarrow SiF_6^{2+} + 2H_2O \quad \text{Reaction 4.4}$$

By reducing the dopant concentration (*n* or *p*) to less than 10^{17} atoms/cm^3, the etch rate of Si in HNA is reduced by ≈150.[80] The doping dependence of the etch rate provides yet another means of patterning a Si surface (see next section). A summary of masks that can be used in acidic etching is presented in Table 4.6.

A simple way to detect the end-point for BHF* etching of a layer of SiO_2 on Si is to monitor the wetting properties of the surface. Whereas SiO_2 wets, Si is hydrophobic.

Dopant Dependence of Silicon Isotropic Etchants The isotropic etching process is fundamentally a charge-transfer mechanism. With isotropic etching Si at open circuit, we will learn further below, the chemical etch may have both chemical and electrochemical components. In the latter case, mobile charge carriers are injected or extracted from the solid, making the reaction sensitive to an applied potential. This also explains the etch rate dependence on dopant type and dopant concentration. Typical etch rates with an HNA system (1:3:8) for n- or p-type dopant concentrations more than 10^{18} cm^3 are 1–3 μm/min. As presented in the preceding section, a reduction of the etch rate by 150 times

* HF is inert in contact with skin but attacks bone.

is obtained in n- or p-type regions with a dopant concentration of 10^{17} cm^{-3} or smaller.[81] This presumably is because of the lower mobile carrier concentration available to contribute to the charge transfer mechanisms. In any event, heavily doped silicon substrates with high conductivity can be etched more readily than lightly doped materials. Dopant-dependent isotropic etching can also be exploited in an electrochemical setup as described in the next section. Although doping changes the chemical etch rate, attempts to exploit these differences for industrial production have failed.[81] This situation is different in the case of electrochemical isotropic etching (see next section).

Preferential Etching A variety of additives to the HNA system, mainly oxidants, can be included to modify the etch rate, surface finish, or isotropy, rendering the etching baths preferential. It is clear that the effect of these additives will only show up in the reaction-controlled regime. Only additives that change the viscosity of the solution can modify the etch rate in the diffusion-limited regime, thereby changing the diffusion coefficient of the reactants.[70,82] We will not review the effect of these additives any further; refer to Table 4.7 and the cited literature for more information.[65–69]

Problems Associated with Isotropic Etchants There are several problems associated with isotropic etching of Si. First, it is difficult to mask with high precision using a desirable and simple mask such as SiO_2 (etch rate is 2–3% of the silicon etch rate). Second, the etch rate is very agitation sensitive in addition to being temperature sensitive. This makes it difficult to control lateral and vertical geometries. Electrochemical isotropic etching (see below) and the development of anisotropic etchants in the late 1960s (see below) were able to overcome many of these problems.

A comprehensive review of isotropic etchants solutions can be found in Kern et al.,[64] including a review of different techniques practiced in chemical etching such as immersion etching, spray etching, electrolytic etching, gas-phase etching, and molten salt etching (fusion techniques). Table 4.7 lists some isotropic and preferential etchants and their specific applications.

Electrochemical Isotropic Silicon Etch—First Example of an Etch Stop For the electrochemical etching of Si, holes are needed at the substrate surface. In p-type Si, these holes are automatically present in the material. In n-type Si or with intrinsic Si, holes can be generated by, for example, illuminating the substrate with photons having energies larger than the bandgap (1.1 eV). The thus-created holes can be forced to the surface by applying a potential to an ohmic contact to the Si. In this case, a high-temperature or extremely aggressive chemical etching process is replaced by an electrochemical procedure using a much milder solution, thus allowing even a simple photoresist mask to be used.[64] In electrochemical acidic etching, with or without illumination of the corroding Si electrode, an electrical power supply is used to drive the chemical reaction by supplying holes to the silicon surface [working electrode (W-EL), see Figure 4.36a]. A voltage is applied across the silicon wafer, and a counter electrode (C-EL, usually platinum) is arranged in the same etching solution. Oxidation is promoted by a positive bias applied to the silicon, causing an accumulation of holes in the silicon at the silicon/electrolyte interface. Under this condition, oxidation at the surface proceeds rapidly while the oxide is readily dissolved by the HF solution. No oxidant such as HNO_3 is needed to supply the holes; excess electron-hole pairs are created by the electrical field at the surface and/or by optical excitation, thereby increasing the etch rate. This technique proved successful in removing heavily doped layers, leaving behind the more lightly doped membranes in all possible dopant configurations: p on p^+, p on n^+, n on p^+, and n on n^+.[87,88] This electrochemical etch-stop technique is demonstrated in Figure 4.36b.[89] A 5% HF solution is used; the electrolyte cell is kept in the dark at room temperature; and the distance between the Si anode and the Pt cathode in the electrochemical cell is 1–5 cm. Instead of using HF, one can substitute NH_4F (5 wt%) for the electrochemical etching as described by Shengliang.[90] Shengliang reports a selectivity of n-type silicon to n^+-type silicon (0.001 Ω cm) of 300 with the latter etchant. In Figure 4.36b, the current density versus applied voltage across the anode and cathode during

Table 4.7 Isotropic and Preferential Defect Etchants and Their Specific Applications

Etchant	Application/Material	Remark/Reference
HF 8 vol%, HNO_3 75 vol%, and CH_3COOH 17 vol%	n- and p-type Si, all planes, general etching	Planar etch; e.g., 5 μm/min at 25°C
1 part 49% HF, 1 part (1.5M-CrO_3) (by volume)	Delineation of defects on (111), (100), and (110) Si without agitation	Yang etch[65]
5 vol parts nitric acid (65%), 3 vol parts HF (48%), 3 vol parts acetic acid (96%), 0.06 parts bromine	Polishing etchant used to remove damage introduced during lapping	So-called CP4 etchant; Heidenreich, US Patent 2619414
HF 48%	SiO_2	Etch rate is 20–2000 nm/min; etch rate for Si is 0.3 Å/min for n-type 2 Ω cm (111); etch rate for Al is 5 nm/min
HF:NH_4F (buffered HF 28 ml HF, 170 ml H_2O, 2113 g NH_4F)	SiO_2	Etch rate is 100–500 nm/min at 25°C
1HF, 3HNO_3, 10 CH_3COOH	Delineates defects in (111) silicon; etches p^+ or n^+ and stops at p^- or n^-	Dash etch[83]; p- and n-type Si at 1300 Å/min in the [100] direction and 46 Å/min in the [111] direction at 25°C
1HF, 1(5M-CrO_3)	Delineates defects in (111); needs agitation; does not reveal etch pits on (100) well	Sirtl etch[84]
HF:H_2O_2	Titanium	880 nm/min
2HF,1(0.15M-$K_2Cr_2O_7$)	Yields circular (100) Si dislocation etch pits; agitation reduces etch time	Secco etch[69]
60 ml HF, 30 ml HNO_3, 30 ml (5M-CrO_3), 2 g $Cu(NO_3)_2$, 60 ml CH_3COOH, 60 ml H_2O	Delineates defects in (100) and (111) Si; requires agitation	Jenkins etch[85]
H_2O_2	Tungsten	20–100 nm/min
34 g KH_2PO_4, 13.4 g KOH, 33 g $K_3FE(CN)_6$ and H_2O to make up 1 L	Tungsten	160 nm/min
1 ml HCl, 9 ml saturated $CeSO_4$ solution	Chromium	80 nm/min
1 ml HCl, 1 ml glycerine	Chromium	80 nm/min
2HF, 1(1M-CrO_3)	Delineates defects in (100) Si without agitation; works well on resistivities 0.6–15.0 Ω cm n- and p-types)	Schimmel etch[86]
2HF, 1(1M-CrO_3), 1.5 H_2O	Works well on heavily doped (100) Si	Modified Schimmel
HF/$KMnO_4$/CH_3COOH	Epitaxial Si	
3 ml HCl, 1 ml HNO_3	Gold	25–50 μm/min Aqua regia
4 g KI, 1 g I_2, and 40 ml H_2O	Gold	0.5–1 μm/min
H_3PO_4 Etchants	Si_3N_4	Etch rate is 5–10 nm/min 160–180°C
KOH + alcohols	Polysilicon	85°C
H_2SO_4/H_2O_2	Organic layers	>1000 nm/min
Acetone	Organic layers	>4000 nm/min
H_3PO_3/ $HNO_3/HC_2H_3O_2$	Aluminum	Etch rate is 660 nm/min 40–50°C
HNO_3/BHF/water	Si and polysilicon	0.1 μm/min for single-crystal Si

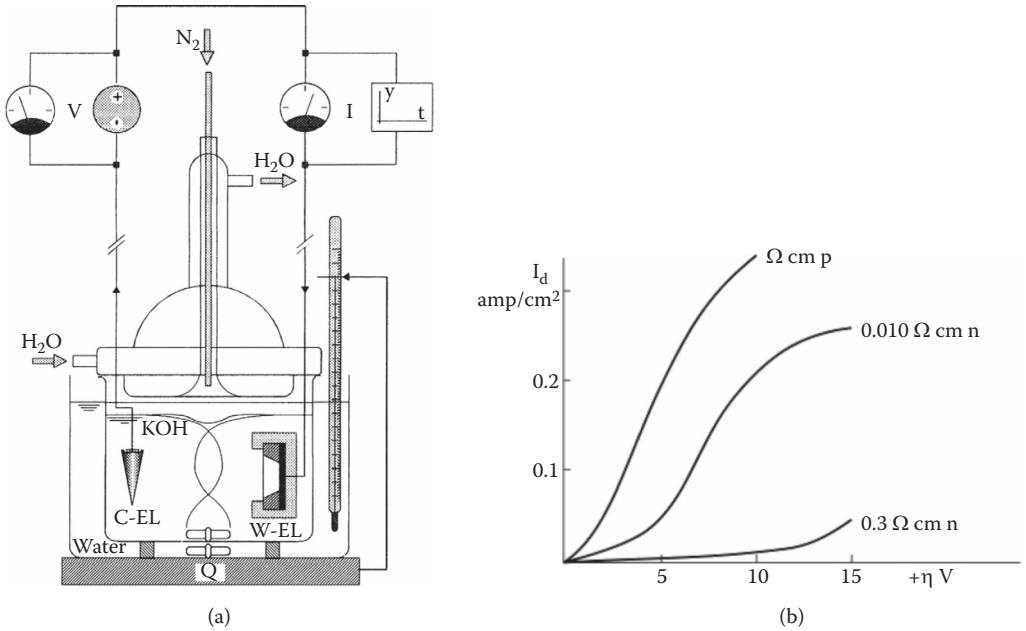


FIGURE 4.36 (a) Electrochemical etching apparatus. W-EL, working electrode (Si); C-EL, counter electrode (e.g., Pt); Q, heat supplied. (b) Current-voltage (*I-V*) curves in electrochemical etching of silicon of various doping. Etch rate dependence on dopant concentration and dopant type for HF-anodic etching of Si. (From van Dijk, H. J. A., and J. de Jonge. 1970. Preparation of thin silicon crystals by electrochemical thinning of epitaxially grown structures. *J Electrochem Soc* 117:553–54. With permission.[89])

dissolution is plotted (*I-V*). The current density is related to the dissolution rate of silicon. It can be seen that p-type and heavily doped n-type materials can be dissolved at relatively low voltages, whereas n-type silicon with a lower doping level does not dissolve at the same low voltages. Experiments in this same setup with homogeneously doped silicon wafers show that n-type silicon of about 3×10^{18} cm^{-3} (<0.01 Ω cm) completely dissolves in these etching conditions, whereas n-type silicon of donor concentrations lower than 2×10^{16} cm^{-3} (>0.3 Ω cm) barely dissolves. For p-type silicon, dissolution is initiated when the acceptor concentration is higher than 5 × 1015 cm–3 (<3 Ω cm), and the dissolution rate further increases with increasing acceptor concentration.

For more precise electrochemical isotropic etching work, the two-electrode system from Figure 4.36a is replaced with a three-electrode system and a potentiostat. The latter is an instrument used to control the potential of the semiconductor working electrode precisely with respect to a third electrode, the so-called reference electrode (REF) (for example, see Figure 4.46 for a three-electrode and potentiostat setup, and for more theoretical electrochemistry background see Volume I, Chapter 7). The current I passing between the W-EL and the C-EL is recorded as a function of the potential. No current passes through the reference electrode. The potential voltage of the W-EL can be scanned over a given potential range at a certain rate (mV/s). A typical *I-V* curve, also called a cyclic voltammogram, for a p-type silicon sample in an HF solution is shown in Figure 4.37, where the reference electrode used is an Ag/AgCl electrode.

The shape of the curve shown here is determined by the speed at which reactions occur at the silicon surface, and the rate of transport of reacting particles. In general the *I-V* curve for a Si electrode in an HF solution can be divided into three ranges: the cathodic range, the porous range, and the electropolishing range. In the cathodic range, i.e., the potential range negative to the open-circuit potential, the surface hole concentration in the Si electrode is low; therefore, no etching can take place. For p-type Si electrodes in the dark there is also no hydrogen evolution for lack of electrons. If an n-type Si electrode were used, copious hydrogen evolution would be observed in this same voltage range because n-type electrodes have plenty of electrons in the conduction band. The porous range is the potential range

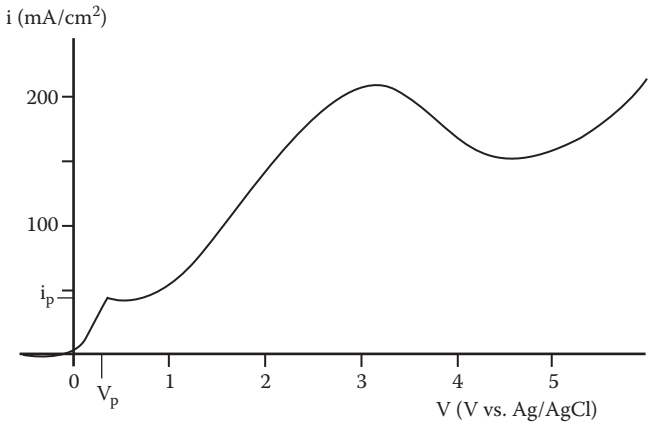


FIGURE 4.37 *I-V* curve of a p-type silicon wafer in a 5% HF solution. Resistivity of the wafer 0.01–0.018 Ωcm, scan rate 500 mV/s. (From Tjerkstra, R. W. 1999. *Isotropic etching of silicon in fluoride containing solutions as a tool for micromachining.* PhD diss., University of Twente, Twente, the Netherlands.[74])

before the potential of the first current peak (V_p in Figure 4.37) where porous Si forms. During the formation of porous Si, hydrogen gas is evolved, and the morphology of the porous Si depends, among other factors, on the doping density of the wafer, the HF concentration, and the current density. The potential range after the first current peak (V_p) is called the electropolishing range. Here, the current density exceeds the so-called critical current density, I_p, marked by the first peak in the *I-V* curve. The silicon etches, leaving behind a smooth (electropolished) surface. Porous Si does not form in this range. The critical current density depends on the fluoride concentration in the electrolyte solution, the dopant concentration in the semiconductor, the hydrodynamics of the system, and the temperature.[75] For a given Si electrode, the critical current density is given as[91]:

$$I_p = Ac^{1.5}e^{\frac{-E_a}{kT}} \qquad (4.1)$$

where the constant, A, is 3300 A cm^{-2} (wt% HF)$^{-1.5}$, c is HF concentration, the activation energy E_a is 0.345 eV, k is Boltzmann constant, and T is the temperature. When the current density is lower than I_p, a porous layer is formed on the Si substrate.

Many researchers have investigated the electrochemical etching of silicon in HF solutions, and initially the following reaction mechanisms were suggested as[9]:

$$Si + 2HF + 2h^+ \rightarrow SiF_2 + 2H^+ \qquad \text{Reaction 4.5}$$

In this scenario, in the porous Si range, two holes are necessary to dissolve each silicon atom. Thus, it was assumed that the silicon dissolves as an Si(II) species, after which it is further oxidized by water or disproportionates to an Si(IV) and Si(0) species. One ends up with the soluble SiF_6^{2-} complex [Si(IV)] and porous Si [Si(0)]. The porous Si deposited on the surface was assumed to be amorphous because it stemmed from the disproportionation reaction of SiF_2. The hydrogen evolution in the porous Si range was believed to be caused by a reaction of the silicon subfluoride:

$$SiF_2 + 2H_2O \rightarrow SiO_2 + 2HF + H_2 \qquad \text{Reaction 4.6}$$

However, porous Si was shown to be monocrystalline, and this theory had to be discarded.[92,93] It was also discovered that the freshly etched Si surface is H-terminated, making the Si surface hydrophobic.[94] Current reaction mechanisms are generally based on the assumption that the presence of a fluoride ion close to the substrate surface causes the Si-Si backbonds to be polarized, thus facilitating the breaking of these bonds.[95–97] The backbonds are broken one at a time, leaving an H-terminated surface and an SiH_xF_y compound that reacts with water and HF to form SiF_6^{2-} and H_2. The overall reaction is then:

$$Si + 2H^+ + 2h^+ \rightarrow Si(IV) + H_2 \qquad \text{Reaction 4.7}$$

At pH < 3, fluoride is mainly present in the form of HF or HF_2^-, so these components are also assumed to be involved in the reactions. It is generally assumed that the electropolishing of silicon takes place in two steps. In the first step an oxide layer is formed, and this oxide layer is subsequently dissolved by HF. Dissolution rates and thickness of the oxide layer have been studied using infrared[98] and current-time measurements.[96,99–102] The second peak in the *I-V* curve in Figure 4.37 is thought to mark a change from one type of oxide to another.[103] The overall reaction of the dissolution of Si in the polishing range is:

$$Si + 2H_2O + 4h^+ \rightarrow SiO_2 + 4H^+ \qquad \text{Reaction 4.8}$$

$$SiO_2 + 6HF \rightarrow SiF_6^{2-} + 2H_2O + 2H^+ \quad \text{Reaction 4.9}$$

For the dissolution of n-type silicon, holes have to be generated by illuminating the sample with photons having energy larger than the bandgap. If the photon flux is high enough, an *I-V* curve similar to that shown in Figure 4.37 can be obtained. If the photon flux is lower than the critical current density, porous Si forms and a potential-independent photocurrent is observed (see further below).

The acidic electrochemical technique has not been used much in micromachining and is primarily used to polish surfaces. Because the etching rate increases with current density, high spots on the surface are more rapidly etched, and very smooth surfaces result. This method of isotropic electrochemical etching has some major advantages that could make it a more important micromachining tool in the future. Such advantages include very smooth etched surfaces (e.g., with an average roughness R_a of 7 nm), room temperature and IC compatible process, usage of simpler resist schemes because the process is much milder than etching in HNA, and etching can be controlled simply by switching a voltage on or off. We will pick up the discussion of anodic polishing, photoetching, and formation of porous Si in HF solutions after gathering more insight into various etching models.

Anisotropic Etching

Introduction Anisotropic etchants shape, also "machine," desired structures in crystalline materials by etching much faster in one direction than another. When carried out properly, anisotropic etching results in geometric shapes bounded by the slowest etching and perfectly defined crystallographic planes. Anisotropic wet etching techniques, dating back to the 1960s at the Bell Laboratories, were developed mainly by trial and error. Figure 4.38 shows the cross-section of a variety of shapes formed in single-crystal Si using anisotropic etching. The thinned membrane shown in Figure 4.38a with diffused resistors may be used for a piezoresistive pressure sensor. In this application, the wafer is selectively thinned from a starting thickness of 300–500 μm to form a diaphragm having a final thickness of 10–20 μm with precisely controlled lateral dimensions and a thickness control of the order of 1 μm or better. A typical procedure involves the steps summarized in Table 4.8.[53]

The development of anisotropic etchants solved the lateral dimension control lacking in isotropic etchants. Lateral mask geometries on planar photoengraved substrates can be controlled with an accuracy and reproducibility of 0.5 μm or better, and the anisotropic nature of the etchant allows this accuracy to be translated into control of the vertical

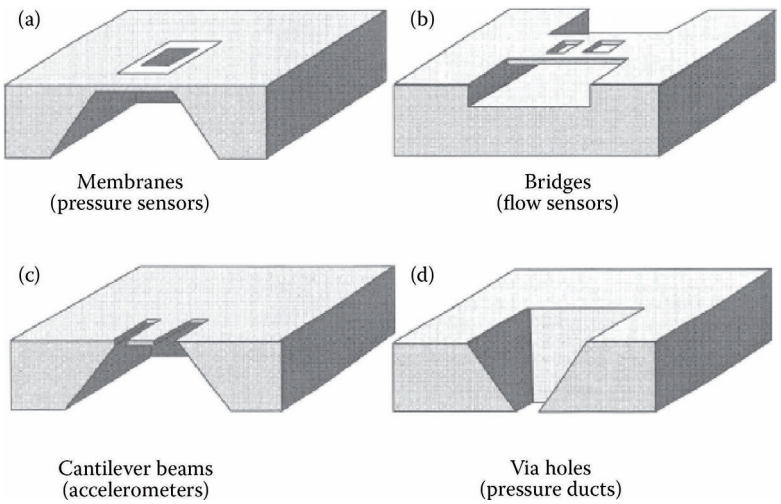


FIGURE 4.38 Anisotropic wet etching in bulk Si. Membranes for (a) pressure sensors, (b) bridges for flow sensors, (c) cantilevers for accelerometers, and (d) via holes for pressure duct are shown.

Table 4.8 Summary of the Process Steps Required for Anisotropic Etching of a Membrane in Si

Process	Duration	Process Temperature (°C)
Oxidation	Variable (h)	900–1200
Spinning resist at 5000 rpm	20–30 s	Room temperature
Prebake	10 min	90
Exposure	20 s	Room temperature
Develop	1 min	Room temperature
Postbake	20 min	120
Strippping of oxide (BHF:1:7)	±10 min	Room temperature
Stripping resist (acetone)	10–30 s	Room temperature
RCA1 (NH_3 (25%)+ H_2O+H_2O_2:1:5:1)	10 min	Boiling
RCA2 (HCl+H_2O+ H_2O_2:1:6:1)	10 min	Boiling
HF-dip (2% HF)	10 s	Room temperature
Anisotropic etch	From minutes up to 1 day	70–100

Source: Elwenspoek, M., H. Gardeniers, M. de Boer, and A. Prak. 1994. *Micromechanics*. Report No. 122830. Twente, the Netherlands: University of Twente.[53]

etch profile. Different etch stop techniques, needed to control the thickness of an active sensor component (such as the membrane or cantilever structures in Figure 4.38), are available. The invention of etch stop techniques truly made the application of the bulk micromachined structures shown in Figure 4.38 commercially manufacturable and viable.

Although anisotropic etchants solve problems of lateral control associated with isotropic etching, they are not problem-free. These etchants are slow—even in the fast etching <100> direction—with etch rates of about 1 µm/min or less. That means that etching through a wafer is a time-consuming process: to etch through a 300-µm-thick wafer, one needs 5 hours. They also must be run hot to achieve these etch rates (80–115°C), precluding many simple masking options. Like the isotropic etchants, their etch rates are temperature sensitive; however, they are not particularly agitation sensitive, which is considered to be a major advantage.

A wide variety of etchants have been used for anisotropic etching of silicon, including alkaline aqueous solutions of KOH, NaOH, LiOH, CsOH, RbOH, NH_4OH, and quaternary ammonium hydroxides, with the possible addition of alcohol. Alkaline organics such as ethylene-diamine, choline (trimethyl-2-hydroxyethyl ammonium hydroxide), hydrazine, and sodium silicates with additives such as pyrocatechol and pyrazine are used as well. Etching of silicon is possible without the application of an external voltage and is dopant insensitive over several orders of magnitude; however, in a curious contradiction to its suggested chemical nature, it has been shown to be bias-dependent.[104,105] This contradiction will be explained with the help of the etching models presented below.

Alcohols such as propanol and isopropanol butanol typically slow the attack on Si.[106,107] The role of pyrocatechol[108] is to speed up the etch rate through complexation or chelation of the reaction products. Additives such as pyrazine and quinone have been described as catalysts by some,[109] but this is contested by others.[110] The etch rate in anisotropic etching is reaction rate controlled and thus temperature-dependent. The etch rate for all planes increases with temperature, and the surface roughness decreases with increasing temperature; thus, etching at higher temperatures gives the best results. In practice, etch temperatures of 80–85°C are used to avoid solvent evaporation and temperature gradients in the solution.

Selected Anisotropic Etchant Systems

Overview In choosing an anisotropic wet etchant, a variety of issues must be considered:

- Ease of handling
- Toxicity
- Etch rate
- Desired topology of the etched bottom surface
- IC compatibility
- Etch stop availability
- Etch selectivity over other materials
- Mask material and thickness of the mask

The principal characteristics of four different anisotropic etchants are listed in Table 4.9. The most commonly used are KOH[10–23,52,110–118] and ethylenediamine pyrocatechol + water (EDP)[108,109,119];

Table 4.9 Principal Characteristics of Four Different Anisotropic Etchants

Etchant/ Diluent/ Additives/ Temperature	Etch Stop	Etch Rate (100) μm/min	Etch Rate Ratio	Remarks	Mask (Etch Rate)
KOH (water) 85°C 44 g/100 ml	$B > 10^{20}$ cm^{-3} reduces etch rate by 20	1.4	400 for (100)/(111) and 600 for (110)/(111)	IC incompatible, avoid in eyes, etches oxide fast, lots of H_2 bubbles	Photoresist (shallow etch at room temperature); Si_3N_4 (<1 nm/min) SiO_2 (28 Å/min)
Ethylenediamine pyrocatechol (water) 115°C 750 ml/120 g/240 ml	$\geq 7 \times 10^{19}$ cm^{-3} reduces the etch rate by 50	1.25	35 for (100)/(111)	Toxic, ages fast, O_2 must be excluded, few H_2 bubbles, silicates may precipitate	SiO_2 (2–5 Å/min), Si_3N_4 (1 Å/min), Ta, Au, Cr, Ag, Cu are not attacked, Al at a 0.33 μm/min
Tetramethyl ammonium hydroxide (TMAH) (water) 90°C	$>4 \times 10^{20}$ cm^{-3} reduces etch rate by 40	1	From 12.5–50 (100)/(111)	IC compatible, easy to handle, smooth surface finish, few studies	SiO_2 etch rate is four orders of magnitude lower than (100) LPCVD Si_3N_4
N_2H_4 (water, isopropyl alcohol) 100°C 100 ml/100 ml	$>1.5 \times 10^{20}$ cm^{-3} practically stops the etch	2.0	10 (100)/(111)	Toxic and explosive, OK at 50% water	SiO_2 (<2 Å/min) and most metallic films; does not attack aluminum

Given the many possible variables, the data in the table are only typical examples.

hydrazine-water is rarely used.[120,121] Quaternary ammonium hydroxide solutions such as tetramethyl ammonium hydroxide-water (TMAHW) and tetraethyl ammonium hydroxide-water (TEAHW) have become popular.[122,123] Each has its advantages and problems. NaOH is not used much anymore.[124]

Potassium Hydroxide The simple KOH water system is the most popular Si anisotropic etchant. A KOH etch, in near-saturated solutions (1:1 in water by weight) at 80°C, produces a uniform and bright surface. Williams and Muller[125] used 50 wt% KOH at 80°C for a reported (100) etch rate of 1.4 μm/min. Nonuniformity of etch rate gets considerably worse at more than 80°C. An abundance of bubbles are seen emerging from the Si wafer while etching in KOH. The etching selectivity between Si and SiO_2 is not very good in KOH because it etches SiO_2 too quickly. KOH is also incompatible with the IC fabrication process (e.g., aluminum bond pads are quickly attacked and damaged) and can cause blindness when it gets in contact with the eyes. The etch rate for low index planes is maximal at around 4 M (see Figure 4.39a and c from Peeters[126] and Lambrechts et al.[127])

Herr[54] found that the high-index crystal planes exhibit the highest etch rates for 6 M KOH and that for lower concentrations the etch bottoms disintegrate into microfacets. In 6 M KOH, the etch-rate order is (311) > (144) > (411) > (133) > (211) > (122). These authors could not correlate the particular etch rate sequence with the measured activation energies. This is in contrast to lower activation energies corresponding to higher etching rates for low index planes as shown further below in Figure 4.40. Their results obtained on large open-area structures differ significantly from previous ones obtained by underetching special mask patterns. The vertical etching rates obtained here are substantially higher than the underetching rates described elsewhere, and the etch rate sequence for different planes is also significantly different. These results suggest that crevice effects may play an important role in anisotropic etching.

To create vertical (100) faces, as shown in Volume I, Chapter 4 (Figure 4.10), in general only KOH works (not EDP or TMAHW), and it has to be carried out in high-selectivity conditions (low temperature, low concentration: 25 wt% KOH, 60°C). Interestingly, high-concentration KOH (45 wt%) at higher temperatures (80°C) produces a smooth sidewall, controllable and repeatable at an angle of 80°. EDP produces 45° angled planes, and TMAHW usually makes a 30° angle.[129]

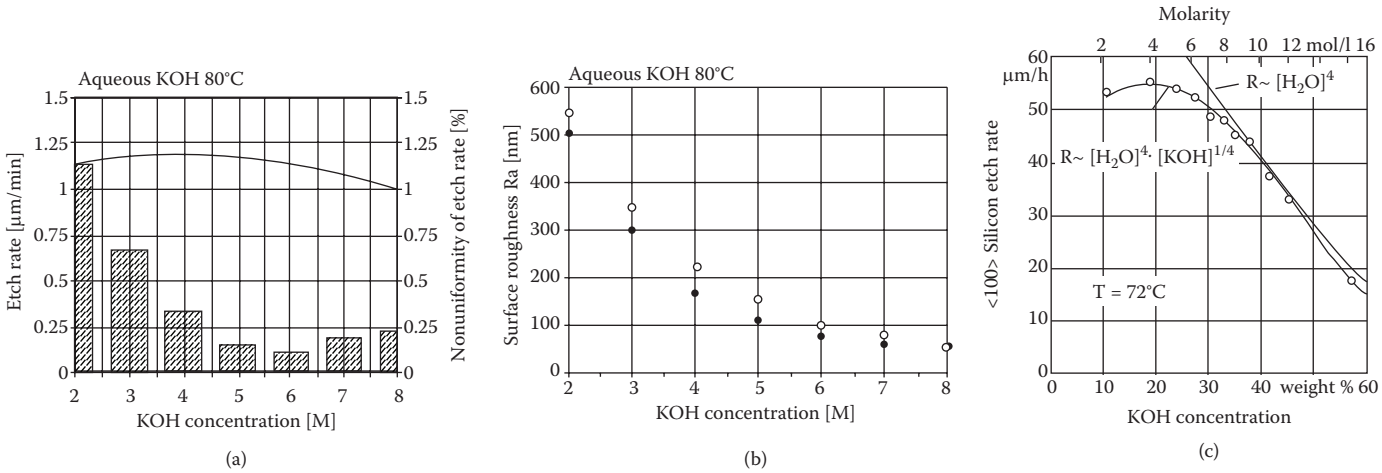


FIGURE 4.39 Anisotropic etching of silicon. (a) Silicon (100) etch rate (line) and nonuniformity of etch rate (column) in KOH at 80°C as a function of KOH concentration. The etch rate for all low index planes is maximal at around 4 M. (b) Silicon (100) surface roughness (Ra) in aqueous KOH at 80°C as a function of concentration for a 1-hour etch time (thin line) and for an etch depth of 60 μm (thick line). (c) Silicon (100) etch rate as a function of KOH concentration at a temperature of 72°C. (a and b from Peeters, E. 1994. Process development for 3D silicon microstructures, with application to mechanical sensor design. Ph.D. diss., Catholic University of Louvain, Belgium. With permission. c from Seidel, H., L. Csepregi, A. Heuberger, and H. Baumgartel. 1990. Anisotropic etching of crystalline silicon in alkaline solutions–part I: orientation dependence and behavior of passivation layers. *J Electrochem Soc* 137:3612–26. With permission.[126,52])

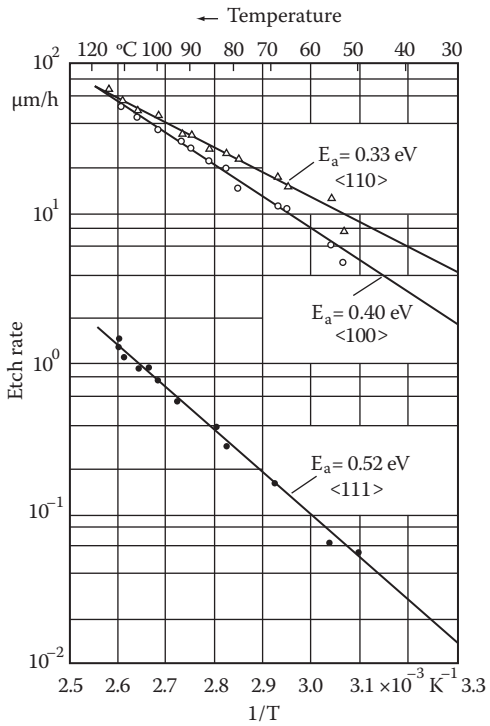


FIGURE 4.40 Vertical etch rates as a function of temperature for different crystal orientations: (100), (110), and (111). Etch solution is EDP (133 ml H_2O, 160 g pyrocatechol, 6 g pyrazine, and 1 l ED). (From Seidel, H., L. Csepregi, A. Heuberger, and H. Baumgartel. 1990. Anisotropic etching of crystalline silicon in alkaline solutions–part I: orientation dependence and behavior of passivation layers. *J Electrochem Soc* 137:3612–26. With permission.[52])

Besides KOH,[129] other hydroxides have been used, including NaOH,[104,124] CsOH,[130] and NH_4OH.[130] A major disadvantage of KOH is the presence of alkali ions, which are detrimental to the fabrication of sensitive electronic parts.

Ethylenediamine Pyrocatechol With ethylenediamine pyrocatechol (EDP; sometimes referred to as EPW, for ethylenediamine, pyrocatechol, and water), a variety of masking materials (e.g., SiO_2, Si_3N_4, Au, Cr, Cu, Ag, and Ta) can be used. EDP is less toxic than hydrazine (see below), no sodium or potassium contamination occurs, and the etch rate of SiO_2 is much slower than with KOH. The ratio of etch rates of Si and SiO_2 using EDP can be as large as 5000:1, corresponding to about 2 Å/min of SiO_2 compared with 1 μm/min of Si, which is much larger than the ratio for Si to SiO_2 in KOH, which is at the highest 400:1.[131] Importantly, the etch rate slows down at a lower boron concentration than with KOH. A typical fastest-to-slowest hierarchy of Si etch rates with EDP at 85°C according to Barth[132] is (110) > (411) > (311) > (511) > (211) > (100) > (331) > (221) > (111). Petersen uses 750 ml of ethylenediamine, 120 g of pyrocatechol, and 100 ml of water. This cocktail, at

115°C, results in an etch rate of 0.75 μm/min, with an (100)/(111) etch-rate ratio of 35:1.[75]

Ethylenediamine in EDP reportedly causes allergic respiratory sensitization, and pyrocatechol is described as a toxic corrosive. The material is also optically dense, making end-point detection harder, and it ages quickly; if the etchant reacts with oxygen, the liquid turns a red-brown color, and it loses its useful properties. If cooled down after etching, precipitation of silicates in the solution will occur. Sometimes, precipitation during etching can happen, spoiling the results. When preparing the solution, the last ingredient added should be water because water addition causes the oxygen sensitivity. All the above make the etchant difficult to handle.

In terms of etching and masking layers, amine gallates are similar to EDP but perhaps safer.[133] Amine gallate etchants are not used much but appear promising. They are composed of a mixture of ethanolamine, gallic acid, water, pyrazine, hydrogen peroxide, and a surfactant. Etch rates as high as 2.3 μm/min have been measured on a (100) Si plane, and etch stops at lower boron concentration than it takes to stop EDP have been observed ($>3 \times 10^{19}$ cm^{-3}). Pyrazine and peroxide can be added to increase the etch rate, but they affect the surface roughness negatively.

Ammonium Hydroxide-Water and Tetramethyl Ammonium Hydroxide-Water Efforts continue to find anisotropic etchants that are more compatible with complementary metal oxide semiconductor (CMOS) processing than alkali hydroxides and that are neither toxic nor harmful. Two examples are ammonium hydroxide-water (AHW) mixtures[130] and tetramethyl ammonium hydroxide-water (TMAHW) mixtures (Tabata et al.[123] and Schnakenberg et al.[130]). Kern[77] used AHW (9.7% in water) and achieved 0.11 μm/min etch rates on (100) Si at temperatures of 85–92°C. Schnakenberg et al.[130] reported their best AHW results with a 3.7 wt% solution at 75°C for well-stirred etching baths. For the same solution, these authors demonstrated a boron-dependent etch stop at 1.3×10^{20} cm^{-3} with a selectivity of 1:8000 (see also under Etch-Stop Techniques, this chapter). Ammonia-based etchants have not become widely used for several reasons, including their slow etch rate, tendency to lead to rough surfaces (hillocks), and rapid evaporative losses.[72] A TMAHW [$(CH_3)_4NOH$] solution, on the other hand, is one of the more useful wet etchants for silicon. TMAHW does not decompose at temperatures less than 130°C, a very important feature from the viewpoint of production; it is nontoxic, not expensive, and can be handled easily. TMAHW solutions also exhibit excellent selectivity to silicon oxide and silicon nitride masks. The etchant is so selective for Si over SiO_2 that it is advisable to remove the thin native oxide of Si in HF before attempting a TMAHW etch. The solution is often already present in the clean room because it is used in many positive photoresist developers. At a solution temperature of 90°C and 22 wt% TMAH, a maximum (100) silicon etch rate of 1.0 μm/min is observed, 1.4 μm/min for (110) planes [this is higher than those observed with EDP, AHW, hydrazine water, and tetraethyl ammonium hydroxide (TEA), but slower than those observed for KOH] and an anisotropy ratio, AR(100)/(111), of between 12.5 and 50.[134] A disadvantage is that TMAHW occasionally results in the formation of pyramidical hillocks at the bottom of the etched cavity. From the viewpoint of fabricating various silicon sensors and actuators, a concentration greater than 22 wt% is preferable because lower concentrations result in more pronounced roughness on the etched surface. However, higher concentrations give a lower etch rate and lower etch ratio (100)/(111). Tabata[135] also studied the etching characteristics of pH-controlled TMAHW. To obtain a low aluminum etching rate of 0.01 μm/min, pH values less than 12 for 22 wt% TMAHW were required. At those pH values the Si (100) etching rate is 0.7 μm/min. The aluminum etch rate can also be reduced by adding silicon powder to the etchant.[136] The etch rate for TMAHW begins to slow down for boron doping levels above approximately 1×10^{19} cm^{-3} and decreases by a factor of 40 for 2×10^{20} cm^{-3}.[137]

Hydrazine (H_2N_4) Etch rates with hydrazine-water mixtures are on the order of 2 μm/min, and similar masking materials can be used as with EDP.[75] The (100)/(111) etch ratios are lower than those for KOH or EDP. Hydrazine-water is very reducing and explosive at high hydrazine concentrations (rocket

fuel), is a suspected carcinogen, and is difficult to dispose so its use should be avoided for safety reasons. However, a 50 wt% hydrazine-water solution is stable and is commonly used, and, according to Mehregany,[121] excellent surface quality and sharply defined corners are obtained in Si. Also on the positive side, the etchant has a very low SiO_2 etch rate and will not attack most metal masks except for Al, Cu, and Zn. According to Wise, on the other hand, Al does not etch in hydrazine either, but the etch produces rough Si surfaces.[138]

Arrhenius Plots for Anisotropic Etching A typical set of Arrhenius plots for <100>, <110>, and <111> silicon etching in an anisotropic etchant (e.g., EDP) is shown in Figure 4.34.[52] It is seen that the temperature dependence of the etch rate is large and that the slope differs for the different planes, that is, (111) > (100) > (110). Lower activation energies in Arrhenius plots correspond to higher etch rates. The anisotropy ratio (AR) derived from this figure is:

$$\mathrm{AR} = \frac{(\mathrm{hkl})_1 \text{ etch rate}}{(\mathrm{hkl})_2 \text{ etch rate}} \tag{4.2}$$

The AR is approximately 1 for isotropic etchants and can be as high as 400:200:1 for (110)/(100)/(111) in 50 wt% KOH/H_2O at 85°C. Generally, the activation energies of the etch rates of EDP are smaller than those of KOH. Price found that (111) planes always etch slowest, and the selectivity with respect to (100) in KOH etching can be greatly increased by adding isopropyl alcohol (IPA), a less polar diluent, used to saturate the solution.[107] The sequence for (100) and (110) etching rate can be reversed, for example, to 50:200:8 in 55 vol% ethylene/diamine/H_2O (also at 85°C). The (110) Si plane etches eight times slower and the (111) eight times faster in KOH/H_2O than in ED/H_2O, whereas the (100) etches at the same rate.[111] Working with alcohols and other organic additives often changes the relative etching rate of the different Si planes. Along this line, Seidel et al.[52,110] found that the decrease in etch rate by adding isopropylalcohol to a KOH solution was 20% for the <100> planes but almost 90% for <110> planes. As a result of the much stronger decrease of the etch rate on a (110) surface, the etch ratio of (100)/(110) is reversed.

Misalignment will change the etch rate greatly; a 1° misalignment on the [111] direction may increase the etch rate on (near) the (111) surface by 300%.[139]

Si Surface Roughness in Anisotropic Etching Aniso-tropic etchants frequently leave too rough a Si surface behind, and a slight isotropic etch is used to "touch up" the surface. The average surface roughness, R_a, of Si continuously decreases with increasing KOH concentration as can be gleaned from Figure 4.39b. The silicon etch rate, R, as a function of KOH concentration is shown in Figure 4.39a and c.[52] Because the difference in etch rates for different KOH concentrations is small, a highly concentrated KOH (e.g., 7 M) is preferred in obtaining a smooth surface on low index planes. Except at very high concentrations of KOH, the etched (100) plane becomes rougher the longer one etches. This is thought to be because of the development of hydrogen bubbles, which hinder the transport of fresh solution to the silicon surface, causing "micromasking," which results in hillock formation.[72,140] Average roughness, R_a, is influenced strongly by fluid agitation. Stirring can reduce the R_a values more than an order of magnitude, probably caused by the more efficient removal of hydrogen bubbles from the etching surface when stirring.[141] Ohwada et al.[142] noted that the use of ultrasonic agitation essentially eliminated surface roughness in KOH etching. Baum et al.[143] investigated the surface finish of (100) Si in KOH etching with an atomic force microscope and confirmed that mild ultrasonic agitation improved the surface finish considerably. Hillock formation may also be suppressed by the addition of a suitable oxidizing agent that competes with hydrogen evolution, such as by adding ferricyanide or peroxydisulfate ions. Bressers et al.[144] report a drastic reduction in hillock formation when adding 18 mM $K_3Fe(CN)_6$ to a 4 M KOH solution at 70°C, and Klaassen et al.[145] accomplish the same by adding 5 g/L ammonium peroxydisulfate to a 5-wt% TMAHW solution. Baum et al.[143] found that the inclusion in the KOH bath of oxygen and/or isopropanol results in root mean square roughness values smaller than 20 nm. The effectiveness of these additives has been related to

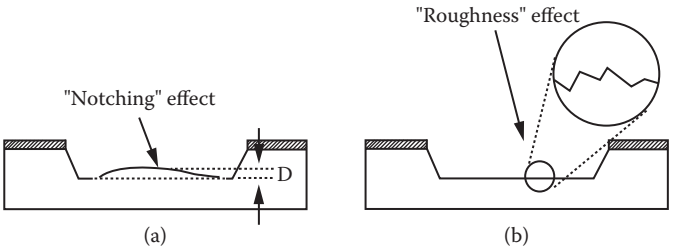


FIGURE 4.41 (a) Macroscopic roughness (notching effect) and (b) microscopic roughness on a Si membrane.

changes of the contact angle between the liquid/gas/etching interface.

A distinction must be made between macroscopic and microscopic roughness (Figure 4.41). Macroscopic roughness, also referred to as notching or pillowing, results when centers of exposed areas etch with a seemingly lower average speed compared with the borders of the areas; therefore, the corners between sidewalls and (100) ground planes are accentuated. Thus, membranes or double-sided clamped beams (microbridges) tend to be thinner close to the clamped edges than in the center of the structure. This difference can be as large as 1–2 µm, which is considerable if one is etching 10–20-µm-thick structures. Notching increases linearly with etch depth but decreases with higher concentrations of KOH. The microscopic smoothness of originally mirror-like polished wafers can also be degraded into microscopic roughness. It is this type of short-range roughness we referred to in discussing Figure 4.39b above. For more background on metrology techniques to measure surface roughness, see Volume III, Chapter 6.

Masking for Anisotropic Etchants Etching through a whole wafer (400–600 µm) takes several hours (a typical wet anisotropic etch rate is 1.1 µm/min), definitely not a fast process. When using KOH, SiO_2 cannot be used as a masking material for features requiring that long an exposure to the etchants. The SiO_2 etch rate as a function of KOH concentration at 60°C is shown in Figure 4.42. There is a distinct maximum at 35 wt% KOH of nearly 80 nm/h. The shape of this curve will be explained further below based on Seidel et al.'s etching model. Experiments have shown that even a 1.5-µm-thick oxide is not sufficient for the complete etching of a 380-µm-thick wafer (6 hours) because of pinholes in the oxide.[128]

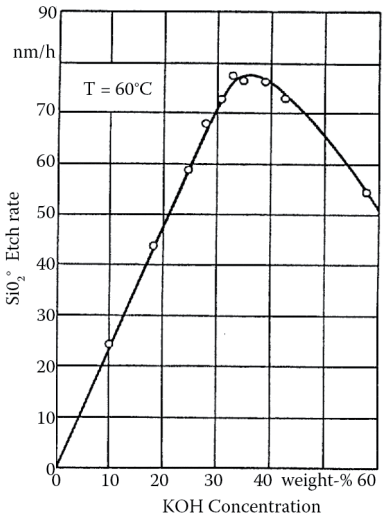


FIGURE 4.42 The SiO_2 etch rate in nanometers per hour as a function of KOH concentration at 60°C. (From Seidel, H., L. Csepregi, A. Heuberger, and H. Baumgartel. 1990. Anisotropic etching of crystalline silicon in alkaline solutions–part I: orientation dependence and behavior of passivation layers. *J Electrochem Soc* 137:3612–26. With permission.[52])

The etch rate of thermally grown SiO_2 in KOH-H_2O somewhat varies and apparently depends not only on the quality of the oxide but also on the etching container and the age of the etching solution, as well as other factors.[48] The Si/SiO_2 selectivity ratio at 80°C in 7 M KOH is 30 ± 5. This ratio increases with decreasing temperature; reducing the temperature from 80 to 60°C increases the selectivity ratio from 30 to 95 in 7 M KOH.[57] Thermal oxides are under strong compressive stress because in the oxide layer one silicon atom takes nearly twice as much space as in single-crystalline Si (see also Volume I, Chapter 4). This might have severe consequences; for example, if the oxide mask is stripped on one side of the wafer, the wafer will bend. Atmospheric pressure chemical vapor-deposited (APCVD) SiO_2 tends to exhibit pinholes and etches much faster than thermal oxide. Annealing of APCVD oxide removes the pinholes, but the etch rate in KOH remains greater by a factor of 2–3 than that of thermal oxide. Low-pressure chemical vapor-deposited (LPCVD) oxide is a mask material of comparable quality as thermal oxide. The etch rate of SiO_2 in EDP is smaller by two orders of magnitude than in KOH.

For prolonged KOH etching, a high-density silicon nitride mask has to be deposited. An LPCVD nitride generally serves better for this purpose than

less-dense plasma-deposited nitride.[146] With an etch rate of less than 0.1 nm/min, a 400-Å layer of LPCVD nitride suffices to mask against KOH etchant. The etch selectivity Si/Si_3N_4 was found to be better than 10^4 in 7 M KOH at 80°C. The nitride also acts as a good ion-diffusion barrier, protecting sensitive electronic parts. Nitride can easily be patterned with photoresist and etched in CF_4/O_2-based plasma or, in a more severe process, in H_3PO_4 at 180°C (10 nm/min).[147] Nitride films are typically under a tensile stress of about 1×10^9 Pa. If, in the overall processing of the devices, nitride deposition does not pose a problem, KOH emerges as the preferential anisotropic wet etchant. For dopant-dependent etching, EDP is the better etchant and generally better suited for deep etching because its oxide etch rate is negligible (<5 Å/min).

Summarizing, oxide and nitride films mask for anisotropic etchants to varying degrees with both mask types being used. When these layers are used to terminate a Si etch in the [100] direction, a low etch rate of the mask layer allows overetching of silicon to compensate for wafer thickness variations. A KOH solution etches SiO_2 at a relatively fast rate of 1.4–3 nm/min; therefore, Si_3N_4 or Au/Cr must be used as a mask against KOH for deep and long etching.

Back-Side Protection of Silicon Wafer during Etching In many cases it is necessary to protect the back of a wafer from an isotropic or anisotropic etchant. The back is either mechanically or chemically protected. In the mechanical method, the wafer is held in a holder, often made from Teflon®. The wafer is fixed between Teflon-coated O-rings that are carefully aligned to avoid mechanical stress in the wafer. In the chemical method, waxes or other organic coatings are spun onto the back of the wafer. Two wafers may also be glued back to back for faster processing.

Etch-Stop Techniques

Introduction

In many cases, it is desirable to stop etching in silicon when a certain cavity depth or a certain membrane thickness is reached. Nonuniformity of etched devices resulting from nonuniformity of the silicon wafer thickness can be high. Taper of double-polished wafers, for example, can be as high as 40 µm![127] Even with the best wafer quality, the wafer taper is still around 2 µm. The taper and variation in etch depth lead to intolerable thickness variations for many applications. High-resolution silicon micromachining relies on the availability of effective etch-stop layers. It is the existence of impurity-based etch stops in silicon that has allowed micromachining to become a high-yield commercial production process. Etch-rate control typically requires monitoring and stabilization of the following parameters:

- Etchant composition
- Etchant aging
- Stabilization with N_2 sparging (especially with EDP and hydrazine)
- Taking account of the total amount of material etched (loading effects)
- Etchant temperature
- Diffusion effects (constant stirring is required, especially for EDP)
- Stirring also leads to a smoother surface through bubble removal
- Trenching (also pillowing) and roughness decrease with increased stirring rate
- Light may affect the etch rate (especially with n-type Si)
- Surface preparation of the sample can have a big effect on etch rate (the native oxide retards etch start; a dip in dilute HF is recommended)

With a good temperature, etchant concentration, and stirring control, the variation in etch depth typically is 1% (see Figure 4.43).[127] A good pretreatment of the surface to be etched is a standard RCA clean (see Volume II, Chapter 1, Wafer Cleaning and Contaminants) combined with a 5% HF dip to remove the native oxide immediately before etching in KOH.

In the early days of micromachining, one of the following techniques was used to etch a Si structure anisotropically to a predetermined thickness. In the simplest mode, the etch time was monitored (in Table 4.9 we listed some etch rates for different etchants) or a bit more sophisticated; the infrared

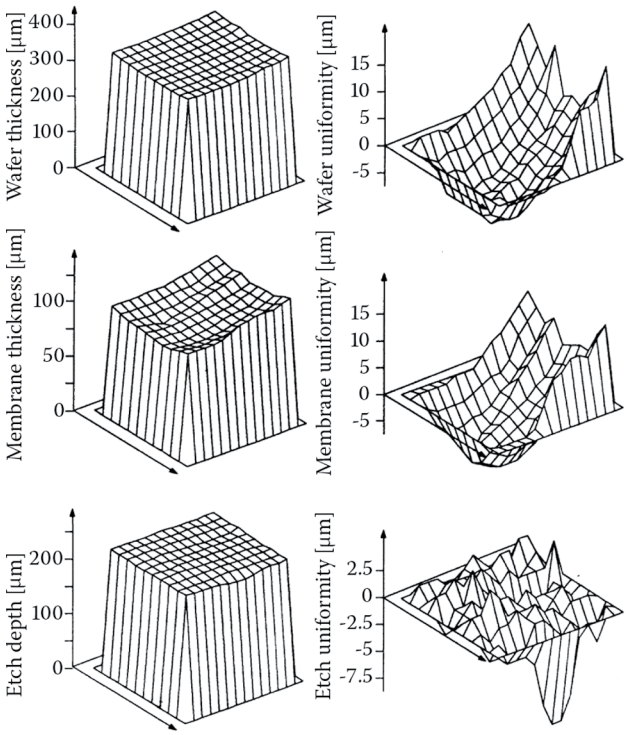


FIGURE 4.43 Map of wafer thickness, membrane thickness, and etch depth. (From Lambrechts, M., and W. Sansen. 1992. *Biosensors: microelectrochemical devices.* Philadelphia: Institute of Physics Publishing. With permission.[127])

transmittance through the etching membrane was followed. For thin membranes, the etch stop cannot be determined by a constant etch time method with sufficient precision. The spread in etch rates becomes critical if one etches membranes down to thicknesses of less than 20 µm; it is almost impossible to etch structures down to less than 10 µm with a timing technique. In the V-groove technique, V-grooves with precise openings (see Volume I, Equation 4.2, $W_0 = W_m - 2z\cot 54.74^\circ = W_m - \sqrt{2}z$) were used such that the V-groove stopped etching at the exact moment a desired membrane thickness was reached (see Figure 4.44).[148]

One can also design wider mask openings on the wafer's edge so that the wafer is etched through at those sites at the moment the membrane has reached the appropriate thickness. Although Nunn and Angell[149] claimed that an accuracy of about 1 µm could be obtained using the V-groove method, none of the mentioned techniques are found to be production worthy. Nowadays the above methods are almost completely replaced

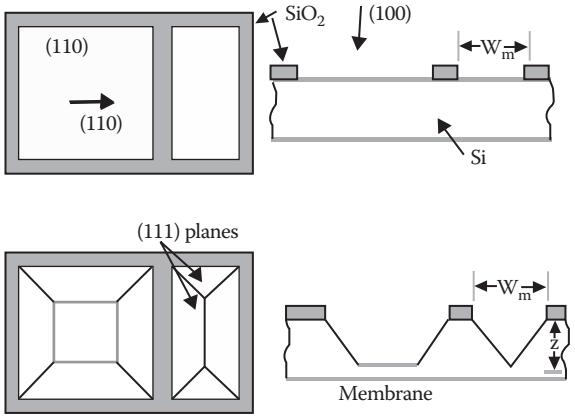


FIGURE 4.44 V-groove technique to monitor the thickness of a membrane. At the precise moment the V-groove is developed, the membrane has reached the desired thickness.

by etch-stop techniques based on a change in etch rate-dependent on doping level or the chemical nature of a stopping layer.

Boron Etch Stop

The most widely used etch-stop technique is based on the fact that anisotropic etchants, especially EDP, do not attack heavily boron-doped (p^{++}) silicon layers. Selective p^{++} doping is typically implemented using gaseous or solid boron diffusion source with a mask (such as silicon dioxide). The maximum practical depth achievable is 15 µm. The boron etch-stop effect was first noticed by Greenwood in 1969.[150] He assumed that the presence of a p-n junction was responsible. In 1971 Bogh[151] found that an impurity concentration of about $7 \times 10^{19}/cm^{-3}$ resulted in the etch rate of Si in EDP decreasing sharply (see also Table 4.9) but without any requirement for a p-n junction. For KOH-based solutions, Price[107] found a significant reduction in etch rate for boron concentrations greater than 5×10^{18} cm^{-3}. The model by Seidel et al. (see below) provides an elegant explanation for the etch stop at high boron concentrations.

From the above, it follows that a simple boron diffusion or implantation, introduced from the front of the wafer, can be used to create beams and diaphragms by etching from the back. A boron etch-stop technique is illustrated in Figure 4.45A for the fabrication of a micromembrane nozzle.[152] The SiO_2 mesa in Figure 4.45Ab leads to the desired boron p^{++} profile. The anisotropic etch from the back clears the

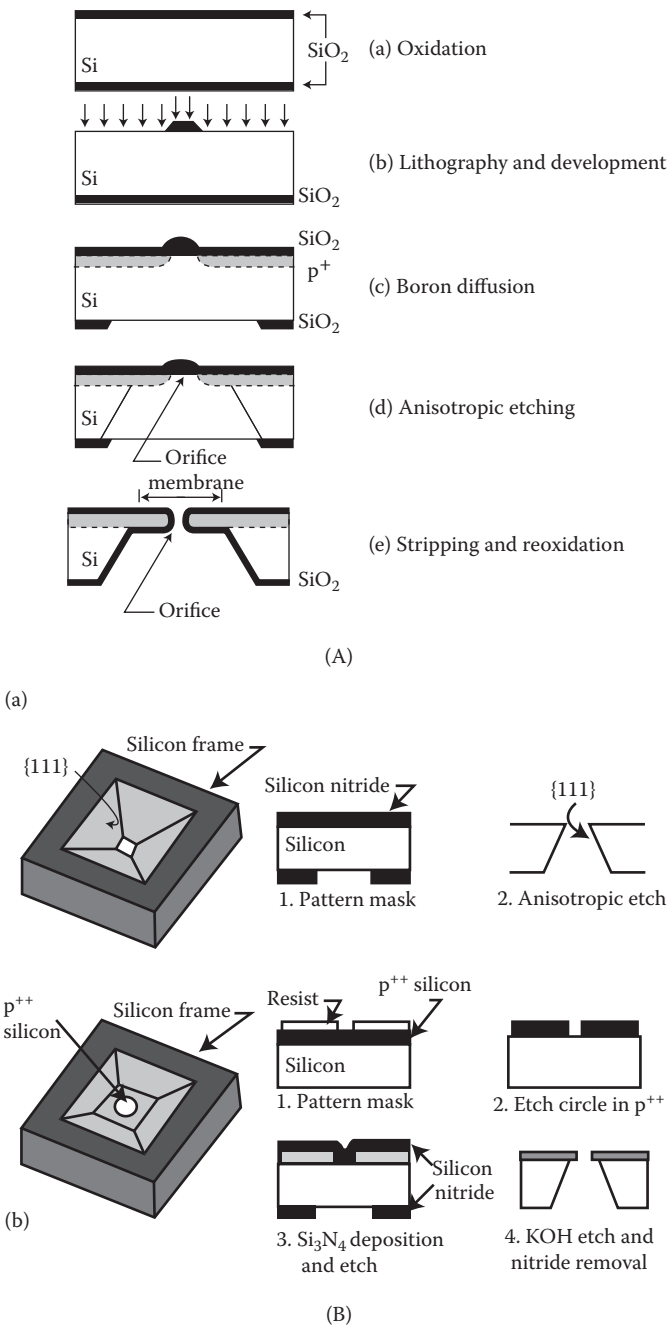


FIGURE 4.45 (A) Illustration of the boron etch stop in the fabrication of a membrane nozzle. (From Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press. With permission.[152]) (B) Two alternate methods to fabricate nozzles. (a) Etch stop is based on silicon nitride. (b) Etch stop is based on boron etch stop.

lightly doped p-type Si (Figure 4.45Ad). Stripping and reoxidation produce an orifice in the suspended membrane (Figure 4.45Ae). Two alternative ways of fabricating nozzles—one square and one circular—are illustrated in Figure 4.45B. The method shown in Figure 4.45Ba uses silicon nitride as an etch-stop layer to etch a square nozzle. The side of the backside opening in the silicon nitride must be larger than 71% of the wafer thickness to etch all the way through the wafer (Volume I, Equation 4.2, $W_0 = W_m - 2z\cot 54.74^\circ = W_m - \sqrt{2}z$). The approach shown in Figure 4.45Bb is again based on a p^{++} etch-stop layer. The difference with the approach in Figure 4.45A is that a uniform p^{++} doping profile is first established here, and that layer is subsequently etched into a circular pattern. Layers of p^{++} silicon with a thickness of 1–20 μm can easily be fabricated, and the boron etch stop is very effective; it is not critical when the operator takes the wafer out of the etchant. One important practical note is that the boron etch stop may become badly degraded in EDP solutions that were allowed to react with atmospheric oxygen. Because boron atoms are smaller than silicon, a highly doped, freely suspended membrane or diaphragm will be stretched; the boron-doped silicon is typically in tensile stress, and the microstructures are flat and do not buckle. Whereas doping with boron decreases the lattice constant, doping with germanium increases the lattice constant. A membrane doped with B and e still etches much more slowly than undoped silicon, and the stress in the layer is reduced. A stress-free, dislocation-free, and slow etching layer (±10 nm/min) is obtained at doping levels of 10^{20} cm^{-3} boron and 10^{21} cm^{-3} germanium.[110,153]

One major disadvantage of the boron etch-stop technique is that the extremely high boron concentrations are not compatible with standard CMOS or bipolar techniques, so they can only be used for microstructures without integrated electronics. Another limitation of this process is the fixed number and angles of (111) planes one can accommodate. The etch stop is less effective in KOH compared with EDP. Besides boron, other impurities have been tried for use in an etch stop in anisotropic etchants. Doping Si with germanium has hardly any influence on the etch rate of either the KOH or EDP solutions. At a doping level as high as 5×10^{21} cm^{-3}, the etch rate is barely reduced by a factor of two.[110]

By burying the highly doped boron layer under an epitaxial layer of lighter doped Si, the problem of incompatibility with active circuitry can be avoided; ±1% thickness uniformity is possible with modern epilayer deposition equipment. A

widely used method of automatically measuring the epi-thickness is with IR instruments, especially Fourier transform infrared (see also epitaxy in Chapter 7).[154]

Electrochemical Etch Stop

For the fabrication of piezoresistive pressure sensors, the doping concentration of the piezoresistor must be kept smaller than 1×10^{19} cm^{-3} because the piezoresistive coefficients decrease considerably above this value, and reverse breakdown becomes an issue. Moreover, high boron levels compromise the quality of the crystal by introducing slip planes and tensile stress and prevent the incorporation of integrated electronics. As a result, a boron stop often cannot be used to produce well-controlled thin membranes unless, as suggested above, the highly doped boron layer is buried underneath a lighter-doped Si epilayer. Alternatively, a second etch-stop method, an electrochemical technique, can be used. In this case, a lightly doped p-n junction is used as an etch stop by applying a bias between the wafer and a counter electrode in the etchant, a technique first proposed by Waggener in 1970.[16] Other early work on electrochemical etch stops with anisotropic etchants such as KOH and EDP was performed by Palik et al.,[155] Jackson et al.,[156] Faust and Palik,[157] and Kim and Wise.[158] In electrochemical anisotropic etching, a p-n junction is made, for example, by the epitaxial growth of an n-type layer (phosphorus-doped, 10^{15} cm^{-3}) on a p-type substrate (boron-doped, 30 Ω-cm). This p-n junction forms a large diode over the whole wafer. The wafer is usually mounted on an inert substrate, such as a sapphire plate, with an acid-resistant wax and is partly or wholly immersed in the solution. An ohmic contact to the n-type epilayer is connected to one pole of a voltage source, and the other pole of the voltage source is connected via a current meter to a counter electrode in the etching solution (see Figure 4.46a). In this arrangement, the p-type substrate can be selectively etched away, and the etching stops at the p-n junction, leaving a membrane with a thickness solely defined by the thickness of the epilayer. The incorporation of a third electrode (a reference electrode) in the three-terminal method, together with the use of a potentiostat depicted in Figure 4.46a, allows for a more precise determination of the silicon potential with respect to the solution than a two-terminal setup as we illustrated in Figure 4.36a.

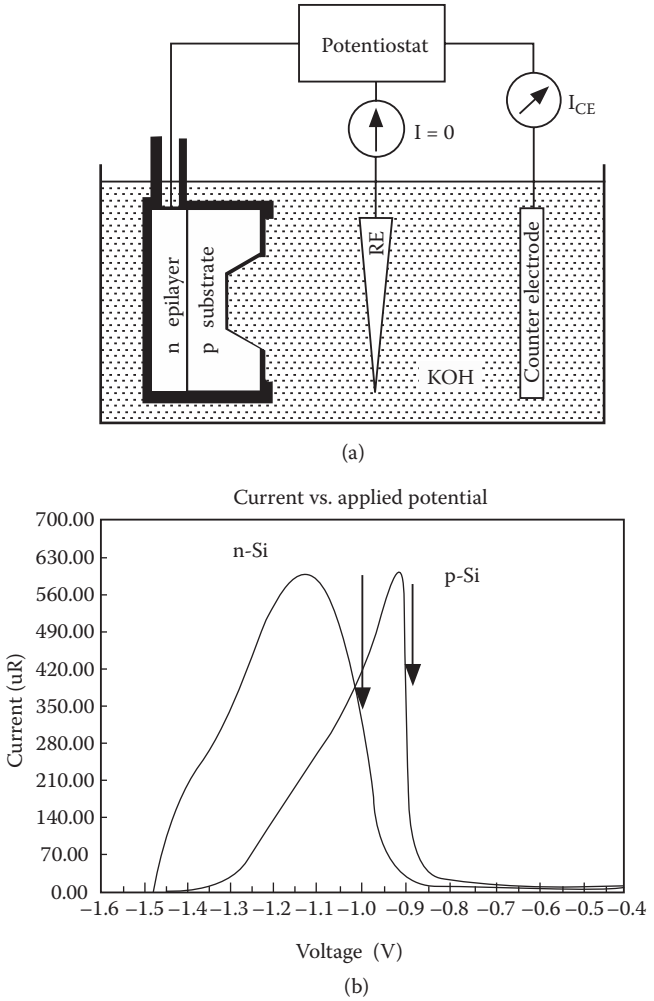


FIGURE 4.46 Electrochemical etch stop. (a) Electrochemical etching setup with potentiostatic control (three-electrode system). Potentiostatic control, mainly used in research studies, enables better control of the potential as it is referenced now to a reference electrode such as an SCE. In industrial settings, electrochemical etching is often carried out in a simpler two-electrode system, that is, a Pt counter electrode and Si working electrode. (From Kloeck, B., S. D. Collins, N. F. de Rooij, and R. L. Smith. 1989. Study of electrochemical etch-stop for high-precision thickness control of silicon membranes. *IEEE Trans Electron Devices* 36:663–69.[159]) (b) Cyclic voltammograms of n- and p-type silicon in an alkaline solution at 60°C. Flade potentials are indicated with an arrow.

At the Flade potential, marked with an arrow in Figure 4.46b, the oxide growth rate equals the oxide etch rate; a further increase of the potential results in a steep fall of the current as a result of complete passivation of the silicon surface. At potentials positive of the Flade potential, all etching stops. At potentials

negative of the Flade potential, the current increases as the potential becomes more positive. This can be explained by the formation of an oxide etching faster than it forms; that is, the silicon is etched away. Whereas electrochemists like to talk about the Flade potential, physicists like to discuss matters in terms of the flat-band potential. The flat-band potential is the applied potential at which there are no more fields within the semiconductor; that is, the energy band diagram is flat throughout the semiconductor (see Volume I, Chapter 4). The passivating SiO_2 layer is assumed to start growing as soon as the negative surface charge on the silicon electrode is cancelled by the externally applied positive bias, a bias corresponding to the flat-band potential. At these potentials, the formation of $Si(OH)_x$ complexes does not lead to further dissolution of silicon because two neighboring Si-OH HO-Si groups will react by splitting off water, leading to the formation of Si-O-Si bonds. As can be learned from Figure 4.46b, the value of the Flade potential depends on the dopant type. Consequently, if a wafer with both n- and p-type regions exposed to the electrolyte is held at a certain potential in the passive range for n-type and the active range for p-type Si, the p-type regions are etched away, whereas the n-type regions are retained. In the case of the diode shown in Figure 4.46a, where, at the start of the experiment, only p-type Si is exposed to the electrolyte, a positive bias is applied to the n-type epilayer (V_n). This reverse biases the diode, and only a reverse bias current can flow. The potential of the p-type layer in this regime is negative to the flat-band potential, and active dissolution takes place. At the moment that the p-n junction is reached, a large anodic current can flow, and the applied positive potential passivates the n-type epilayer. Etching continues on the areas where the wafer is thicker until the membrane is reached there, too. The thickness of the Si membrane is thereby solely defined by the thickness of the epilayer; neither the etch uniformity nor the wafer taper will influence the result. A uniformity of better than 1% can be obtained on a 10-µm-thick membrane. The current versus time curve can be used to monitor the etching process; at first, the current is relatively low, limited by the reverse bias current of the diode. Then, as the p-n junction is reached, a larger anodic current can flow until all the p-type material is consumed and the current decreases again to a low plateau value. The plateau indicates that a current associated with a passivated n-type Si electrode has been reached. The etch procedure can be stopped at the moment that the current plateau has been reached, and because the etch stop is basically one of anodic passivation, it is sometimes called an *anodic oxidation* etch stop. Registering an *I* versus *V* curve as in Figure 4.46b will establish an upper limit on the applied voltage, *V*,[160] and such curves are used for in situ monitoring and controlling of the etch stop. A crude endpoint monitoring can be accomplished by the visual observation of cessation of hydrogen bubble formation accompanying silicon etching. Palik et al.[161] presented a detailed characterization of Si membranes made by electrochemical etch stop.

Using the same anodic oxidation etch stop in a hydrazine-water solution at 90 ± 5°C and a simple two-terminal electrochemical cell, Hirata et al.[162] obtained a pressure sensitivity variation of less than 20% from wafer to wafer (pressure sensitivity is inversely proportional to the square of membrane thickness; see Volume III, Chapter 8). A great advantage of this etch-stop technique is that it works with Si at low doping levels of the order of 10^{16} cm^{-3}. Because of the low doping levels, it is possible to fabricate structures with a very low, or controllable, intrinsic stress. Moreover, active electronics and piezomembranes can be built into the Si without problems. A disadvantage is that the back of the wafer with the aluminum ohmic contact has to be sealed hermetically from the etchant solution, which requires complex fixturing and manual wafer handling. The fabrication of a suitable etch holder is no trivial matter. The holder 1) must protect the epicontact from the etchant, 2) must provide a low-resistance ohmic contact to the epilayer, and 3) must not introduce stress into wafers during etching.[53,126] Stress introduced by etch holders easily leads to diaphragm or wafer fracture and etchant seepage through to the episide.

Using a four-electrode electrochemical cell, controlling the potentials of the epitaxial layer and the silicon wafer separately, as shown in Figure 4.47, can further improve the thickness control of the resulting membrane by directly controlling the p/n bias

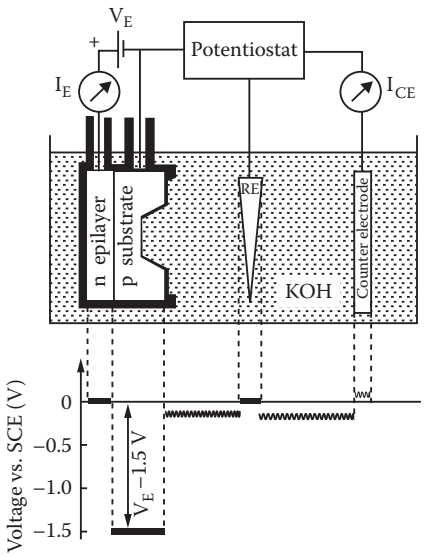


FIGURE 4.47 Four-electrode electrochemical etch-stop configuration. Voltage distribution with respect to the SCE reference electrode (RE) for the four-electrode case. The fourth electrode enables an external potential to be applied between the epitaxial layer and the substrate, thus maintaining the substrate at etching potentials. (From Kloeck, B., S. D. Collins, N. F. de Rooij, and R. L. Smith. 1989. Study of electrochemical etch-stop for high-precision thickness control of silicon membranes. *IEEE Trans Electron Devices* 36:663–69. With permission.[159])

voltage. The potential required to passivate n-type Si can be measured using the three-electrode system in Figure 4.46a, but this system does not take into account the diode leakage. If the reverse leakage is too large, the potential of the n-type Si, V_n, will approach the potential of the p-type Si, V_p. If there is a large amount of reverse diode leakage, the p-type region may passivate before reaching the n-type region, and etching will cease. In the four-electrode configuration, the reverse leakage current is measured separately via a p-type region contact, and the counter electrode current (I_{CE}) may be monitored for endpoint detection. The four-electrode approach allows etch stopping on lower quality epilayers (larger leakage current) and should also enable etch stopping of p-type epi on n-type substrates. Kloeck et al.[159] demonstrated that, using such an electrochemical etch-stop technique and with current monitoring, the sensitivity of pressure sensors fabricated on the same wafer could be controlled to within 4% standard deviation. These authors used a 40% KOH solution at 60 ± 0.5°C. Without the etch stop, the sensitivity from sensor to sensor on one wafer varied by a factor of two.

The etching solution used in electrochemical etching can be either isotropic or anisotropic. An electrochemical etch stop used in isotropic media was discussed above. In this case, HF/H_2O mixtures are used to etch the highly doped regions of p^+p, n^+n, n^+p, and p^+n systems.[87,88,163–165] The rate-determining step in etching with isotropic etchants does not involve reducing water with electrons from the conduction band, as it does in anisotropic etchants, and the etch-stop mechanism, as we will learn below, is very different. In isotropic media, the etch stop is simply a consequence of the fact that higher conductivity leads to higher corrosion currents, and the etching slows down on lower conductivity layers. A major advantage of the KOH electrochemical etch is that it retains the anisotropic characteristics of KOH without needing a heavily doped p^+ layer to stop the etching.[166]

In Figure 4.48 we illustrate the etch-stop techniques discussed so far. Figure 4.48a and b, represents two types of boron-dependent etch stops: diffused boron etch stop and buried layer boron

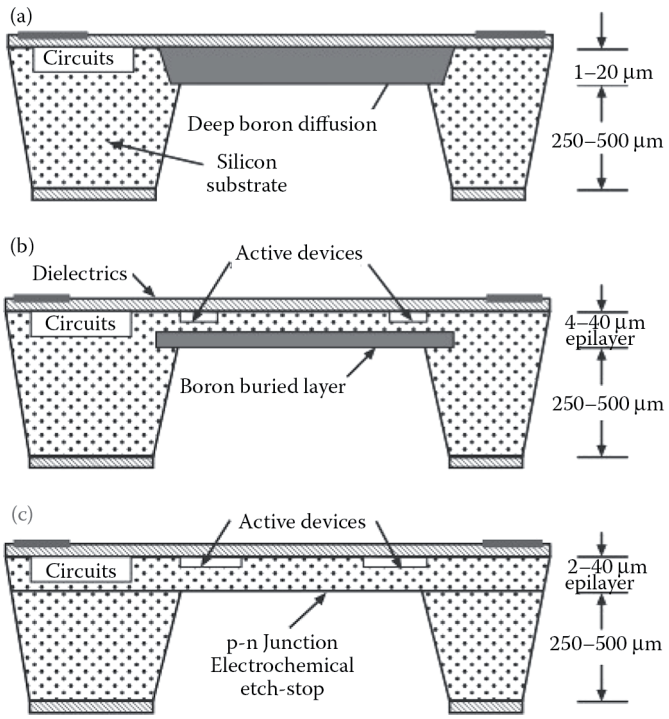


FIGURE 4.48 Typical cross-sections of bulk micromachined wafers with various methods for etch-stop formation shown. (a) Diffused boron etch stop. (b) Boron etch stop as a buried layer. (c) Electrochemical etch stop. (From Wise, K. D. 1985. *Micromachining and micropackaging of transducers.* Eds. C. D. Fung, P. W. Cheung, W. H. Ko, and D. G. Fleming. New York: Elsevier. With permission.[138])

etch stop. In Figure 4.48c we illustrate the electrochemical etch stop.[138] A comparison of the boron etch stop and the electrochemical etch stop reveals that the IC compatibility and the absence of built-in stresses, both results of the low dopant concentration, are the main assets of the electrochemical etch stop.

Photoassisted Electrochemical Etch Stop (for n-Type Silicon)

A variation on the electrochemical diode etch-stop technique is the photoassisted electrochemical etch-stop method illustrated in Figure 4.49.[168] An n-type silicon region on a wafer may be selectively etched in an HF solution by illuminating and applying a reverse bias across a p-n junction, driving the p-type layer cathodic and the n-type layer anodic. Etch rates up to 10 μm/min for the n-type material and a high-resolution etch stop render this an attractive potential micromachining process. Advantages also include the use of lightly doped n-type Si, bias- and illumination intensity-controlled etch rates, in situ process monitoring using the cell current, and the ability to spatially control etching with optical masking or laser writing. Using this method, Mlcak et al.[167] prepared stress-free cantilever beam test samples. They diffused boron into a 10^{15} cm^{-3} (100) n-type Si substrate through a patterned oxide mask, leaving exposed a small n-type region that defines two p-type cantilever beams (see Figure 4.49).

The boron diffusion resulted in a junction 3.3 μm underneath the surface. An ohmic contact on the back of the wafer was used to apply a variable voltage across the p-n junction, and both p and n areas were exposed to the HF electrolyte. The exposed n-type region was etched to a depth of 150 μm by shining light on the whole sample. The resulting n-type Si surface was at first found to be rough because porous Si up to 5 μm in height forms readily in HF solutions. The Si surface could be made smoother by etching at higher bias [4.3 V vs. saturated calomel electrode (SCE)] and higher light intensity (2 W/cm^2) to a finish with features of the order of 0.4 μm in height. Smoothing could also be accomplished by a 5-second dip in HNO_3:HF:H_3COOH or a 30-second dip in 25 wt%, 25°C, KOH. Yet another way of removing unwanted porous Si is a 1000°C wet oxidation to make oxidized porous Si (OPS) followed by an HF etch.[168] The formation of porous Si and OPS is detailed further below.

The photoassisted electrochemical etching of n-type Si is commercially exploited at Boston Microsystems Inc. (http://www.bostonmicrosystems.com).

Photoinduced Preferential Anodization (for p-Type Si)

Electrochemical etching requires the application of a metal electrode to apply the bias. The application of such a metal electrode often induces contamination

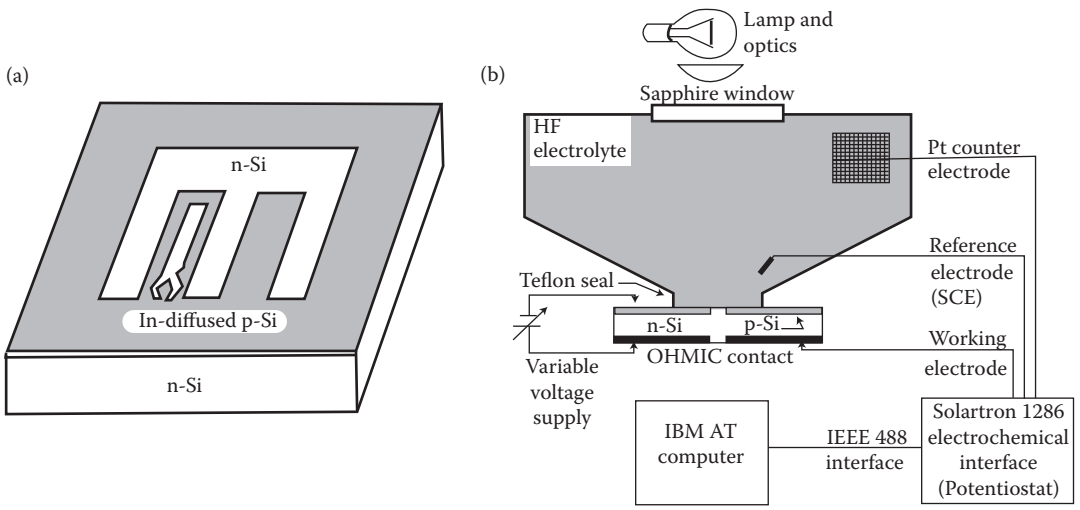


FIGURE 4.49 Photoelectrochemical etching. (a) Schematic of the photoelectrochemical etching experiment apparatus. (b) Schematic of the spatial geometry of the diffused p-type silicon layer into n-type Si used to form cantilever beam structures. (From Mlcak, R., H. L. Tuller, P. Greiff, and J. Sohn. *Proceedings: IEEE micro electro mechanical systems (MEMS '93).* Fort Lauderdale, FL: IEEE. With permission.[167])

and constitutes at least one extra process step, and extra fixturing is needed. With photoinduced preferential anodization (PIPA), it is not necessary to deposit metal electrodes. Here, one relies on the illumination of a p-n junction to bias the p-type Si anodically, and the p-type Si converts automatically into porous Si, whereas the n-type Si acts as a cathode for the reaction. The principle of photobiasing for etching purposes was known and patented by Shockley as far back as 1963.[169] In US patent 3,096,262, he writes that "light can be used in place of electrical connections … for biasing of the sample … This means a small isolated area of p-type material on an n-type body may be preferentially biased for removal of material beyond the junction by etching." The method was "reinvented" by Yoshida et al.[168] in 1992 and in 1993 by Peeters et al.[170,171] The latter group called the method PHET, for photovoltaic electrochemical etch-stop technique, and the former group coined the PIPA acronym.

In PIPA, etch rates of up to 5 µm/min result in the formation of porous layers that are readily removed with Si etching solutions. An important advantage of the technique is that very small and isolated p-type islands can be anodized at the same time. The method also lends itself to fabrication of 3D structures using p-type Si as sacrificial layers.[168] A disadvantage of the technology is that one cannot control the process very well because the current cannot be measured for endpoint detection. Application of PIPA to form a microbridge is shown in Figure 4.50. First, a buried p-type layer doped to 10^{18} cm^{-3} and an n-type layer doped to 10^{15} cm^{-3} are formed on an n-type substrate using epitaxy (Figure 4.50a). Then, the p-type layer is preferentially anodized in 10% HF solution under 30 mW/cm^2 light intensity for 180 minutes (Figure 4.50b), forming porous Si. The porous Si is then oxidized in wet oxygen at 1000°C (Figure 4.50c). Finally, the sacrificial layer of oxidized porous Si (OPS) is etched and removed with an HF solution (Figure 4.50d). The resulting surfaces of the n-type silicon are very smooth. It is interesting to consider making complicated 3D structures by going immediately to the electropolishing regime instead. Yoshida et al.[168] believe that porous Si as a sacrificial intermediate is more suitable for fabricating complicated structures.

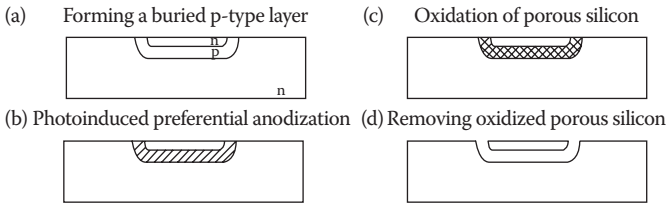


FIGURE 4.50 Photoassisted electrochemical etch stop (for n-type Si). Fabrication process for a microbridge and SEM picture of Si structure before and after PIPA. (From Yoshida, T., T. Kudo, and K. Ikeda. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92).* Travemunde, Germany: IEEE. With permission.[168])

The authors are probably referring to the fact that electropolishing is much more aggressive and could not be expected to lead to the same retention of the shape of the buried, sacrificial p-type layers. Peeters et al.[160] carry out their photovoltaic etching in KOH, thus skipping the porous Si stage of Yoshida et al. They, like Yosida et al., stress the fact that, in one single etch step, this technique can make a variety of complex shapes that would be impossible with electrochemical etching techniques. These authors found it necessary to coat the n-type Si part of the wafer with Pt to get enough photovoltaic drive for the anodic dissolution; this metallization step makes the process more akin to photoelectrochemical etching (PEC), and some advantages of the photobiasing process are lost.

Etch Stop at Thin Insoluble Films

Yet another distinct way (the fourth) to stop etching is by using a change in composition of material. An example is an etch stopped at a Si_3N_4 diaphragm (see Figure 4.45Ba for an example). Silicon nitride is very strong, hard, and chemically inert, and the stress in the film can be controlled by changing the Si/N ratio in the LPCVD deposition process. The stress turns from tensile in stoichiometric films to compressive in silicon-rich films (for details, see Volume III, Chapter 5). A great number of materials are not attacked by anisotropic etchants. Hence a thin film of such a material can be used as an etch stop.

Another example is the SiO_2 layer in a silicon-on-insulator (SOI) structure. A buried layer of SiO_2, sandwiched between two layers of crystalline silicon, forms an excellent etch stop because of the

good selectivity of many etchants of Si over SiO_2. The oxide does not exhibit the good mechanical properties of silicon nitride and consequently is rarely used as a mechanical member in a microdevice. As with PIPA, no metal contacts are needed with an SOI etch stop, greatly simplifying the process over an electrochemical etch-stop technique.

We classified SOI micromachining under surface micromachining in Chapter 7, where more details about this promising micromachining alternative are presented.

Chemical Etching Models for Single-Crystal Si

Introduction

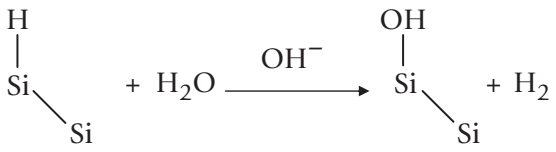
Conflicting data abound in the literature for the anisotropic etch rates of the different Si planes, especially for the higher index planes. This is not too surprising, given the multiple parameters influencing individual results: temperature, stirring, size of etching feature (i.e., crevice effects and loading effects), etchant concentration, addition of alcohols and other organics, surface defects, complexing agents, surfactants, pH, cation influence, and so on. Even the reason(s) why acidic media lead to isotropic etching whereas alkaline media lead to anisotropic etching was, until recently, not explicitly addressed in any of the models surveyed. More rigorous experimentation and standardization will be needed, as well as improved etching models, to better understand the influence of all these parameters on etch rates. In the following section we will present recent atomistic and fluctuating energy level models (see Volume I, Chapter 4) to explain anisotropic and isotropic etching of silicon. It is our hope that the reading of this section will inspire more detailed semiconductor electrochemistry work on Si electrodes. The refining of an etching model will be of invaluable help in writing more predictive Si etching software code.

Atomistic Model—Gosalvez-Kelly

The origin of the macroscopic anisotropy in the etching of Si lies in the site-specificity of the etch rates at the atomistic scale. Surface sites on different crystallographic planes react with an etchant in a different manner because their surface orientations are different. As a consequence, as pointed out by Gosalvez, the macroscopic etch rate is different for different crystal orientations.[47] The removal of Si surface atoms is a complex process that involves chemical and electrochemical reactions. As shown in Figure 4.51, the etching process takes place through sequential oxidation and etching reactions. During etching of Si, the surface, as we show here, is predominantly H-terminated, and this is true for both isotropic[94] and anisotropic etching.[172] The H-termination of the surface was established from both ex situ (after etching the surface) and in situ (during etching) infrared spectroscopy by Kelly et al.[172] In other words, removal of the native oxide from Si leaves the surface hydrogen-terminated.

Once the substitution of H (hydrophobic) by OH (hydrophilic) has taken place, a fast sequence of chemical steps (represented as "etching" in Figure 4.51) leads to the removal of the silicon atom as a $Si(OH)_4$ product. The triggering effect of the reaction by the OH ligand is attributed to the difference in electronegativities between the Si and O atoms, resulting in the polarization and weakening of the backbonds of the Si atom, which as a result, becomes more vulnerable to further attack. Although the hydroxyl group plays an important role in catalyzing the removal of the surface atom by favoring backbond attack, the real active species attacking the backbonds are polar molecules: water in alkaline etchants and HF in acidic etchants.[172]

In an alkaline medium, the first step involving the nucleophilic attack by an OH^- ion on the Si-H bond may be represented as[172]:



Reaction 4.10

The departing hydride reacts with water to give hydrogen gas, and the Si surface atom is hydroxylated. It is the presence of the OH group that polarizes the Si-Si backbond. With the OH adding to the positively polarized surface atom, the backbond is

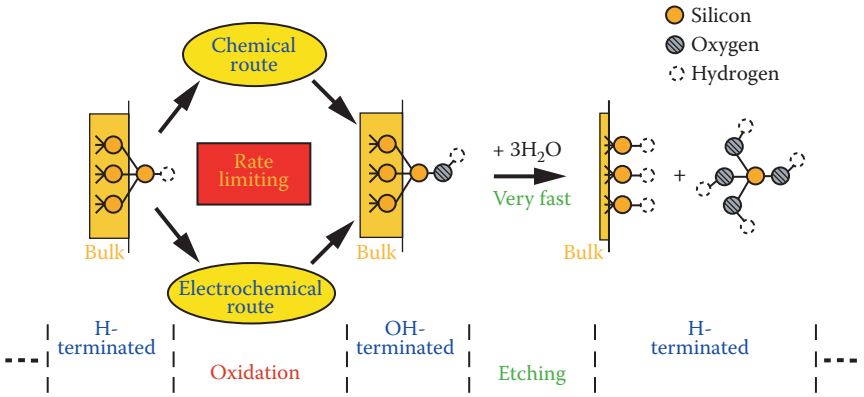


FIGURE 4.51 Schematic representation of the oxidation and etching reactions that sequentially occur for the removal of Si surface atoms. The rate limiting process (oxidation) can take place through chemical and electrochemical routes. Only the nearest underlying bulk atoms in the neighborhood of the surface site are depicted. (Gonsalvez, M. 2003. *Atomistic modelling of anisotropic etching of crystalline silicon.* PhD diss., Helsinki University of Technology.[47])

attacked by water, and an H atom attaches itself to the subsurface atom, restoring the hydrogen termination of the Si surface as shown in Reaction 4.11[172]:

Reaction 4.11

Whether the surface is hydrogenated (hydrophobic) or hydroxylated (hydrophilic) during etching depends on the relative rates of Reactions 4.10 and 4.11. The fact that in practice a predominantly H-termination is found is determined by kinetics rather than thermodynamics, and this implies that the site specificity of the whole process is a result of the rate-limiting oxidation reaction alone.[47]

The number of backbonds associated with the attacked Si atom is universally accepted as a major source for surface specificity. In the case of a (100) surface, each atom has a dihydride termination and is bonded to two subsurface Si atoms. In the (111) case, the surface is monohydride terminated with three backbonds. A dihydride is chemically less stable than a monohydride, so one expects Reaction 4.10 to be faster for a (100) surface.[173] Because it should be easier to dislodge a (100) Si atom with only two backbonds, the rate of Reaction 4.11 also should be faster than for a (111) surface. This qualitatively explains already why one expects a (111) plane to etch much more slowly than a (100) plane in alkaline media.[172] In the model of Seidel et al.[52] (see below) the energy levels of the electrons in the backbonds are associated with surface states within the bandgap of Si. The energy level of these surface states is assumed to vary for different surface orientations, being lowest for {111} planes.

We return now for a moment to Figure 4.51 and draw the readers' attention to the fact that chemical and electrochemical reaction routes provide two alternative mechanisms for the oxidation of the hydrogen-terminated sites before the removal of the resulting hydroxyl-terminated silicon. For open-circuit etching of Si, i.e., with no external circuit connecting the Si to a counter electrode, there are two possible mechanisms: chemical and electroless. The two are distinguished by the fact that the latter involves the participation of free charge carriers, giving rise to measurable currents and allowing for the possibility of controlling the etching process with a biasing potential.[172] In the electrochemical route of Si oxidation, an oxidizing agent in solution extracts/injects mobile charge carriers from/into the solid. When, for example, an oxidizing agent extracts bonding electrons from the valence band, i.e., injects holes, and the holes stay localized at the surface, they will cause bond rupture. It is important to realize that these holes are mobile so that the two reactions, the reduction of the oxidant and the oxidation of the solid, can be considered as two spatially and temporally independent electrochemical reactions. This form of etching is generally referred to as electroless because it does not involve two separate electrodes

(the same process is involved in metal corrosion). A reaction like this, involving the exchange of charge carriers, can lead to external currents when an external circuit is fashioned and can be influenced by a bias applied between the etching electrode (Si) and a counter electrode (e.g., Pt). In chemical etching, bonds in the etching agent and the solid are broken, and new bonds are formed in a localized and synchronous reaction, not involving free carriers. In this case, the reaction rate cannot be influenced by an applied bias. In the oxidation subprocess of Figure 4.51, the main difference between the chemical and the electrochemical routes lies in the size of the activated complex. In chemical oxidation, we saw (Reaction 4.10) that the hydroxyl ion acts as a catalyst. It is postulated that the hydroxyl ion reduces the energy barrier of a sizable pentavalent transition state as shown schematically in Figure 4.52.[174]

The fact that both water (reactant) and OH^- (catalyst) are involved in the reaction depicted in Figure 4.52 is made clear from the fact that etching of single-crystal silicon is possible in pure boiling water, where there is a small concentration of OH^- present, but not in concentrated HF, where the concentration of OH^- is too low to enable the oxidation subprocess in Figure 4.52. In this case, the reaction stops after the H-termination.[47] Besides the number of backbonds of the site under attack, Gosalvez concludes that the site specificity of anisotropic etching is also determined by the geometry of the transition complex; in other words, the steric hindrance of the OH-termination controls the reaction rate as well.[47] The larger the size and the more rigid the transition complex, the larger one expects the anisotropy to be. As a result of these constraints, the chemical oxidation route is very anisotropic.

The Si etching process has been found to depend on the applied potential; thus, the chemical reaction in Figure 4.52 cannot be the only oxidation subprocess. Under anodic potential, the dissolution rate becomes enhanced, and the ratio of H_2/Si is reduced because of the contribution of the electrochemical reaction path.[105] Therefore, there also must be an electrochemical route involving free charge carriers. We will detail our understanding of the electrochemical route when analyzing the Seidel et al. model (see below); here we just take advantage of the experimental evidence that the anodic current is found to be only weakly dependent on the OH^- concentration but strongly so on the water concentration (see Equation 4.3, page 275),[52,110] suggesting that water is the only active species. This is illustrated in Figure 4.53, where the process is depicted as occuring through the dissociation of the surface hydride group and subsequent reaction with a water molecule.

In contrast with the sizable pentavalent transition state involved in the chemical oxidation route (see Figure 4.52), in the electrochemical oxidation route the small surface hydride leaves plenty of space for the reaction with water molecules, and the dissociation might occur at any surface site, making the electrochemical route rather isotropic compared with the chemical route.[47]

The surface morphology of the etched surfaces also turns out to be orientation-dependent when examined at the micrometer and smaller length scales. Pyramidal hillocks and round shallow pits,

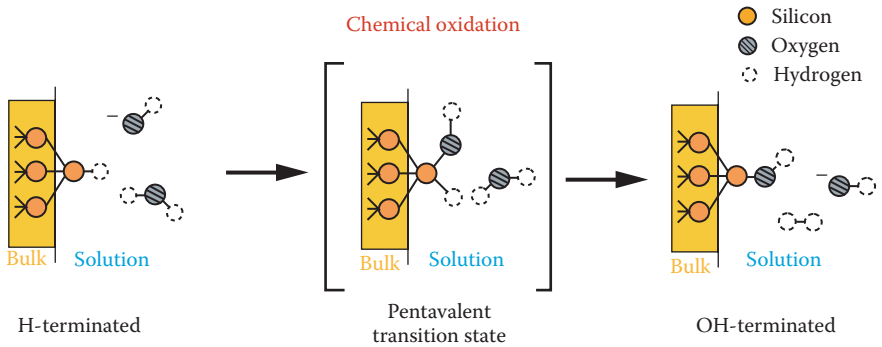


FIGURE 4.52 Schematic representation of the chemical oxidation reaction. Only the nearest underlying bulk atoms in the neighborhood of the surface site are shown. (Gonsalvez, M. 2003. *Atomistic modelling of anisotropic etching of crystalline silicon.* PhD diss., Helsinki University of Technology.[47])

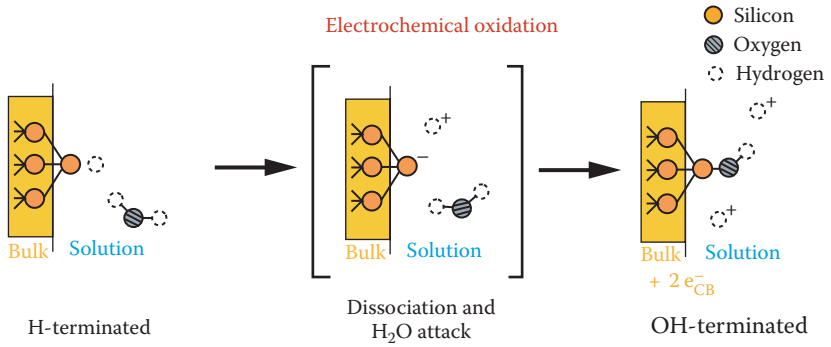


FIGURE 4.53 Schematic representation of the electrochemical oxidation reaction. Only the nearest underlying bulk atoms in the neighborhood of the surface site are shown. (Gonsalvez, M. 2003. *Atomistic modelling of anisotropic etching of crystalline silicon.* PhD diss., Helsinki University of Technology.[47])

for example, are typical for {100} planes, whereas triangular pits are found on {111} surfaces. Gosalvez writes that the macroscopic anisotropy and the surface morphology are two manifestations of the same microscopic site specificity observed at different length scales.[47] The surface microstructure during wet etching of Si then depends on the relative importance of the chemical and electrochemical oxidation routes. The etching process is most anisotropic under cathodic bias, where the etching process is completely chemical, and it becomes increasingly more isotropic as the anodic bias is increased. At sufficiently negative bias one can even eliminate the surface pits on {111} planes, or a totally chemical oxidation process results in perfectly flat {111} surfaces. In Gosalvez's view, under the usual alkaline etching conditions, without biasing potentials, the etching process is very anisotropic, and he claims that the electrochemical oxidation route has only minor effects (mainly the nucleation of pits on {111} planes). Within Gosalvez's elegant model, the etch stop at passivating anodic potentials is a result of an increased OH-termination. The high number of hydroxyl ligands on the Si surface produces the formation of silicon oxide through oxygen insertion into the Si-Si backbonds. The etching mechanism of SiO_2 in alkaline media is slow compared with that of silicon, so the etching process is stopped. This is the same effect observed when increasing the amount of dissolved oxygen in the solution. Very revealing, the effect of oxygen and other oxidizing agents, added to the etching solution, is to make the etching process more isotropic, improving the surface finish of {100} surfaces but negatively affecting that of (111), just in the same manner as the biasing potentials. In both cases, mobile charge carriers make the process more dependent on water molecules only rather than requiring the water/OH^- complex that makes a chemical process more anisotropic.

Seidel et al.'s Model

Anisotropy of Etch Rates Seidel et al.'s model is based on the fluctuating energy level model of the silicon/electrolyte interface and assumes the injection of electrons in the conduction band during the etching process. The construction of energy diagrams of the type discussed here was introduced in Volume I, Chapter 7. Consider the situation of a piece of Si immersed in an alkaline solution without applied bias (open circuit conditions). After immersion of the silicon crystal into the alkaline electrolyte, a negative excess charge builds up on the surface as a result of the higher original Fermi level of the H_2O/OH^- redox couple in the etchant as compared with the Fermi level of the solid. In other words, as explained in Volume I, Chapter 7, the electrochemical potential difference is equalized across the interface; the Fermi levels line up; or also, the work function difference is equalized. The Fermi level in a given material is equivalent to the electrochemical potential of the electrons in that material expressed per particle, or:

$$E_F = \frac{\bar{\mu}_e}{N_A} \tag{7.79}$$

(see Volume I, Chapter 7). This leads to a downward bending of the energy bands on the solid surface for

both p- and n-type silicon (Figure 4.54a and b). The downward bending is more pronounced for p-type than for n-type because of the initially larger difference of the Fermi levels between the solid and the electrolyte.

Next, hydroxyl ions cause the Si surface to oxidize, consuming water and liberating hydrogen in the process. From the measurement of the amount of Si dissolution and H_2 generation, a stoichiometric ratio of $2H_2/1Si$ was found,[108] and thus the following detailed steps, based on Palik,[128] were suggested:

$$Si + 2OH^- \rightarrow Si(OH)_2^{2+} + 2e^-$$

$$Si(OH)_2^{2+} + 2OH^- \rightarrow Si(OH)_4 + 2e^-$$

$$Si(OH)_4 + 4e^- + 4H_2O \rightarrow Si(OH)_6^{2-} + 2H_2$$

Reaction 4.12

Silicate species were observed by Raman spectroscopy.[175] Thus, the overall silicon oxidation reaction consumes four electrons first injected into the conduction band, where they stay near the surface as a result of the downward bending of the energy bands (see Figure 4.54). Raley et al.[176] first presented evidence for injection of four electrons rather than a mixed hole and electron mechanism. These authors could explain the measured etch-rate dependence on hole concentration only by assuming that the proton or water reduction reaction is rate determining and that a four-electron injection mechanism with the conduction band was involved. The injected electrons are highly "reducing" and react with water to form hydroxide ions and hydrogen:

$$4H_2O + 4e^- \rightarrow 4H_2O^- \qquad \text{Reaction 4.13}$$

$$4H_2O^- \rightarrow 4OH^- + 4H^+ + 4e^- + 2H_2 \qquad \text{Reaction 4.14}$$

It is thought that the hydroxide ions in Reaction 4.14, generated directly at the silicon surface, react in the oxidation step. The hydroxide ions from the bulk of the solution may not play a major role because they will be repelled by the negatively charged Si surface, whereas the hydroxide ions formed in situ do not need to overcome this repelling force. This would explain why the etch rates for an EDP solution with an OH^- concentration of 0.034 M are nearly as large as those for KOH solutions with a hundred-fold higher OH^- concentration of 5–10 M.[52] The hydrogen formed in Reaction 4.14 can inhibit the reaction, and surfactants may be added to displace the hydrogen (IBM, US Patent 4,113,551, 1978). Additional support for the involvement of four water molecules (Reaction 4.13) comes from the experimentally observed correlation between the fourth power of the water concentration and the silicon etch rate for concentrated KOH solutions. The Si etch rate, *R*, as a function of KOH concentration at 72°C, was shown in Figure 4.39a and c.[52] The etch rate has a maximum at about 20 wt% KOH of about 55 µm/h. The best fit for this experimentally determined etch rate, for most KOH concentrations, is[52,110]:

$$R = k[H_2O]^4\,[KOH]^{\frac{1}{4}} \qquad (4.2)$$

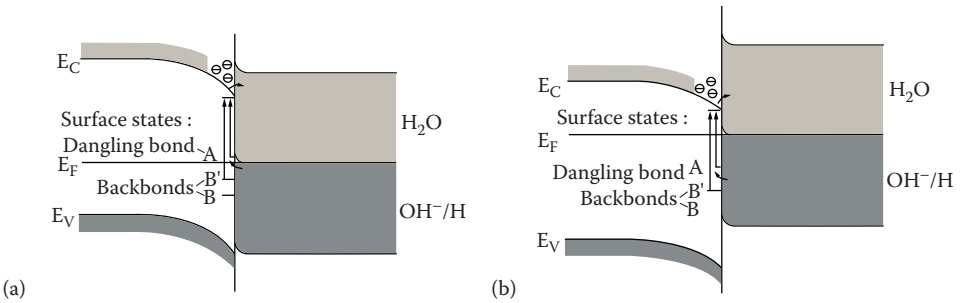


FIGURE 4.54 Band model of the silicon/electrolyte interface for moderately doped silicon (electrolyte at pH > 12): p-type silicon (a) and n-type silicon (b). We assume no applied bias and no illumination. The energy scale functions in respect to the saturated calomel electrode (SCE), an often-used reference in electrochemistry. Notice that p-type Si exhibits more band bending because its Fermi level is lower in the bandgap. For simplicity, we show only one energetic position for surface states associated with dangling bonds and backbonds; in reality, there will be new surface states arising during reactions as the individual dangling bonds and backbonds are taking on different energies as new Si-OH bonds are introduced.

The weak dependence of the etching curve on the KOH concentration (~1/4 power) supports the assumption that the hydroxide ions involved in the oxidation reactions are mostly generated from water. A strong influence of water on the silicon etch rate was also observed for EDP solutions. In molar water concentrations of up to 60%, a large increase of the etch rate occurs.[108] The driving force for the overall Reaction 4.12 is given by the larger Si-O binding energy of 193 kcal/mol as compared with a Si-Si binding energy of only 78 kcal/mol. The role of cations, K^+, Na^+, and Li^+, and even complicated cations such as $NH_2(CH_2)_2NH_3^+$ can probably be neglected.[52]

The four electrons in Reaction 4.13 are injected into the conduction band in two steps. In the case of {100} planes, there are two dangling bonds per surface atom for the first two of the four hydroxide ions to react with, injecting two electrons into the conduction band in the process. As a consequence of the strong electronegativity of the oxygen atoms, the two bonded hydroxide groups on the silicon atom reduce the strength of the two silicon backbonds (Figure 4.55). With two new hydroxide ions approaching, two more electrons (now stemming from the Si-Si backbonds) are injected into the conduction band, and the silicon-hydroxide complex reacts with the two additional hydroxide ions. Seidel et al.[52] claim that the step of activating the second two electrons from the backbonds into the conduction band is the rate-limiting step, with an associated thermal activation energy of about 0.6 eV for {100} planes.

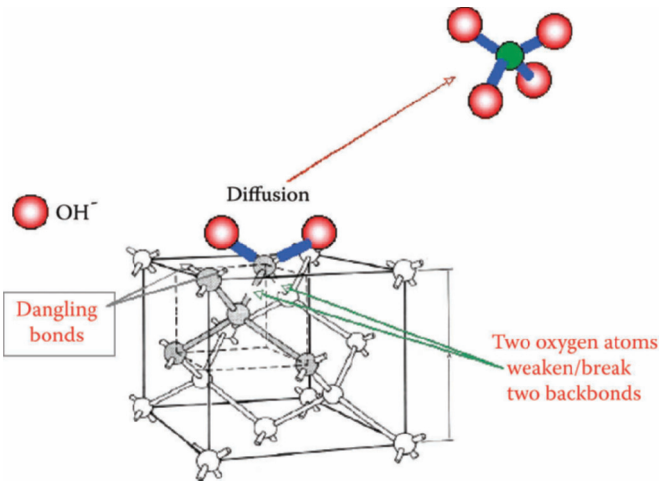


FIGURE 4.55 Two backbonds on Si {100} surfaces.

The electrons in the backbonds are associated with surface states within the bandgap (see Figure 4.54). The energy level of these surface states is assumed to vary for different surface orientations, the lowest being for {111} planes. The thermal activation of the backbonds corresponds to an excitation of the electrons out of these surface states into the conduction band. Because the energy for the backbond surface state level is the lowest within the bandgap for {111} planes, these planes will be hardest to etch. The {111} planes have only one dangling bond for a first hydroxide ion to react with. The second step, which is the rate-limiting step, involves breaking three lower energy backbonds (Figure 4.56).

The lower energy of the backbond surface states within the Si bandgap for {111} Si atoms can be understood from the simple argument that their energy level is raised less by the electronegativity of a single binding hydroxide ion compared with two in the case of the silicon atoms in {100} planes. The high etch rate generally observed on {110} surfaces is similarly explained by a high energy level of the backbond-associated surface states for these planes. Elwenspoek et al.[55,56] do not accept this "two versus three backbonds" argument (see below). They point out that the silicon atoms in the {110} planes also have three backbonds, and activation energy in these crystallographic directions should be comparable with that of {111} planes in contrast to experimental evidence. Seidel et al. would probably counterargue here that the backbonds and the energy levels of the associated surface states are not necessarily the same for {111} and {110} planes because that energy will also be influenced by the effect of the orientation of these bonds. Another argument in favor of the relatively high etching rates of {110} planes is the easier

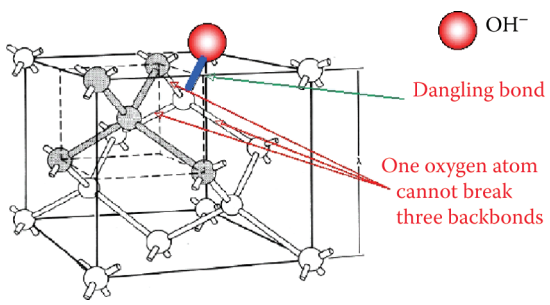


FIGURE 4.56 Three backbonds on Si {111} surfaces.

penetrability of {110} surfaces for water molecules along channels in that plane.

The final step in the anisotropic etching is the removal of the reaction product $Si(OH)_4$ by diffusion. If the production of $Si(OH)_4$ is too fast for solutions with a high water concentration, the $Si(OH)_4$ leads to the formation of a SiO_2-like complex before $Si(OH)_4$ can diffuse away. This might be observed experimentally as a white residue on the wafer surface.[177] The high pH values in anisotropic etching are required to obtain adequate solubility of the $Si(OH)_4$ reaction product and to remove the native oxide from the silicon surface. From silicate chemistry it is known that for pH values greater than 12 the $Si(OH)_4$ complex will undergo the following reaction by the detachment of two protons:

$$Si(OH)_4 \rightarrow SiO_2\,(OH)_2^{2-} + 2H^+ \qquad \text{Reaction 4.15}$$

$$2H^+ + 2OH^- \rightarrow 2H_2O \qquad \text{Reaction 4.16}$$

Pyrocatechol in an ethylenediamine etchant acts as a complex-forming agent for reaction products such as $Si(OH)_4$, converting these products into more complex anions:

$$Si(OH)_4 + 2OH^- + 3C_6H_4(OH)_2 \rightarrow$$

$$Si(C_6H_4O_2)_3^{2-} + 6H_2O \qquad \text{Reaction 4.17}$$

There is evidence by Abu-Zeid et al.[178] of diffusion control contribution to the etch rate in EDP, probably because the hydroxide ion must diffuse through the layer of complex silicon reaction products (see Reaction 4.17). The same authors also found that the etch rate depends on the effective silicon area being exposed and on its geometry (crevice and loading effects). That is why the silicon wafer is placed in a holder, and the solution is vigorously agitated to minimize the diffusion layer thickness. For KOH solutions, no effect of stirring on etching rate was noticed. Stirring here is mainly used to decrease the surface roughness, probably through removal of hydrogen bubbles.

The influence of alcohol on the KOH etching rate in Equation 4.3 is the result of a change in the relative water concentration and its concomitant pH change; it does not participate in the reaction (this was confirmed by Raman studies by Palik et al.[175]). The reversal of etch rates for {110} and {100} planes through alcohol addition to KOH/water etchants can be understood by assuming that the alcohol covers the silicon surface,[175] thus canceling the "channeling advantage" of the {110} planes. In the case of EDP, alcohol has no effect because the water concentration can be freely adjusted without significantly influencing the pH value as a result of the incomplete dissociation of EDP.

For the etching of SiO_2 shown in Figure 4.42, Seidel et al. propose the following reaction:

$$SiO_2 + 2OH^- \rightarrow SiO_2(OH)_2^{2-} \qquad \text{Reaction 4.18}$$

At KOH concentrations up to 35%, a linear correlation occurs between SiO_2 etch rate and KOH concentration. The SiO_2 etch rate in KOH solutions exceeds those in EDP by close to three orders of magnitude. For higher concentrations, the etch rate decreases with the square of the water concentration, indicating that water plays a role in this reaction. Seidel et al. speculate that at high pH values the silicon electrode is highly negatively charged (the point of zero charge of SiO_2 is 2.8), repelling the hydroxide ions while water takes over as reaction partner. An additional reason for the decrease is that the hydroxide concentration does not continue to increase with increasing KOH concentration for very concentrated solutions. The decrease of the Si/SiO_2 etch rate ratio with increasing temperature and pH of the solution (up to 35%) follows out of the larger activation energy associated with the SiO_2 etch rate (0.85 eV) and its linear correlation with the hydroxide concentration, whereas the silicon etch rate mainly depends on the water concentration.

The effect of water concentration and pH value on the etching process in the Seidel et al. model is summarized in Table 4.10.[179]

Doping Dependence of Si Anisotropic Etching A good model for anisotropic etching of Si should also be able to explain the fact that the anisotropic etchant systems of Table 4.9 exhibit drastically reduced etch rates for high boron concentrations in silicon ($\geq 5 \times 10^{19}$ cm^{-3} solid solubility limit). Other

Table 4.10 Effect of Water Concentration and pH Value on the Characteristics of Si Etching

	$-$ H_2O $+$	$-$ pH $+$
SiO_2 etch rate	No effect	$- \Leftrightarrow +$
Si etch rate	$- \Leftrightarrow +$	Little effect
Solubility	No effect	$- \Leftrightarrow +$
Si/SiO_2 ratio	$- \Leftrightarrow +$	$+ \Leftrightarrow -$
Diffusion effects	$- \Leftrightarrow +$	$+ \Leftrightarrow -$
Residue formation	$- \Leftrightarrow +$	$+ \Leftrightarrow -$
p^+ etch stop	$- \Leftrightarrow +$	$+ \Leftrightarrow -$
p, n etch stop	$- \Leftrightarrow +$	$+ \Leftrightarrow -$

Source: Seidel, H. 1990. *Technical digest: IEEE solid-state sensor and actuator workshop.* Hilton Head Island, SC.[179]

impurities (P, Ge) also reduce the etch rate but at much higher concentrations (see Figure 4.57[52]). Boron typically is incorporated using ion implantation (thin layers) or liquid/solid source deposition (thick layers >1 µm). These doped layers are used as very effective etch stop layers (see above). Hydrazine or EDP, which display a smaller (100) to (111) etch rate ratio (~35) than KOH, exhibits a stronger boron concentration dependency. The etch rate in KOH is reduced by a factor of 5–100 for a boron concentration larger than 10^{20} cm^{-3}. When etching in EDP, the factor climbs to 250.[151] With TMAHW solutions, the Si etch rate decreases to 0.01 µm/min for boron concentrations of about 4×10^{20} cm^{-3}.[180]

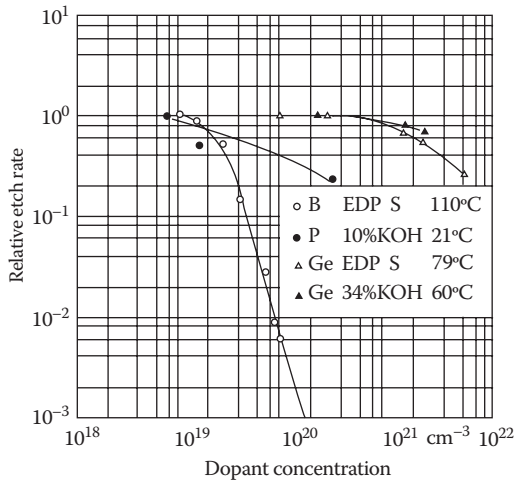


FIGURE 4.57 Relative etch rate for (100) Si in EDP and KOH solutions as a function of concentration of boron, phosphorus, and germanium. (From Seidel, H., L. Csepregi, A. Heuberger, and H. Baumgartel. 1990. Anisotropic etching of crystalline silicon in alkaline solutions–part I: orientation dependence and behavior of passivation layers. *J Electrochem Soc* 137:3612–26. With permission.)

Seidel et al.'s model, we shall see, gives the most plausible explanation for this effect.

Some of the different mechanisms that have been invoked to explain the doping dependence of Si anisotropic etching follow.

1. Several observations suggest that doping leads to a more readily oxidized Si surface. Highly boron- or phosphorus-doped silicon in aqueous KOH spontaneously can form a thin passivating oxide layer.[128,155] The boron oxides and hydroxides initially generated on the silicon surface are not soluble in KOH or EDP etchants.[75] The substitutional boron creates local tensile stress in the silicon, increasing the bond strength so that a passivating oxide might be more readily formed at higher boron concentrations. Boron-doped silicon has a high defect density (slip planes), encouraging oxide growth.
2. Electrons produced during the oxidation of silicon are needed in a subsequent reduction step (hydrogen evolution in Reaction 4.14). When the hole density passes 10^{19} cm^{-3}, these electrons combine with holes instead, thus stopping the reduction process.[155] Seidel et al.'s model follows this explanation (see below).
3. Silicon doped with boron is under tension as the smaller boron atoms enter the lattice substitutionally. The large local tensile stress at high boron concentration makes it energetically more favorable for the excess boron (>5×10^{19} cm^{-3}) to enter interstitial sites. The strong B-Si bonds bind the lattice rigidly. With high enough doping, the high binding energy can stop etching.[75] This hypothesis is similar to item 1, except that no oxide formation is invoked.

The model by Seidel et al. provides an elegant explanation for the etch stop at high boron concentrations. At moderate dopant concentration, we saw that the electrons injected into the conduction band stay localized near the semiconductor surface as a result of the downward bending of the bands (Figure 4.58). The electrons there have a small probability of recombining with holes deeper into the crystal even for p-type Si. This situation changes when the doping level in the silicon increases

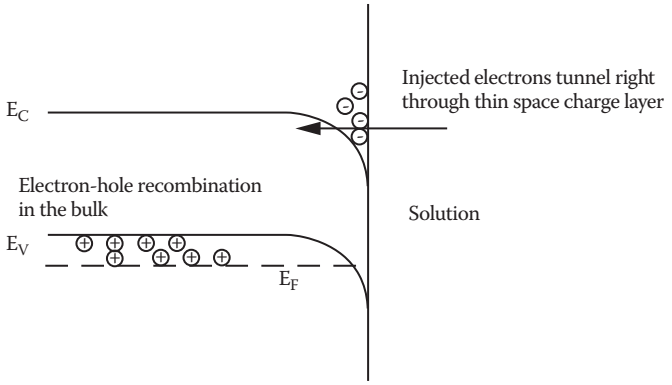


FIGURE 4.58 Si/electrolyte interface energetics at high doping level explaining etch-stop behavior.

further. At a high dopant concentration, silicon degenerates and starts to behave like a metal. For a degenerate p-type semiconductor, the space charge thickness shrinks, and the Fermi level drops into the valence band as indicated in Figure 4.58.

The injected electrons shoot (tunnel) right through the thin surface charge layer into deeper regions of the crystal, where they recombine with holes from the valence band. Consequently, these electrons are not available for the subsequent reaction with water molecules (Reaction 4.13), the reduction of which is necessary for providing new hydroxide ions in close proximity to the negatively charged silicon surface. These hydroxide ions are required for the dissolution of the silicon as $Si(OH)_4$. The remaining etch rate observed within the etch stop region is then determined by the number of electrons still available in the conduction band at the silicon surface. This number is assumed to be inversely proportional to the number of holes and thus the boron concentration. Experiments show that the decrease in etch rate is nearly independent of the crystallographic orientation, and the etch rate is proportional to the inverse fourth power of the boron concentration in all alkaline etchants.

Model of Seidel et al. Summary The key points of the Seidel et al. model can now be summarized as follows (see also Table 4.10)[179]:

1. The rate-limiting step in anisotropic etching of Si is the water reduction.
2. Hydroxide ions required for oxidation of the silicon are generated through reduction of water at the silicon surface. The hydroxide ions in the bulk do not contribute to the etching because they are repelled from the negatively charged surface. This implies that the silicon etch rate will depend on the molar concentration of water and that cations will have little effect on the silicon etch rate.
3. The dissolution of silicon dioxide is assumed to be purely chemical with hydroxide ions. The SiO_2 etch rate depends on the pH of the bulk electrolyte.
4. For boron concentrations in excess of 3×10^{19} cm^{-3}, the silicon becomes degenerate, and the electrons are no longer confined to the surface. This prevents the formation of the hydroxide ions at the surface and thus causes the etching to stop.
5. Anodic biases will prevent the confinement of electrons near the surface as well and lead to etch stop as in the case of a p^+ material.

This model applies well for lower index planes (i.e., {nnn} with $n < 2$), where high etch rates always correspond to low activation energies. However, for higher index planes (i.e., {n11} and {1nn} with n = 2, 3, 4), Herr et al.[54] found no correlation between activation energies and etch rates. For higher index planes, we must rely mainly on empirical data.

Several chemical models explaining the anisotropy in etching rates for the different Si orientations have been proposed. In the different etching models, different Si crystal properties have been correlated with the anisotropy in silicon etching:

1. According to Gandhi[181] and Kern,[77] the silicon oxidation reaction steps involve injection of holes into the Si (raising the oxidation state of Si), hydroxylation of the oxidized Si species, complexation of the silicon reaction products, and dissolution of the reaction products in the etchant solution. In this reaction scheme, etching solutions must provide a source of holes and hydroxide ions, and they must contain a complexing agent that solubilizes reacted Si species in the etchant solution, for example, pyrocatechol forming the soluble $Si(C_6H_4O_2)_3^{2-}$ species. This older model still seems to guide the thinking

of many micromachinists, although Gosalvez's atomistic model and Seidel et al.'s energy level-based model of the silicon/electrolyte interface are more powerful and predictive.

2. Because the {111} Si planes present the highest density of atoms/cm^2 to the etchant and because the atoms are oriented such that three bonds are below the plane (Figure 4.56), it is possible that these bonds become chemically shielded by surface-bonded OH or oxygen, thereby slowing the etch rate for {111} planes. In this model, it is assumed that the anisotropy is caused by differences in activation energies and backbond geometries on different Si surfaces.[182] Seidel et al.[52,110] support this explanation. They detail a process to explain anisotropy based on the difference in energy levels of backbond-associated surface states for different crystal orientations.
3. It also has been suggested that etch rate correlates with available bond density; thus, the surfaces with the highest bond density etch faster.[107] The available bond densities in Si and other diamond structures follow the sequence 1:0.71:0.58 for the {100}/{110}/{111} surfaces. However, Kendall[48] commented that bond density alone is an unlikely explanation because of the magnitude of etching anisotropy (e.g., a factor of 400) compared with the bond density variations of at most a factor of two.
4. Kendall[48] explains the slow etching of {111} planes based on their faster oxidation during etching; this does not happen on the other faces, he speculates, because of greater distance of the atoms on planes other than (111). Because the {111} planes oxidize faster, these planes may be better protected against etching. The oxidation rate in particular in oxygen follows the sequence {111} > {110} > {100}, and the etch rate often follows the reverse sequence (see also Volume I, Chapter 4 on Si oxidation in an oxygen gas stream). In the most used KOH-H_2O wet etching system, however, the sequence is {110} > {100} > {111}.
5. Besides the number of backbonds of the silicon atom under attack, Gosalvez concludes that the site specificity of anisotropic etching is also determined by the geometry of the transition complex; in other words, the steric hindrance of the OH-termination controls the reaction rate as well. The larger the size and the more rigid the transition complex, the larger one expects the anisotropy to be.
6. Madou (see below) explains why Si etches anisotropically in alkaline etchants and isotropically in acidic etchants based on the pH dependency of the energy levels at the silicon/electrolyte interface.
7. Finally, Elwenspoek et al.[55,56] (see below) propose that it is the degree of atomic smoothness of the various surfaces that is responsible for the anisotropy of the etch rates. Basically, this research group argues for the kinetics of smooth faces [the (111) plane is atomically flat] being controlled by a nucleation barrier absent on rough surfaces. The latter, therefore, would etch faster by orders of magnitude. This model also contributes to the understanding why certain etchants etch isotropically and others etch anisotropically.

Elwenspoek et al.'s Model

Elwenspoek et al.[55,56] introduced an alternative model for anisotropic etching of Si, a model built on theories derived from crystal growth. According to these authors, the Seidel et al. model does not clearly explain the fast etching of {110} planes. Those planes, having three backbonds like the {111} planes, should etch equally slowly. The activation energy of the anisotropic etch rate depends on the etching system used; for example, etching in KOH is faster than in EDP, even when the pH of the solution is the same. Seidel et al. attribute this dependence to diffusion that plays a greater role in EDP than in KOH solutions. However, Elwenspoek et al. point out that, at least for slow etching, the etch rate should not be diffusion controlled but governed by surface reactions. With surface reactions, diffusion should have a minor effect, analogous to growth at low pressure in an LPCVD reactor (see Chapter 7). Another comment focuses on the lack of understanding why certain etchants etch isotropically and others etch anisotropically.

Elwenspoek et al. note that crystal surfaces are not ideal and highlight the parallels in the process

of etching and growing of crystals; slowly growing crystal planes also etch slowly! A key to understanding both processes, growing or dissolution (etching), pertains to the concept of the energy associated with the creation of a critical nucleus on a single-crystalline smooth surface, that is, the free energy associated with the creation of an island (growth) or a cavity (etching). Etching or growing of a material starts at active kink sites on steps. Kink sites are atoms with as many bonds to the crystal as to the liquid. Kinetics depends critically on the number of such kink sites (see Figure 4.59). Remember that in Volume I, Chapter 2 we calculated that at room temperature a certain number of vacancies (n_{eq}) must be present in an ideal crystal at any temperature above 0 K (Volume I, Equation 2.59 and Figure 2.35).

A group of vacancies and adatoms can lead to a cavity (etching) or an island (growth), respectively. This aspect has remained neglected in the discussion of etch rates of single crystals until now. The free energy change, ΔG, involved in creating an island or digging a cavity (of circular shape in an isotropic material) of radius r on or in an atomically smooth surface, is given by:

$$\Delta G = -n\Delta\mu + 2\pi r\gamma \tag{4.3}$$

where n is the number of atoms forming the island or the number of atoms removed from the cavity, $\Delta\mu$ is the chemical potential difference between silicon atoms in the solid state and in the solution, and γ is the step free energy. The step free energy in Equation 4.3 will be different at different crystallographic surfaces. This can be easily understood from the following example. A perfectly flat {111} surface in the Si diamond lattice has no kink positions (three backbonds, one dangling bond per atom), whereas on the {001} face every atom has two backbonds and two dangling bonds; that is, every position is a kink position. Consequently, creating an adatom-cavity pair on {111} surfaces costs energy: three bonds must be broken and only one is reformed. In the case of {001} faces, the picture is different. Creating an adatom-cavity pair now costs no energy because two bonds must be broken to remove an atom from the {001} face, but the energy is returned by placing the atom back on the surface. In Equation 4.3 n can be further written out as:

$$n = \pi r^2 h\rho \tag{4.4}$$

where h is the height of the step (down for a cavity and up for an island), r is the diameter of the hole or island, and ρ is the density (atoms/cm^3) of the solid material. The result is:

$$\Delta G = -\pi r^2 h\rho\Delta\mu + 2\pi r\gamma \tag{4.5}$$

where $\Delta\mu$ is counted positive, and γ is positive in any case. In Figure 4.60 we show a plot of ΔG versus r. Equation 4.6 exhibits a maximum at:

$$r^* = \frac{\gamma}{h\rho\Delta\mu} \tag{4.6}$$

At r^*, the free energy is:

$$\Delta G^* = \Delta G(r^*) = \frac{\pi\gamma^2}{h\rho\Delta\mu} \tag{4.7}$$

Consequently, an island or an etched cavity of critical size exists on a smooth face. If by chance a cavity

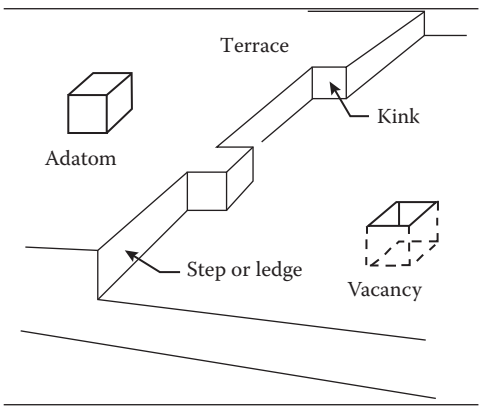


FIGURE 4.59 Perspective drawing of a crystal showing terraces, step or ledge, kinks, adatoms, and vacancies.

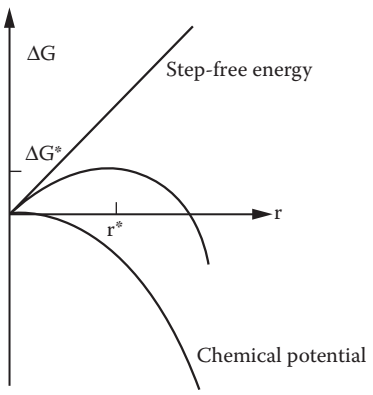


FIGURE 4.60 A plot of ΔG versus r based on Equation 4.5 exhibits a maximum.

is dug into a crystal plane smaller than r^*, it will be filled rather than allowed to grow, and an island that is too small will dissolve rather than continue to grow because that is the easy way to decrease the free energy. With $r = r^*$, islands or cavities do not have any course of action, but in case of $r > r^*$, the islands or cavities can grow until the whole layer is filled or removed. Importantly, if r^* is equal or smaller than the diameter of a single atom, every atom that leaves the surface will lead to a whole layer filled (a terrace) or removed. In the case of crystal growth, the reverse of etching, this phenomenon is called kinetic roughening. For isotropic etching $\Delta\mu$ is very large; therefore, for all crystal orientations r^* is very small. On the other hand, in light of this nucleation barrier theory, to remove atoms directly from flat crystal faces, such as the {111} Si faces, seems very difficult because the created cavities increase the free energy of the system, and filling of adjacent atoms is more probable than removal; in other words, a nucleation barrier has to be overcome. That is, in anisotropic etching, $\Delta\mu$ is small, or, based on Equation 4.7, a sizable nucleation barrier must be overcome. The growth and etch rates, R, of flat faces are proportional to:

$$R \sim e^{-\frac{\Delta G^*}{kT}} \tag{4.8}$$

Because ΔG^* is proportional to γ^2 (see Equation 4.7), the activation energy is different for different crystallographic faces, and both the etch rate and the activation energy are anisotropic. If $\Delta G^*/kT$ is large, the etch rate will be very slow, as is the case for large step free energies. Both $\Delta\mu$ and γ depend on the temperature and type of etchant, and these parameters might provide clues to understanding the variation of etch rate, degree of anisotropy, temperature dependence, and so on, giving this model perhaps more bandwidth than Seidel et al.'s model. According to Elwenspoek et al., the chemical reaction energy barrier and the transport in the liquid are isotropic, and the most prominent anisotropy effect is caused by the step free energy (absent on rough surfaces) rather than the surface free energy. However, the surface free energy and the step free energies are related: when comparing flat faces, those having a large surface free energy have a small step free energy, and vice versa. The most important difference in these two parameters is that the step free energy is zero for a rough surface, whereas the surface energy remains finite.

Flat faces grow and etch with a rate proportional to ΔG^*, which predicts that faces with a large free energy to form a step will grow and etch much slower than faces with a smaller free energy. Elementary analysis indicates that the only smooth face of the diamond lattice is the (111) plane. There may be other flat faces, but with lower activation energies, as a result of reconstruction and/or adsorption, prominent candidates in this category are {100} and {110} planes. On the other hand, a rough crystal face grows and etches with a rate directly proportional to $\Delta\mu$. The temperature at which γ vanishes and a face transitions from smooth to rough is called the *roughening transition temperature* T_R.[183,184] Above T_R, the crystal is rough on the microscopic scale. Rough crystal faces grow and dissolve with a rate proportional to $\Delta\mu$ and therefore proceed faster than flat surfaces. Imperfect crystals, for example, surfaces with screw dislocations, etch even faster with R proportional to $\Delta\mu^2$.

For the state of a surface slightly above or below the roughening temperature, T_R, thermal equilibrium conditions apply. Etching, in most practical cases, is far from equilibrium, and kinetic roughening might occur. Kinetic roughening occurs if the super- or undersaturation of the solution is so large that the thermally created islands or cavities are the size of the critical nucleus.[184] One can show that if the super- or undersaturation is larger than $\Delta\mu_c$, given by:

$$\Delta\mu_c = \frac{\pi f_0 \gamma^2}{kT} \tag{4.9}$$

(f_0 being the area one atom occupies in a given crystal plane), the growth and etch mechanism changes from a nucleation barrier-controlled mechanism to an isotropic growth/etch mechanism. The growth rate and etch rate become proportional to the chemical potential difference ($\Delta\mu$). Thus, it can be expected that if the undersaturation becomes high enough, even the {111} faces could etch isotropically, as they do in acidic etchants. If the undersaturation becomes so large that $\Delta\mu \ll kT$, the nucleation barrier breaks down. Each single-atom cavity acts as a nucleation site made in vast numbers by thermal fluctuations. The face in question

etches with a rate comparable to the etch rate of a rough surface. This situation is called *kinetic roughening*. If all faces are kinetically rough, the etch rate becomes isotropic.

The binding energy ΔE of an atom in a crystal slice with orientation (*hkl*) divided by *kT* (Boltzmann constant × absolute temperature) is known as the α factor of Jackson of that crystal face,[185] or:

$$\alpha = \frac{\Delta E}{kT} \tag{4.10}$$

At sufficiently low temperature, where entropy effects can be ignored, $kT\alpha$ is proportional to the step free energy γ, and the number of adatom-cavity pairs is then proportional to $\exp(-\alpha)$. This number is very small on the {111} silicon faces at low temperature, but 1 on the {001} silicon faces at any temperature. The consequence for {111} and {001} planes is that, at sufficiently low temperatures, the first are atomically smooth and the latter are atomically rough.

Isotropic etching requires conditions of kinetic roughening because the etch rate is no longer dominated by a nucleation barrier but by transport processes in solution and the chemical reaction. To test this aspect of the model, Elwenspoek et al. show that there is a transition from isotropic to anisotropic etching if the undersaturation becomes too small. This can occur if one etches with an acidic etchant very long or if one etches through very small holes in a mask (crevice effect). In both cases, anisotropic behavior becomes evident as aging or limited transport of the solution causes the undersaturation to become very small. No proof is available to indicate that acidic etchants are much more undersaturated than the alkaline etchants. Still, the above explains some phenomena that Seidel et al.'s model fails to address. Another nice confirmation of Elwenspoek et al.'s model is in the effect of misalignment on etch rate. A misalignment of the mask close to smooth faces implies steps; there is no need for nucleation to etch. Because the density of steps is proportional to the angle of misalignment, the etch rate should be proportional to the misalignment angle, provided the distance between steps is not too large. Nucleation of new cavities becomes very probable. This has been observed for the etch rate close to the <111> directions.[52]

Where Elwenspoek et al.'s model becomes a bit murky is in the classification of which surfaces are smooth and which ones are rough. As mentioned above for single-crystal Si, only the {111} planes are smooth at low temperatures. At this stage, the model does not explain anything more than other models; every model has an explanation for the slower {111} etch rate. However, as pointed out, these authors invoke the possibility of surface reconstruction and/or adsorption of surface species, which, by decreasing the surface free energy, could make faces such as {001} and {110} flat as well but with lower activation energies. The authors also take heart in the fact that CVD experiments often end up showing flat {110}, {100}, {331}, and, strongest of all, {111} planes. Si etching models are improving fast, but especially where the influence of the etchant is concerned, more convincing thermodynamic data are needed to estimate $\Delta\mu$ and γ.

Madou Model for Isotropic versus Anisotropic Etching of Silicon

In contrast to alkaline etching, with an acidic etchant such as HF, holes are needed for etching Si. An n-type Si electrode immersed in HF in the dark will not etch because of lack of holes. The same electrode in an alkaline medium etches readily. A p-type electrode in an HF solution, where holes are available under the proper bias, will etch even in the dark. For HF etchants, one might assume that the Ghandi and Kern model,[77,181] relying on the injection of holes, applies. In terms of the band model, this must mean that the silicon/electrolyte interface in acidic solutions exhibits different energetics from the alkaline case. It is not directly obvious why the energetics of the silicon/electrolyte interface would be pH-dependent. To the contrary, because the flat-band potential of most oxide semiconductors and most oxide-covered semiconductors (such as Si) and the Fermi level of the H_2O/OH^- redox couple in an aqueous solution are both expected to change by 59 mV for each pH unit change,[186] one would expect the energetics of the interface to be pH independent. Because the electronegativity in a Si-F bond is higher than for a Si-OH bond, one might even expect the back-bond surface states to be raised higher in an HF medium, making an electron injection mechanism

even more likely than in alkaline media. To clarify this contradiction, we analyze the band model of a Si electrode in an acidic medium in more detail. The band model shown in Figure 4.61 was constructed based on a set of impedance measurements on an n-type Si electrode in a set of aqueous solutions at different pH values. Mott-Schottky plots were constructed from the impedance measurements to determine the pH dependency of the flat-band potential. From that, the position of the conduction band E_{cs}, as well as valence band edges E_{vs}, of the Si electrode in an acidic medium at a fixed pH of 2.2 (the point of zero charge) was calculated at 0.74 eV versus SCE (an often-used reference electrode in electrochemistry) for E_{cs}, and −0.36 eV versus SCE for E_{vs}.[187,188]

We have assumed in Figure 4.61 that the bands are bent upward at open circuit (see below for justification); thus, holes in the valence band are driven to the interface where they can react with Si atoms or with competing reducing agents from the electrolyte. Because we want to etch Si, we are only interested in the reactions where Si itself is consumed. Reactions of holes with reducing agents are of great importance in photoelectrochemical solar cells.[188] For n-type Si, holes can be 1) injected by oxidants from the solution (e.g., by adding nitric acid to the HF solution); 2) supplied at the electrolyte/semiconductor interface by shining light on a properly biased n-type Si wafer; or 3) created by impact ionization, that is, Zener breakdown, of a sufficiently high reverse biased n-type Si electrode.[189] With a p-type Si wafer, a small forward bias supplies all the holes needed for the oxidation of the lattice even without light because the conduction happens via a hole mechanism. An important finding, explaining the different reaction paths in acidic and alkaline media, comes from plotting the flat-band potential as a function of pH. It was found that the band diagram of Si shifts with less than 59 mV per pH unit. Actually, the average shift is only about 30 mV per pH unit.[188] As shown in Figure 4.62, with increasing pH, the energy levels of the solution increase faster than the energy levels in the semiconductor. As a consequence, it is more likely that electron injection takes place in alkaline media as the filled levels associated with the OH^- are closer to the conduction band, whereas in acidic media the filled levels of the redox system overlap better with the valence band, favoring a hole reaction. A lower position of the redox couple with respect to the conduction band edge, E_{cs}, in acidic media explains the upward bending of the bands as drawn in Figure 4.62. With isotropic etching in acidic media, the reaction starts with a hole in the valence band, equivalent to a broken Si-Si bond. In this case, the relative position of backbond-related surface states in different crystal orientations is of no consequence, and all planes etch at the same rate. A study of the interfacial energetics helps to understand why isotropic etching occurs in acidic media and anisotropic etching in alkaline media.

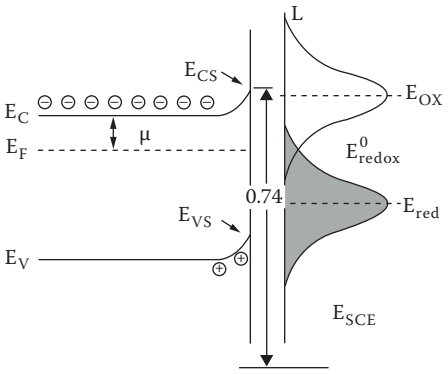


FIGURE 4.61 Band diagram for n-type Si in pH = 2.2 (no bias or illumination). Reference is the SCE. In Figure 4.40 no energy values were given; here we provide positions of the conduction band edge, E_{cs} = 0.74 eV versus SCE, and the valence band edge, E_{vs} = 0.74 eV –1.1 eV = –0.36 eV versus SCE (1.1 eV is the bandgap of Si). These values were determined by means of Mott-Schottky plots. (Madou, M. J., K. W. Frese, and S. R. Morrison. 1980. Photoelectrochemical corrosion of semiconductors for solar cells. *SPIE* 248:88–95.[187]) The separation between the Fermi energy and the bottom of the conduction band is indicated by μ.

A few words of caution regarding our explanation for the reactivity difference between acidic and alkaline media are in order. Little is known about the width of the bell-shaped curves describing the redox levels in solution.[186] Not knowing the surface concentrations of the reactive redox species involved in the etching reactions further hinders a better understanding of the surface energetics because the bell-shaped curves for oxidant and reductant will only be the same height (as drawn in Figures 4.61 and 4.62) if the concentrations of oxidant and reductant are the same. Clearly, the above picture is oversimplified;

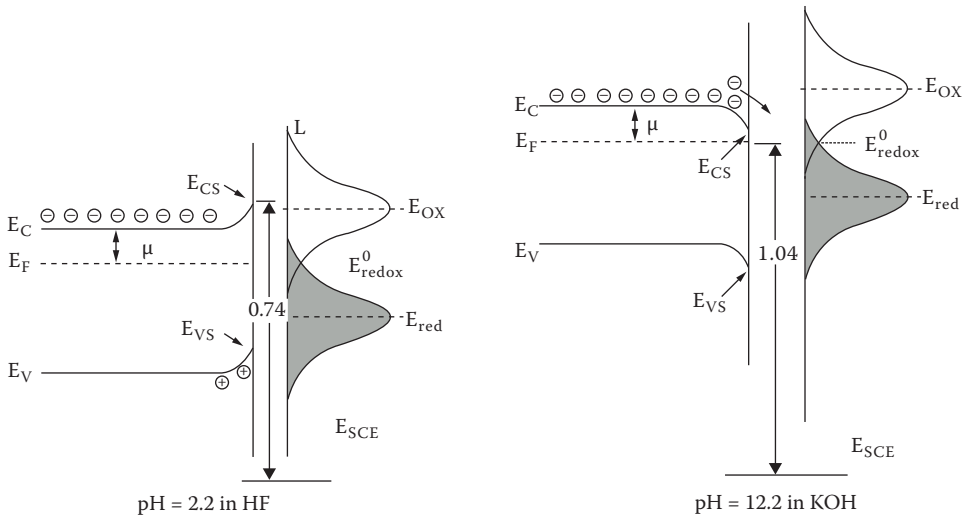


FIGURE 4.62 Band model comparison of the Si/electrolyte interface at low and high pH. Increasing the pH by 10 units shifts the redox levels up by 600 mV, whereas the Si bands only move up by 300 mV. This leads to a different band-bending and a different reaction mechanism, that is, electron injection in the alkaline media (anisotropic) and hole injection in acidic media (isotropic).

several authors have found that the dissolution of Si in HF might involve both the conduction and valence band, a claim confirmed by photocurrent multiplication experiments.[102,190] These photocurrent multiplication experiments showed that one or two holes generated by light in the Si valence band were sufficient to dislodge one Si unit, meaning that the rest of the charges were injected into the conduction band.[191] Our contention here is only that the low pH dependence of the flat-band potential of a silicon electrode makes a conduction band mechanism more favorable with alkaline-type etchants and a hole mechanism more favorable in acidic media.

Continued attempts at modeling the etch rates of all Si planes are underway. For example, Hesketh et al.[192] attempted to model the etch rates of the different planes developing on silicon spheres in etching experiments with KOH and CsOH by calculating the surface free energy. The density of surface bonds on a Si plane is indicative of the surface free energy, which can be estimated by counting the bond density and multiplying by the bond energy. Using the unit cell dimension a of Si of 5.431 Å and a silicon-to-silicon bond energy of 42.2 kcal/mol, the surface free energy, ΔG, can be related to N_B, the bond density, by the following expression:

$$\Delta G = \frac{N_B}{2} \times 2.94\ 10^{-19}\ \frac{J}{m^2} \qquad (4.11)$$

Although Hesketh et al. could not explain the etching differences observed between CsOH and KOH (these authors identified a cation effect on the etch rate!), a plot of the calculated surface free energy versus orientation yielded minima for all low index planes such as {100}, {110}, and {111}, as well as for the high index {522} planes. Fewer bonds per unit area on the low index planes produce a lower surface energy and lower etching rate. When Hesketh et al. added the in-plane bond density to the surface bond density, producing a total bond density, a correlation with the hierarchy of etch rates in CsOH and KOH was found, that is, {311}, {522} > {100} > {111}. The surfaces with the higher bond density etched faster, suggesting that the etch rate might be a function of the number of electrons available at the surface. Hesketh et al. imply that their result falls in line with the Seidel et al.'s model, although it is unclear how the total bond density relates to surface state energies of backbonds. Moreover, Hesketh's model does not take into account the angles of the bonds, and in Elwenspoek's view, the surface free energy does not determine the anisotropy.

More research should focus on the modeling of Si etch rates. The semiconductor electrochemistry of corroding Si electrodes will be a major tool in further developments. Interested readers may consult Sundaram et al.[193] on Si etching in hydrazine and Palik et al.[128] on the etch-stop mechanism in

heavily doped silicon; both explain in some detail the silicon/electrolyte energetics. A more generic treatise on semiconductor electrochemistry can be found in Morrison.[194] A free etch simulator from the University of Illinois can be found at http://mass.micro.uiuc.edu/research/completed/aces/pages/home.html. For more modeling software of Si etching, consult Volume III, Chapter 6.

Etching with Bias and/or Illumination of the Semiconductor

Introduction

The isotropic and anisotropic etching of Si we discussed thus far requires neither a bias nor illumination of the semiconductor. The etching in such cases proceeds at open circuit, and the semiconductor is shielded from light. In a cyclic voltammogram, as shown in Figure 4.63, the operational potential in this case is the rest potential, V_r, where anodic and cathodic currents are equal in magnitude and opposite in sign, resulting in the absence of flow of current in an external circuit. This does not mean that macroscopic changes do not occur at the electrode surface because the anodic and cathodic currents may be part of different chemical reactions just as in the corrosion of a metal exposed to, for example, seawater. Earlier, in the context of Si etching, we called this type of open circuit etching "electroless" etching (see page 266). As an example of "electroless" etching, consider again the isotropic etching of Si in an HF/HNO_3 etchant at open circuit where the local anodic reaction is associated with corrosion of the semiconductor:

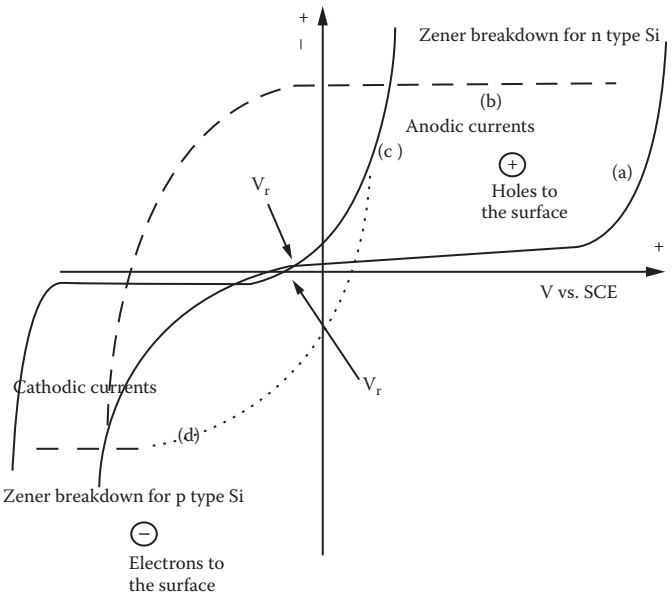


FIGURE 4.63 Basic cyclic voltammograms (*I* vs. *V*) for n- and p-type Si in an HF solution in the absence of a hole-injecting oxidant: n-type Si without illumination (a); n-type Si under illumination (b); p-type Si without illumination (c); p-type Si under illumination (d). If the reactions on the dark Si electrode determining the rest potential V_r for n- and p-type are the same, then V_r is expected to be the same as well. For clarity of the figure we have chosen V_r differently here; in practice, V_r for n- and p-type Si in HF are found to be identical.

$$\text{Local anode: } Si + 2H_2O + nh^+ \rightarrow SiO_2 + 4H^+ + (4\text{-}n)e^- \qquad \text{Reaction 4.19}$$

$$SiO_2 + 6HF \rightarrow H_2SiF_6 + 2H_2O \qquad \text{Reaction 4.20}$$

(the involvement of holes h^+ in the acidic reaction was discussed in the preceding sections), whereas the local cathodic reaction could be associated with reduction of HNO_3:

$$\text{Local cathode: } NO_3^- + 4H^+ + 3e^- \rightarrow NO + 2H_2O \qquad \text{Reaction 4.21}$$

We now explore Si etching while illuminating and/or applying a bias to the silicon sample. To simplify the situation, we first consider the case of an oxidant-free solution so that all the holes must come from within the semiconductor. Anodic dissolution of n-type Si in an HF-containing solution requires a supply of holes to the surface. For an n-type wafer under reverse bias, very few holes will show at the surface unless the high reverse (anodic) bias is sufficient to induce impact ionization or Zener breakdown (see Figure 4.63a). Alternatively, the interface can be illuminated, creating holes in the space charge region (SCR) that the field pushes toward the semiconductor/electrolyte interface (Figure 4.63b). In the forward direction, electrons from the Si conduction band (majority carriers) reduce oxidizing species in the solution (e.g., reduction of protons to hydrogen). A p-type Si sample exhibits high anodic currents even without illumination at small anodic (forward) bias (Figure 4.63c). Here, the current is carried by holes. A p-type electrode illuminated under reverse

bias gives rise to a cathodic photocurrent (Figure 4.63d). At relatively low light intensities, the photocurrent plateaus for both n- and p-types (Figure 4.63b and d respectively) depend linearly on light intensity. The photocurrent is cathodic for p-type Si (species are reduced by photoproduced electrons at the surface, e.g., hydrogen formation) and anodic for n-type Si (species are oxidized by photoproduced holes at the surface; either the lattice itself is consumed or reducing compounds in solution are). In Figure 4.64, we show the cyclic voltammograms of n-type and p-type Si in the presence of a hole-injecting oxidant. The most obvious effect is on the dark p-type silicon electrode. The injection of holes in the valence band increases the cathodic dark current dramatically. The current level measured in this manner for varying oxidant concentration or different oxidants could be used to estimate the efficiency of different isotropic Si etchants, a pointer to the fact that semiconductor electrochemistry has been underused as a tool to study Si etching. When n-type Si is consumed under illumination, we experience photocorrosion (see Figure 4.63b). This photocorrosion phenomenon has been a major barrier to the long-term viability of photoelectrochemical cells.[187]

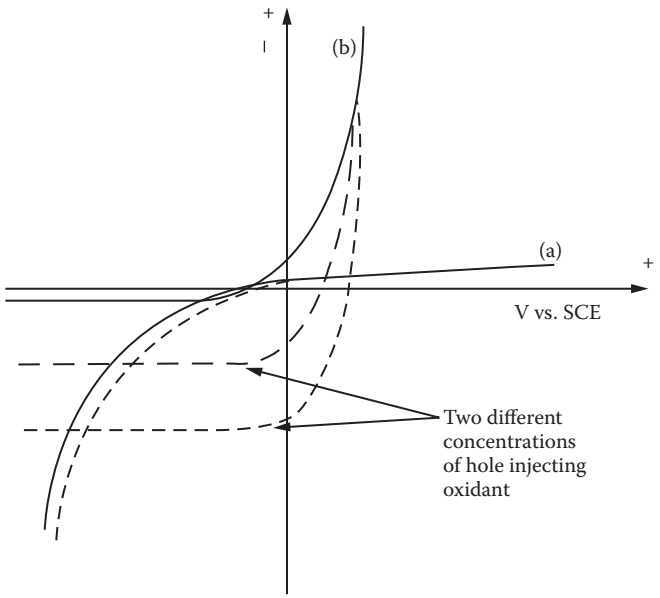


FIGURE 4.64 Basic cyclic voltammograms for n- and p-type Si in an HF solution and in the presence of a hole-injecting oxidant, for example, HNO_3: n-type Si in the dark (a); p-type Si in the dark (b). An increase of the cathodic dark current on the p-type electrodes is most obvious. The current level is proportional to the oxidant concentration.

In what follows, photocorrosion is put to good use for electropolishing and formation of microporous and macroporous layers.[189]

Electropolishing and Porous Silicon

Electropolishing In Figure 4.65 we show cyclic voltammograms for p- and n-type Si in an HF solution in the dark and under illumination. The anodic current in the dark for a p-type Si electrode presents two peaks, characterized by i_{CRIT} and v_p (first peak) and i_{MAX} and v_{MAX} (second peak). This was also illustrated, in somewhat less detail, in Figure 4.37, where we only marked the first peak with i_p and v_p (i_{CRIT} was called i_p in that graph) and did not show the oscillations at potentials positive of v_{MAX}. As can be deduced from this figure, at high light intensities, in so-called photoelectrochemical etching (PEC), the anodic curves for n-type and p-type Si are similar. The two differences are 1) a potential shift of a few hundred millivolts, and 2) the illuminated n-type Si in the anodic range only shows the first peak and then transitions into a photoplateau. At very high illuminations, a second anodic peak also appears in this case. Because of this equivalence, most of the processes described below apply for both forward-biased p-type and for n-type Si under strong illumination. PEC etching involves photocorrosion in

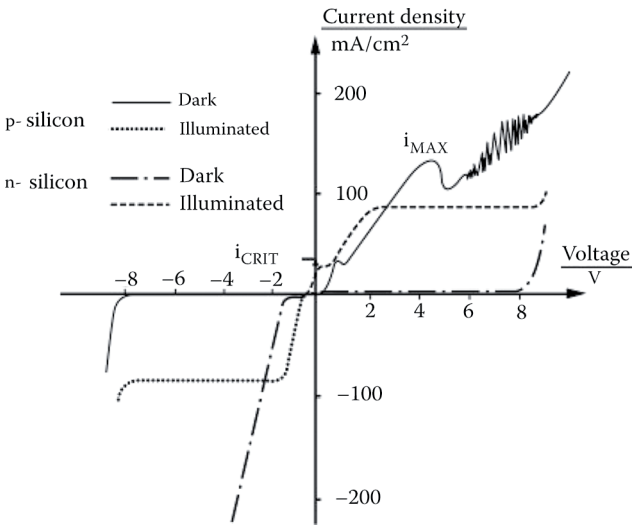


FIGURE 4.65 Cyclic voltammograms identifying porous Si formation regime and electropolishing regime. (From Levy-Clement, C., A. Lagoubi, R. Tenne, and M. Neumann-Spallart. 1992. Photoelectrochemical etching of silicon. *Electrochim Acta* 37:877–88. With permission.[189])

an electrolyte in which the semiconductor is generally chemically stable in the dark; that is, no hole-injecting oxidants are present, as in the case of an HF solution (see also experiments described in Figure 4.63). For carrying out PEC, a setup as shown in Figure 4.66a may be used with a provision to illuminate the semiconductor electrode. The Al strip shown here is used to make ohmic contact to the back side of the Si electrode. For fabricating porous Si and electropolishing, one often uses a so-called electrochemical double cell, also referred to as double tank lateral cell, as shown in Figure 4.66b. Such a cell has two platinum working electrodes, one in each part of the cell, with the Si wafer in the middle separating the two parts. When current is passed through the wafer, its anodic side will be etched. The main advantages compared with the cell in Figure 4.66a are that no back-side metallization is required and that the uniformity of the process is typically better.

The potential range to the left (more positive) of the first current peak (v_p) in Figure 4.65 was clarified earlier as the electropolishing range. Here, the current density exceeds the so-called critical current density, i_{CRIT} (see also Equation 4.1), marked by the first peak in the *I-V* curve. The silicon etches in this domain, leaving behind a smooth (electropolished) surface. It is generally assumed that the electropolishing of silicon takes place in two steps. In the first step an oxide layer forms, and this oxide layer subsequently dissolves in the HF (see Reactions 4.8 and 4.9). Hills on the surface dissolve faster than depressions because the current density is higher on high spots. As a result, the surface becomes smoother, that is, electropolishing takes place.[91] Electropolishing in this regime can be used to smooth silicon surfaces or to thin epitaxially grown silicon layers. The second peak and the oscillations in Figure 4.65 are explained as follows: at current densities exceeding i_{MAX} another type of oxide grows on top of the silicon, leading to a decrease of the anodic photocurrent (explaining the i_{MAX} peak), until a steady state is reached in which dissolution of the oxide by HF through formation of a fluoride complex in solution (SiF_6^{2-}) equals again the oxide growth rate. The oscillations sometimes observed in the anodic curve in Figure 4.65 can be explained by a nonlinear correlation between formation and dissolution of the oxide.[189]

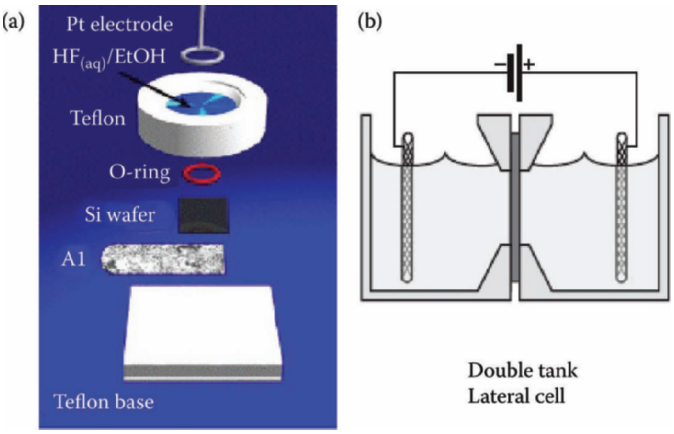


FIGURE 4.66 Cells for electrochemical and photoelectrochemical etching of Si. (a) The cell comes with a provision to illuminate the semiconductor. The Al is used for ohmic contact to the Si electrode. (Picture from http://sailorgroup.ucsd.edu/research/porous_Si_intro.html.) (b) Electrochemical double cell (also referred to as double tank lateral cell). (Gaburro, Z., N. Daldosso, and L. Pavesi, eds. 2005. *Encyclopedia of condensed matter physics: porous silicon.* Amsterdam: Elsevier Ltd.[195])

Electropolishing is also possible under a layer of porous Si already formed at the porous Si/bulk silicon interface, and the porous layer may be underetched this way. The latter is the enabling technology in the Bosch advanced porous Si membrane process (APSM) (see page 293). The APSM technology makes use of porous Si to form a cavity underneath the surface of monocrystalline silicon and turns the traditional bulk micromachining process into a surface micromachining technology.

Porous Silicon

Historical Note Uhlir first discovered porous Si (PS) in 1956 at the Bell Laboratories.[7] He found that, with the appropriate current and solution composition, silicon does not dissolve uniformly, and pores are produced instead. These pores, he found, propagate primarily in the <100> direction of the Si wafer. This did not provide the smooth polish for Si that he sought, and after his curious results were reported in a Bell Labs technical note, the work was pretty much forgotten. Renewed interest in porous Si came about in the 1970s and 1980s: its high surface area was found to be useful as a model in the spectroscopic studies of chemical adsorption/

FIGURE 4.67 Bright photoluminescence of porous Si at room temperature. (Science and art from Mikael Östling, KTH.)

desorption kinetics on crystalline Si surfaces, and the material also found application as a precursor to generate thick oxide layers on Si and as a dielectric layer in capacitance-based chemical sensors (see above and further below under "Porous Si Applications").

In the late 1980s, Leigh Canham, from the Defense Research Agency in the United Kingdom, argued that the thin Si filaments generated when the pores in porous Si become large and numerous enough might display quantum confinement effects. His intuition turned out to be correct, and by 1990 Canham found that electrochemically etched PS material was fluorescent at room temperature with a bright red-orange color (see Figure 4.67).[196,197] Shortly thereafter, in 1991, Richter demonstrated electroluminescence on porous Si in the 560–480-nm range.[198]

The observation that porous Si features a large photoluminescence (PL) effect in the visible range (at room temperature!) was a very surprising result because the PL efficiency of bulk Si (Si) is very low as a result of its indirect energy bandgap (typical room temperature light emission efficiencies for silicon are much less than 0.001%; in contrast, GaAs LEDs routinely have quantum efficiencies of 1–10%, and approaching 30% for specialized LEDs). As expected from the quantum confinement-induced blue shift, the emitted color is at an energy that is significantly larger than the bandgap energy for bulk Si (which occurs in the infrared region of the spectrum $E_g = 1.1$ eV) (see Figure 4.68 and Volume I, Chapter 3). The underlying mechanism for this blue shift is the result of the quantum-confined structures formed in the porous Si pore walls. High-porosity porous Si material emits green/blue light, whereas low porosity emits red light.

The discovery of efficient visible light emission from porous Si initiated an explosion of work focused on creating Si-based optoelectronic switches, displays, lasers, and so on. However, problems with the material's chemical and mechanical stability and the aging of its optoelectronic properties (see Figure 4.73) led to a waningof the interest in porous Si as an optoelectronic material by the mid-1990s (http://chem-faculty.ucsd.edu/sailor/research/porous_Si_intro.html). By the late 1990s, optoelectronic properties research of porous Si got a new lease on life, however, because of the potential of porous Si for photonic crystals (see below and Volume I, Chapter 5).[197] Porous Si might represent a simple way of making photonic crystals as explained in the section on Macroporus Silicon. Other unique features of porous Si, such as its large surface area, its controllable pore size, its convenient surface chemistry, and its compatibility with conventional silicon microfabrication technologies, also inspired research into applications far outside the optoelectronics field. Additional applications include biocompatible interfaces, biosensing, and drug delivery. An overview of porous

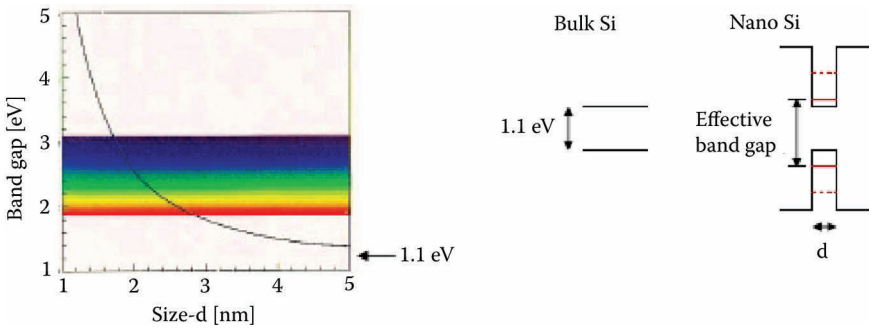


FIGURE 4.68 Bandgap widening of silicon in porous silicon.

Table 4.11 Porous Si Application Fields and Examples[195]

Application Area	Role of PS	Key Property
Optoelectronics	LED Waveguide Field emitter Optical memory	Efficient electroluminescence Tunability of refractive index Hot carrier emission Nonlinear properties
Microoptics	Fabry-Perot filters Photonic band gap structures All optical switching	Refractive index modulation Regular macropore array
Energy conversion	Antireflection coatings Photoelectrochemical cells Solar cells	Low refractive index Photocorrosion cells
Environmental monitoring	Gas sensing	Ambient sensitive properties
Microelectronics	Microcapacitor Insulator layer Low-k material	High specific surface area High resistance Electrical properties
Wafer technology	Buffer layer in heteroepitaxy SOI wafers; see also Eltran process	Variable lattice parameter High etch selectivity
Micromachining	Thick sacrificial layer	Highly controllable etching patameters
Biotechnology	Tissue bonding Biosensors	Tunable chemical reactivity Enzyme immobilization

Si applications is listed in Table 4.11. Today, porous Si, together with carbon nanotubes, is often used to illustrate nanotechnology products/applications that are relatively easy to make/implement.

Porosity in Si Porosity P of a material is the volumetric fraction of air in the material and is defined as:

$$P = \frac{V_{voids}}{V_{material}} = \frac{m_{original} - m_{anodized}}{m_{original} - m_{dissolved}} \qquad (4.12)$$

For example, porous Si with $P = 10\%$ means that there is 10% air and 90% silicon in the material. The tunable pore size range in porous Si can be divided into (International Union of Pure and Applied Chemistry definition)[195]:

1. Microporous: $d < 2$ nm
2. Mesoporous: 2 nm $< d <$ 50 nm
3. Macroporous: $d > 50$ nm

The geometry of porous Si varies from a sponge-like layer consisting of nanoscale branch-shaped fractal-like pores to straight pore arrays. These geometries depend on various conditions, such as the doping level and doping type of the Si substrate, the current density, HF concentration, and crystal orientation. The porous Si porosity can vary between 5% and 85%, and pore size can differ by three orders of magnitude, 20 Å to 10 µm, and the exact underlying formation mechanism, we shall see, is still under debate. The word *nanoporous* is sometimes used for the smallest-pore regime.

Accurate determination of pore size distribution in porous Si is usually performed by the analysis of adsorption isotherms of gases at low temperature in the so-called Barret-Joyner-Halenda, or BJH, method.[199] The physical adsorption of gases by a porous surface is increased relative to a nonporous one because of capillary condensation in the pores. This increase in adsorption starts when the gas pressure is high enough to fill the smallest pores. The specific surface area A of porous Si, determined by Brunauer-Emmett-Teller, or BET, ranges from A_{BET} = 1–1000 m^2/cm^3.[200] The structure of the pores in nanoporous Si can best be observed by transmission electron microscopy (TEM), and its thickness can be monitored with an infrared (IR) microscope. Structural properties of porous Si have also been investigated by a wide variety of other techniques such as electron microscopy, scanning probe microscopy, x-ray scattering techniques, x-ray absorption

techniques, and Raman spectroscopy. For example, Raman measurements were used to demonstrate the crystallinity of the porous Si skeleton and infer information on the size of the Si nanocrystals.

The very reactive porous Si material etches or oxidizes very rapidly. Heat treatment in an oxidizing atmosphere leads to oxidized porous Si (OPS). The oxidation occurs throughout the whole porous volume, and SiO_2 layers several micrometers thick can be obtained in times that correspond to the growth of a few hundred nanometers on regular Si surfaces (1100°C in oxygen for 30 minutes is sufficient to make a 4-µm-thick film).[201] Porous Si is low-density and remains single-crystalline, providing a suitable substrate for epitaxial Si film growth. These properties have been used to obtain dielectric isolation (fully isolated by porous oxidized silicon, or FIPOS) and to make silicon-on-insulator (SOI) wafers (see porous Si applications in Table 4.11); in dry etching (see Chapter 3), as prepared, porous Si etches six to seven times faster compared with single-crystal Si.[202] A typical etch rate of porous Si in a reactive ion etching (RIE) station, with SF_6 as the gas, is 6.8 µm/min (whereas for nonporous Si it is 1.5 µm/min), and in a high-density plasma station the etch rate is 66 µm/min (Si, 10 µm/min). The dry etching rate of porous Si depends on the porous Si porosity, aging of the layer, and thermal treatment. The etch rate of thermally treated porous Si layers is significantly smaller than that of freshly etched layers (0.33 µm/min in a high-density plasma).[203] By treatment of the porous Si with NH_3 at high temperatures, it is possible to make thick Si_3N_4 films. Even at 13 µm, these films show little evidence of stress in contrast to stoichiometric LPCVD nitride films.[204]

Porous Silicon Fabrication

Microporous Silicon Whether one is in the regime of electropolishing or porous Si formation depends on both the anodic current density and the HF concentration. The surface morphology produced by the Si dissolution process critically depends on whether mass transport or hole supply is the rate-limiting step. Porous Si formation is favored for high HF concentrations and low currents (low anodic bias for p-type and weak light intensities for n-type Si), where the positive charge supply (holes) is limiting, whereas smooth etching is favored for low HF concentration and high currents (higher anodic bias for p-type Si and strong light intensity for n-type Si), where mass transport is limiting.

For current densities less than i_{CRIT} (Figure 4.65), holes are depleted at the surface, and HF accumulates at the electrode/electrolyte interface. As a result, a dense network of fine holes forms.[202,205] Thus, porous Si typically forms when using low current densities in a highly concentrated HF/water or an HF/ethanol solution—in other words, by limiting the oxidation of silicon resulting from hole and OH^- deficiency. In view of the Si etching models introduced earlier, one observes that aqueous HF is less suitable for the porous etching process because the silicon surface is hydrophobic in water. The porous layer can be made more structurally uniform if an ethanolic solution (e.g., 1 part of 48% HF and 1 part ethanol) is used—this increases the wettability of the silicon and allows better surface wetting by the acid. Ethanolic etch solutions also reduce the formation of hydrogen gas bubbles because ethanol acts as a surfactant and prevents bubbles sticking to the silicon surface. The potential range in Figure 4.65 where porous Si forms is at potentials below (toward negative potentials) the first current peak (v_p; see also Figure 4.37). During the formation of porous Si, hydrogen gas evolves, and the morphology of the porous Si depends, among other factors, on the doping density of the wafer, orientation, the HF concentration, and the current density (and thus the light intensity for n-type Si and the voltage for p-type Si). As pointed out earlier, porous Si was shown to be monocrystalline,[92] and the suggested reaction mechanism is based on the assumption that fluoride ions close to the surface cause the Si-Si backbonds to be polarized, thus facilitating the breaking of these bonds.[206] The Si backbonds are broken, leaving an H-terminated surface and a SiH_xF_y compound that reacts with water and HF to form SiF_6^{2-} and H_2. The overall process, as introduced earlier, is given by Reaction 4.7.

As sketched in Figure 4.69, microporous silicon is easy to obtain: a few minutes of anodization in an HF solution suffices (a typical anodization rate is 0.3 µm/min).

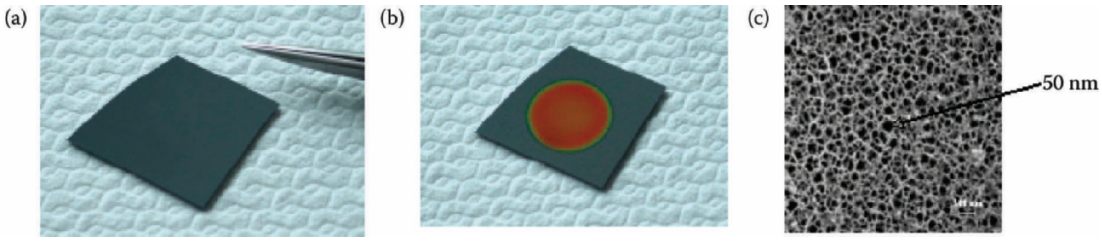


FIGURE 4.69 (a) Si piece. (b) Porous Si in circular area. (c) SEM of porous Si nanosponge.

Porous Si can also be formed chemically through Reactions 4.19 and 4.21. This is a much less common technique to form porous Si. The purely chemical process is called stain etching, or chemical etching (with no current flow), and is performed in an HF-HNO_3 solution. In this case, the difference between chemical polishing and porous Si formation conditions is more subtle. In the chemical polishing case, all reacting surface sites switch constantly from being local anode (Reaction 4.19) to being local cathode (Reaction 4.21), resulting in nonpreferential etching. If surface sites do not switch fast between being local anode and cathode, charges have time to migrate over the surface. In this case, the original local cathode site remains a cathode for a longer time, and the corresponding local anode site, somewhere else on the surface, also remains an anode to keep the overall reaction neutral. A preferential etching results at the localized anode sites, making the surface rough and causing porous Si to form.[207] Any inhomogeneity, for example, some oxide or a kink site at the surface, might increase such preferential etching.[207] Unlike porous Si fabricated by electrochemical means, chemically etched porous Si film thickness is self-limiting.

Depending on porosity and/or pore size, porous Si has the general properties summarized in Table 4.12 (single-crystal Si values are listed as reference). Remember that these properties can be tailored by the process conditions.

The electrical resistivity of porous Si is up to five orders of magnitude higher than that of intrinsic nonporous Si. The reason is that porous Si is depleted of free carriers. This depletion occurs either because 1) energy gap widening caused by quantum confinement, which reduces the thermal generation of free carriers, or 2) trapping of free carriers. Trapping of free carriers takes place during the preparation of the porous Si, and again there are two possible explanations: 1) the binding energy of the dopant impurities is increased, or 2) the formation of more surface states occurs. It has been shown that the dopants are still present in the porous Si in a concentration unchanged from the original Si, but that the dopants are in a neutral state. The electrical transport through porous Si is very much affected by the disordered structure of the porous Si skeleton that restricts the conduction paths to a highly constrained geometry, which for certain types of porosities forms a percolated or fractal geometry. As a consequence, conductivity is thermally activated, strongly frequency-dependent, and highly dispersive. Several models have been proposed to explain the electrical transport properties in porous Si. They differ on the

TABLE 4.12 General Properties of Porous Si

Property	Porous Si Value	Single-Crystal Si Value
Young's Modules E	83 GPa (P = 20%)–0.87 GPa (P = 90%)	160 GPa
Thermal conductivity	1.2 W/mK (nano)–80 W/mK (meso)	157 W/mK
Electrical resistivity	10^{10}–10^{12} Ω cm nanoporous	
Optical bandgap	1.4 eV (P = 70%)–2.0 eV (P = 90%)	1.12 eV
Refractive index	1.2–2.8	3.94
Oxidation 1100°C for 30 min in oxygen	4 μm FASTER	<0.5 μm SLOWER
Dry etching (SF_6 gas)	66 μm/min FASTER	10 μm/min SLOWER

Source: Gaburro, Z., N. Daldosso, and L. Pavesi, eds. 2005. *Encyclopedia of condensed matter physics: porous silicon.* Amsterdam: Elsevier Ltd.[195]

transport paths and mechanisms assumed. The proposed transport paths range from transport in the Si nanocrystals (with diffusion or tunneling between the Si nanocrystals or at their surface) to transport in the amorphous and disordered matrix surrounding the nanocrystals, or through both.

The nanoporous Si structures may have dimensions in the low nanometer range. If the structure size reaches a value less than, e.g., 3 nm, quantum effects can occur; therefore, nanoporous samples can exhibit strong visible photoluminescence and electroluminescence. Raman spectroscopy gives indirect information about the nanostructure of porous Si and has shown that the nanocrystals alter the selection rules relating to the interaction of optical phonons with incident photons (see Volume I, Chapter 5). This results in the efficient photoluminescence and electroluminescence of porous Si compared with nonporous Si. The first and most favored explanation for the visible light emission in porous Si is the quantum confinement of excitons in nanometer-sized Si nanostructures. The nature of the energy gap is still indirect even though a quasidirect bandgap can be formed in ultrasmall Si nanocrystals. The other properties of porous Si listed in Table 4.12 will be highlighted in the text below.

Macroporous Silicon In addition to micropores, well-defined macropores ($d > 50$ nm) can also be made in Si by photo and/or bias etching in HF solutions. The macroporous material exhibits very low values of porosity ($P < 10\%$) and high levels of roughness. Macropores have sizes as large as 10 μm, visible with a scanning electron microscope (SEM) rather than the transmission electron microscope (TEM) typically used for nanoporous material. The two types of pores often coexist, with micropores covering the walls of macropores. This is not a matter of a broad fractal-type distribution of pores because the formation mechanism for the large pores is different.

Conditions of lower HF concentration and higher current density tend to offer larger diameter pores. Electrochemical etching of macropores or macroholes has been reported, for example, with n-type silicon in 2.5–5% HF under high voltage (>10 V), low current density (10 $mAcm^{-2}$), under illumination,

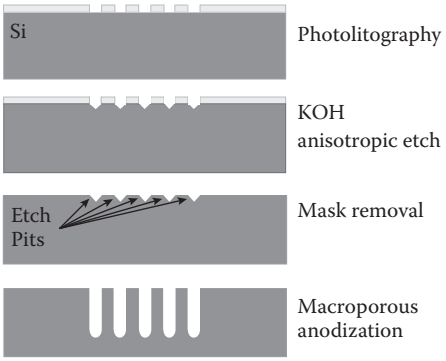


FIGURE 4.70 Schematic lithographic procedure resulting in the definition of pore-starting etch pits. (Gaburro, Z., N. Daldosso, and L. Pavesi, eds. 2005. *Encyclopedia of condensed matter physics: porous silicon.* Amsterdam: Elsevier Ltd.[195])

and in the dark.[189] In the latter case, Zener breakdown in silicon (electric field strength in excess of 3×10^5 V/cm) causes the hole formation (see also Volume I, Chapter 4 on Zener diodes and Volume I, Figure 4.43). In the case of lightly doped n-type silicon, the macropores are formed at much higher anodic potentials than those used for micropore formation.[208] More details on models explaining porous Si formation follow below. By using a pore initiation pattern, the macropores can be localized at any desired location. This option is limited to macroporous porous Si because present lithographic resolution lowest limits are typically in the range of 65 nm (Figure 4.70).

The dramatic effect of the use of a pore initiation pattern is illustrated in the comparison of Figures 4.71 and 4.72.[91,209] In Figure 4.71 we show representative porous Si layers without the use of a pore initiation pattern. Figure 4.71a illustrates nanoscale pores (mesoporous Si),[203] and Figure 4.71b shows microscale pores (macroporous Si).[91] The 45° bevel cross-section in Figure 4.71b is of an n-type sample ($10^{15}/cm^3$ phosphorus-doped) and shows a random pattern of macropores. In contrast, the macroscale pores in Figure 4.72 all start at the micropit initiation points. Pores orthogonal to the surface with depths up to a whole wafer thickness can be made, and aspect ratios as large as 250 are possible.

The periodic arrays of macropores shown in Figure 4.72 can form a 2D photonic crystal. Photonic crystals, we recall, are materials with a periodically varying index of refraction, and in porous

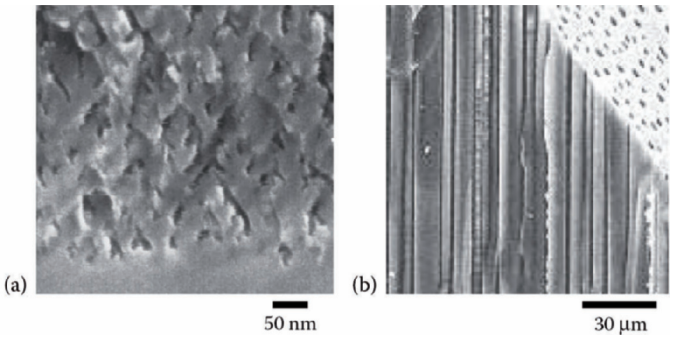


FIGURE 4.71 Representative porous Si layers. (a) Nanoscale pores (mesoporous Si). (Tserepi, A., et al., 2003. Dry etching of porous silicon in high density plasmas. *Physica Status Solidi (A)* 197:163–67.[203]) (b) Microscale pores (macroporous Si). (Lehmann, V. 1993. The physics of macropore formation in low doped n-type silicon. *J Electrochem Soc* 140:2836–43.[91]) Random: surface, cross-section, and a 45° bevel of an n-type sample (10^{15}/cm^3 phosphorus-doped) showing a random pattern of macropores. Pore initiation was enhanced by applying 10-V bias in the first minute of anodization followed by 149 minutes at 3 V. The current density was kept constant at 10 mA/cm^2 by adjusting the back-side illumination. A 6% aqueous solution of HF was used as an electrolyte.

Si, air alternates with Si (see Volume I, Chapter 5). Grüning et al.[210] realized such a photonic crystal with a photonic bandgap in the near IR using macropores in n-type silicon etched under back-side illumination.

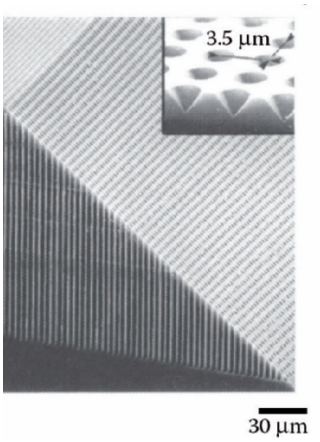


FIGURE 4.72 SEM image of well-ordered localized or arrayed straight macropores. Inset, SEM image is an artificial micropit array formed in advance of Si electrochemical etching. Localized: surface, cross-section, and a 45° bevel of an n-type sample (10^{15}/cm^3 phosphorus-doped) showing a predetermined pattern of macropores (3 V, 350 min, 2.5% HF). Pore growth was induced by a regular pattern of pits produced by standard lithography and subsequent alkaline etching (inset top right). (From Lehmann, V. 1993. The physics of macropore formation in low doped n-type silicon. *J Electrochem Soc* 140:2836–43. With permission.[91])

Summarizing, we have learned that single-crystal silicon can be turned into a porous material with porosity and pore sizes determined by the current density, type, and concentration of the dopant, as well as the hydrofluoric acid concentration. Also, a transition from pore formation to electropolishing is reached by increasing the current density and/or by decreasing the hydrofluoric acid concentration. Electropolishing is even possible under a layer of porous silicon (PS) already formed. The latter is the enabling technology in the Bosch APSM process (see page 293), which makes use of PS to form a cavity underneath the surface of monocrystalline silicon.

In Chapter 7 we learn that PS can also be formed under similar conditions from LPCVD polysilicon, with pores roughly following the grain boundaries of the polysilicon. We will see how, by changing the conditions from pore formation to electropolishing and back to PS, enclosed cavities may be formed with porous polysilicon walls (plugs) and "floor and ceiling" of silicon nitride layers.

Porous Silicon Drying and Aging When drying a porous Si film in air, a water meniscus forms in the pores, and this results in stress exerted on the pore walls. Depending on the size of the structure (porosity) and the surface tension of the liquid, cracking of the porous Si layer is always a possibility. This limits the maximum porosity obtainable and also suggests different procedures for drying, for example, by working with very low surface tension liquids in the pores when drying. This may be accomplished by rinsing the freshly made porous Si with pure ethanol that has a lower surface tension than water (22 mJ/m^2 compared with 72 mJ/m^2); an alternative is to use pentane (with a surface tension of 14 mJ/m^2). The best results, however, have been obtained by supercritical drying in CO_2 (>95% porosity), where drying is performed above the supercritical point of a liquid, usually CO_2.[211] Other methods include freeze drying and using very slow evaporation rates. Pentane drying is the easiest to implement. Pentane has a very low surface tension and shows no chemical interaction with porous Si (unlike ethanol). Using pentane as drying liquid strongly reduces the

capillary tension, but because water and pentane are nonmiscible, ethanol or methanol has to be used as intermediate rinsing liquid. Using this drying technique, PS layers with porosity values up to 90% and thickness up to 5 µm exhibit no cracking after drying. Supercritical drying requires a specific apparatus but is more effective in preventing cracking of the film; thus, porous layers of porosity values up to 95% were demonstrated.[211]

In the electrolyte and immediately after drying, the pore walls are mostly H-terminated. The hydrogen termination is replaced by oxygen to form native oxide rapidly in air.[211] This changes the properties of the porous Si over time, making it one of the major problems preventing the implementation of porous Si in practical applications. An example of the nefarious effect of porous Si aging is its effect on photo- and electroluminescence of porous Si. The so-called luminescence fatigue shown in Figure 4.73 is explained by photochemical reactions occurring on the surface of the Si nanocrystals. As the time passes, a shift to the blue region in the peak position and both decreases and increases in the PL intensity have been reported. The blue shift has been attributed to the decrease in dimension of porous Si pillars resulting from the oxidization of the surface with time, which supports the quantum confinement effect.

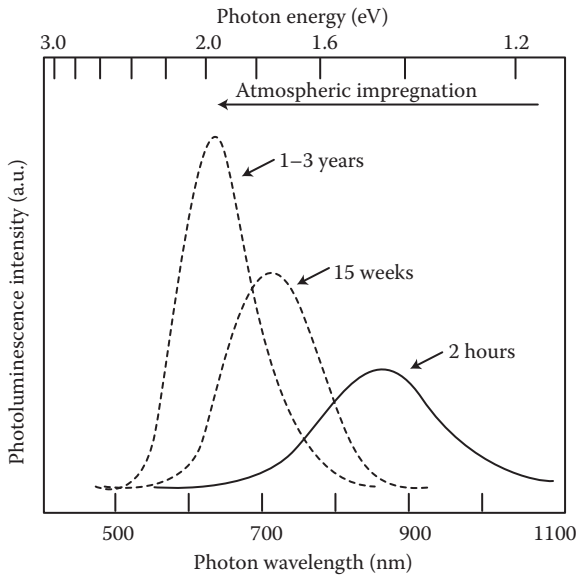


FIGURE 4.73 Aging of the luminescence of porous Si. (Thönissen, M., and M. G. Burger, eds. 1997. *Properties of porous silicon*. Washington, DC: IEEE.[211])

Because of the large surface area of the porous Si, the silicon/oxygen ratio may be large, resulting in a significant impact on the properties of porous Si. Several methods for surface passivation have been reported in the literature, and they seem to lead to more stable luminescence.[211] To name a few, controlled oxidation by anodic or chemical oxidation, rapid thermal oxidation, capping of the porous Si layer by a dielectric or metal, thermal nitridation, and thermal carbonization.[211]

Porous Si Models The critical question about any kind of porous Si formation, nano- or macroporous, is why there is a difference in dissolution rate between pore tip and pore walls? Unagami[205] proposed that insoluble silicates passivate the pore walls, resulting in the formation of a porous Si layer, but this does not explain why such selective passivation takes place in the first case. Parkhutik et al.[212] compared porous Si formation with that of porous alumina formation (see Volume III, Chapter 5). In this model a passive layer (e.g., an oxide) exists at the bottom of each pore and results in the electric field enhancement at the pore tips. Poreformation in this case is a consequence of the dissolution of the passive layer enhanced by the local electrical field, which is stronger at the sharp tips. The competition between growth and dissolution of the passive layer gives rise to a time-evolving electrolyte/passive film interface defining the surface of the pore. This model predicts the correct dependence of pore size on HF concentration. In 1970, Theunissen et al. proposed that porous Si forms in n-type Si as a result of Zener breakdown because of electric field enhancement at the pore tip.[87,163] However, Theunissen's model cannot explain other kinds of porous Si formation without this breakdown condition (e.g., what about porous p-type Si). In 1985, Beale et al.[211] expanded Theunissen's model by introducing the involvement of a space charge region (SCR), which forms at the surface region of the Si in contact with the electrolyte and acts as a Shottky barrier, as illustrated for p-type Si in Figure 4.74. The SCR is depleted of holes and prevents charge transfer at the interface between the electrolyte and the Si surface. When a pore wall becomes sufficiently thin (small d in Figure 4.74), space charge regions from opposite sides of the pore

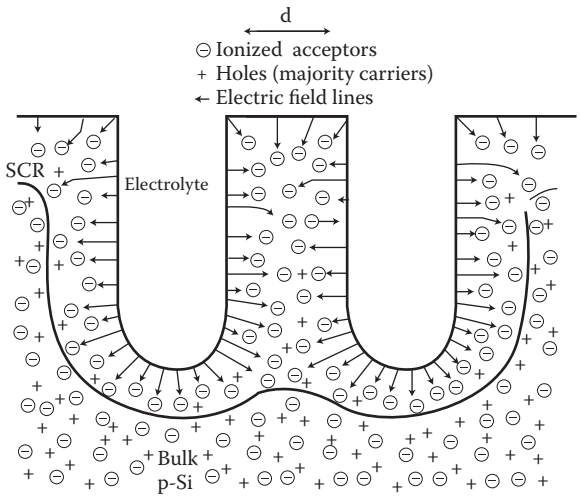


FIGURE 4.74 Space charge regions (SCRs) at the semiconductor/electrolyte interface form an insulating layer that prevents charge transfer. As two space charge regions in a wall pore wall of thickness *d* overlap that wall will passivate, and current is directed to the bottom of the pore. The example shown here involves p-type Si. (Lehmann, V., et al. 1999. *J Electrochem Soc* 146:2968–75.)

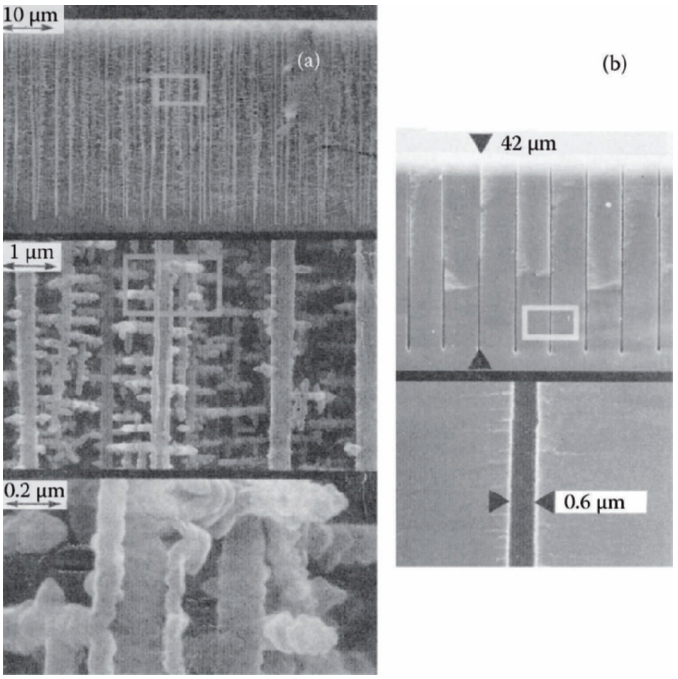


FIGURE 4.75 Comparison of macropores made with breakdown holes (a) and macropores made with light-created holes (b). (a) An oxide replica of pores etched under weak back-side illumination visualizes the branching of pores produced by generation of charge carriers resulting from electrical breakdown (5 V, 3% HF, room temperature, 10^{15}/cm^3 phosphorus-doped). (From Lehmann, V. 1993. The physics of macropore formation in low doped n-type silicon. *J Electrochem Soc* 140:2836–43. With permission.[91]) (b) Single pore associated with KOH pit. (From Lehmann, V., and H. Foll. 1990. Formation mechanism and properties of electrochemically etched trenches in n-type silicon. *J Electrochem Soc* 137:653–59. With permission.[213])

wall overlap, thus decreasing charge transfer at the interface even further. Consequently, current is selectively directed to the pore tips' region, and pore walls stop thinning when $d < 1/2$ SCR. Pore formation in this model is a self-adjusting electrochemical mechanism. The limitation is a result of the depletion of holes in the pore walls, causing them to stop etching, whereas holes continue to be collected at the pore tip, where they continue to promote Si dissolution. No passivating layer is involved to protect the pore wall in this model. The only decisive differences between pore tips and pore walls are their geometry and their location. Holes generated by light or Zener breakdown are collected at pore tips. Every depression or pit in the surface initiates pore growth because the electrical field at a curved pore bottom is much larger than that of a flat surface as a result of the effect of the radius of curvature. The latter leads to higher current and enhances local etching.[208] Zener breakdown and illumination of n-type Si lead to different types of pore geometry.[91,163,213] Branched pores with sharp tips form if holes are generated by breakdown (see Figure 4.75a).[91] Unbranched pores with larger tip radii result from holes created by illumination (see Figure 4.75b).[213] The latter difference can be understood as follows: the electric field strength is a function of bias, doping density, and geometry. High doping level density or sharp pore tips will lower the required bias for breakdown; therefore, macropores will tend to follow pores with the sharpest tips. Because every tip causes a new breakdown and hole generation, the position of the original pore tip becomes independent of the other pores, branching of the pores is possible, and fir tree-type fractal-like pores can be observed. With illumination the pore radius may be larger because the breakdown field strength is not necessary to generate charge carriers; thus, the pores remain unbranched.

To obtain the necessary depletion layers in both n- and p-type Si under anodic bias, Beale et al. had to assume Fermi level pinning at midgap (for example, see the energy diagram for n-Si in HF in Figure 4.62).

Yet another model, the diffusion-limited model developed by Smith et al.,[214] describes the formation

of porous Si in terms of a stochastic random walk. During pore formation, a hole diffuses to the silicon/electrolyte interface and reacts with a Si surface atom. The nature of the random walk presents the pore tip as the most likely contact site for a hole diffusing from the bulk Si region and provides a similar selective dissolution mechanism of surface irregularities as the Beale model. The rate of pore growth, in this model, is limited by the diffusion of holes to the growing pore tip. A characteristic diffusion length, function of dopant concentration, current density, and so on control the different pore morphologies. Unlike the Beale model, this model applies very generally to all electrochemical phenomena. Smith et al. confirmed that the porous Si geometry simulated by random walk and diffusion-limited model is similar to the geometry of formed porous Si.

In 1990, Lehmann et al.[197] proposed a model for nanoporous Si formation based on the quantum confinement effect. As described above, it is well known that porous Si is luminescent in the visible region. Lehmann et al. expanded the quantum confinement effect to explain porous Si formation. The increase in bandgap for quantum wires and quantum dots in the pore walls decreases the mobile carrier concentration, i.e., produces the barrier to charge transfer.[197] As a result, the current is again limited to the pore tips, resulting in porous layer formation. Obviously this model can only account for nanoporous Si formation.

As for the direction of the pores in porous Si, there will be a natural tendency to follow the direction of the strongest field with the highest supply of holes, i.e., in the <001> direction on a (001) substrate. The <100> and <113> directions have been found to be the preferred growth directions of the macropores in both p- and n-type silicon.[215] The growth direction depends on the surface orientation. When using a {111} surface, three equivalent <113> macropores start to grow into the substrate. The macropore walls consist of (111) facets that can be observed by TEM. These pores can intersect and form a 3D pore network in the silicon, making for a possible 3D photonic crystal. The resulting structure is slightly different from the Yablonovite photonic crystal introduced in Volume I, Chapter 5; the pores are tilted about 29° off the vertical axes, whereas in the Yablonovite they have an angle of about 35° (see Volume I, Chapter 5). Large pore formation for (100) and (111) Si is illustrated in Figure 4.76.

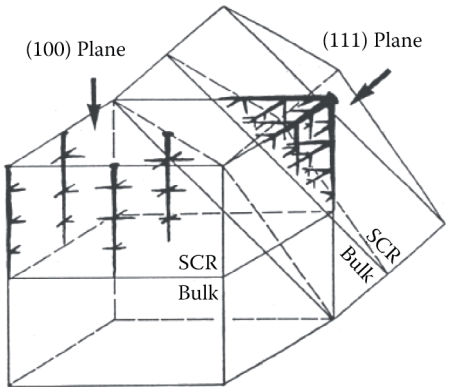


FIGURE 4.76 Large pore-formation for (100) and (111) Si; SCR, space charge region.

Lehmann has shown that square pore cross-sections and pore arrays of high porosity can be produced by subsequent alkaline etching of macroporous Si.[216] With the proper choice of etchants, the rounded pore walls can be etched to reveal (100) or (110) faces.

It was also determined that the macrohole formation in porous Si depends on the wavelength of the light used. It was found that no pore formation occurs below ~800 nm. Depending on the wavelength, the shape of the pores can be manipulated as well.[213] For wavelengths above 867 nm, the depth profile of the pores changed from conical to cylindrical. The latter was interpreted in terms of the influence of the local minority carrier generation rate. Carriers generated deep in the bulk would promote the pore growth at the tips, whereas near-surface generation would lead to lateral growth.

In 1993, Cahill et al.[217] reported the creation of 1–5-μm-diameter pores with pore spacings (center to center) from 200–1000 μm. Until this finding, the pores typically formed were spaced in the range of 4–30 μm center to center and 0.6–10 μm in diameter. Making highly isolated pores presents another challenge. In the previous work, the relatively close spacing of the pores allowed the authors to conclude that the regions between the pores were almost totally depleted and that practically all carriers were collected by the pore tips. In such a case, neither the pore sidewalls nor the wafer surface etched as all

holes was swept to the pore tips. Because the surface was not attacked by pore-forming holes, the quality of the pore initiation mask lost its relevance. In Cahill et al.'s case, on the other hand, a long-lived mask (>20 h) is needed to help prevent pore formation everywhere except at initiation pits. The mask used was a SiC layer sandwiched between two layers of Si nitride. The silicon nitride (insulator) directly atop the Si (semiconductor) serves to insulate the silicon carbide (semiconductor) from the underlying silicon substrate. Because the silicon carbide proves very resistant to HF, loss of thickness does not show during the procedure. The top nitride serves to protect the carbide during anisotropic pit formation. By lowering the bias applied to the Si substrate to less than 2 V with respect to SCE, side branching is avoided.

The Si anisotropy common with alkaline etchants surprisingly shows up here with acidic isotropic etchants. For example, with breakdown-supplied holes, <100>-directed macroholes with <110> branches form (see Figure 4.76),[91] and with <111>-directed wafers, <113> branches form (see Figure 4.76), leading to a complex network beneath the silicon surface. Pyramidal pore tips were also observed when the current density was limited by the bias ($i < i_p$).[91] Isotropic pore pits form when the current is larger than the critical current density, that is, isotropy in HF etching can be changed into anisotropy when the supply of holes is limited. We refer here to the Elwenspoek et al. model (see above), which predicts that in confined spaces etching will tend to be more anisotropic, even when using a normally isotropic etchant such as HF.

The enhanced photoluminescence efficiency of porous Si—compared with Si—has motivated active research toward other porous semiconductor materials. For example, highly porous InP, SiC, GaP, $Si_{1-x}Ge_x$, and Ge structures have all been investigated. Among the first examples of strong anisotropy in porous etching of semiconductors other than Si was that reported by Takizawa et al.,[203] who reported pores in <111> A-oriented InP in HCl with an aspect ratio larger than 100. Some other preliminary evidence was collected on GaAs[218] and GaP.[210] Pore formation in TiO_2 anodic and alumina films is described in more detail in Volume III, Chapter 5.

Porous Silicon Applications Uhlir's discovery has led to all types of interesting new devices from quantum structures, chemical sensors and biosensors, permeable membranes, and photoluminescent and electroluminescent devices to a basis for making thick SiO_2 and Si_3N_4 films.[201,204] Some of these applications are listed in Table 4.11 introduced earlier. Besides its use for dielectric isolation and the fabrication of SOI wafers, porous Si has been introduced in a wide variety of other applications: Luggin capillaries for electrochemical reference electrodes, high surface area gas sensors, humidity sensors, sacrificial layers in silicon micromachining, and so on.

We already highlighted the fact that porous Si exhibits photoluminescent and electroluminescent behavior; light-emitting porous Si (LEPOS) was demonstrated. If a LEPOS device could be integrated monolithically with other structures on silicon, a big step in micro-optics, photon data transmission, and processing would be achieved. To explain the blue shift of the absorption edge of LEPOS of about 0.5 eV compared with bulk silicon[197] and room temperature photoluminescence,[196] Searson et al.[219] proposed an energy-level diagram for porous Si where the valence band is lowered with respect to bulk silicon to give a bandgap of about 1.8 eV (see Figure 4.68).

Porous Si can easily be etched in relatively harmless solutions like diluted aqueous KOH solutions, making it an attractive sacrificial material. The surface and bulk properties of porous Si are easily modified; for example, oxidization of the material makes it useful as an insulating material in gas sensors; metallization enables its use in microelectronics; and nitridation of the material is possible as well.

The large specific surface area (e.g., 200 m^2/cm^3) and the very small diameters of the pores make porous Si an interesting material for several analytical MEMS and NEMS devices. Given the large specific area per unit volume, the electrical transport is strongly influenced by external factors such as residual electrolyte and ambient atmosphere. The latter property is very interesting for sensor applications. Certain gases—for example, NO—have the capability of modifying the free carrier population. Changes in the electrical conductivity in the presence of subparts per million concentrations of such gases can be detected at room temperature operation. Other

gaseous species (for example, polar liquid vapors) also affect electrical transport via electric field interactions with charge carriers.

An optical biosensor based on porous Si also has been demonstrated. PS samples were prepared in such a way that the porous Si films displayed Fabry-Perot fringes in their white-light reflection spectrum. When biological molecules are then chemically attached as recognition elements to the porous Si surface and exposed to the appropriate complementary binding pair, binding occurs and is observed as a shift in the Fabry-Perot fringes.[220,221]

Lammel et al.[222] fabricated tunable optical filters based on porous Si—filters that can be used in reflective or transmission mode. The ratio of voids to total volume of porous Si determines its refractive index, and because porosity can be adjusted by varying the current density, these authors were able to micromachine a multilayer stack of porous Si of different indices by modulating the current density during porosification (in principle, indices between 1.6 and 2.1 can be obtained this way). The process was applied to an area of single-crystal Si delineated by a Si nitride window. The porosified Si plate was released from the substrate by subsurface electropolishing. During chemical release, two suspension arms lift the plate out of the plane automatically by internal stress release. The suspension arms are also provided with Cr/Ni heater wires to actuate the filters. By tilting the freestanding plate of porous Si with the thermal bimorph suspension arms, a wavelength scan is possible. Using this process, Lammel et al. achieved a 20-nm wavelength resolution in the visible part of the spectrum. Figure 4.77 is an SEM of the described porous Si plate optical filter.

Another more recent interest in porous Si stems from the ability to culture mammalian cells directly onto porous Si. Initial research findings from DeMonfort University suggest that the material may be biocompatible (http://www.findarticles.com/m0WVI/ 1999_April_6/54338205/p1/article.jhtml and L.T. Canham[223]). In this vein porous Si can also be used for sieving to filter very small particles. For more on biocompatibility of single-crystal Si, polycrystalline Si, and porous Si, see also Volume III, Chapter 5.

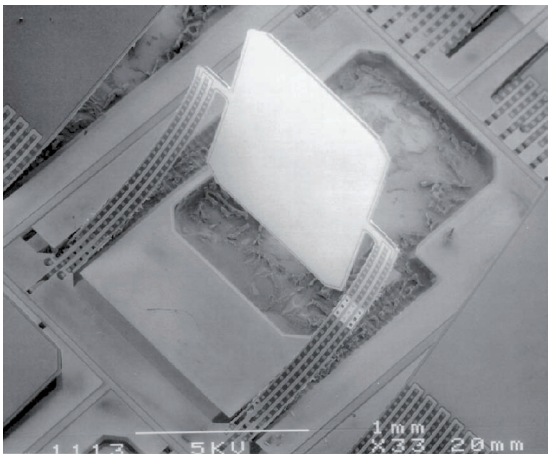


FIGURE 4.77 Optical filter of porous Si. SEM of a freestanding porous Si microplate containing a multilayered optical interference filter. The wavelength can be tuned by tilting the microplate using the integrated thermal bimorph microactuator. The tilt angle is a function of the actuator DC voltage. This filter element can be used to build a microspectrometer. (Courtesy of Dr. Gerhard Lammel, EPFL, Lausanne, Switzerland.[222])

Porous Si also enabled an elegant new micromachining technology at Bosch with their advanced porous silicon membrane (APSM) process. This technology makes use of porous Si to form a cavity underneath the surface of a single-crystalline silicon membrane, which helps to make the wafer manufacturing process fully CMOS-compatible and turns the traditional bulk micromachining process into a surface micromachining technology. Armbruster et al.[224] reported this new micromachining technology for the fabrication of monocrystalline silicon membranes covering a vacuum cavity for applications such as piezoresistive pressure sensors. In Figure 4.78 we compare a silicon-based absolute pressure sensor made with traditional bulk micromachining with the same device made with the new porous Si-based APSM process. The new process evidently produces not only cavities inside a device in a shorter amount of time, uses no wafer-bonding steps, and constitutes an all-silicon process, but it is also is CMOS-compatible. The main process steps, illustrated in Figure 4.79, are (a) local anodic etching of a layered porous Si (LPCVD Si_3N_4 is used as a mask). (b) One of the following processes is used: 1) conversion of part of the porous material to oxidized porous Si and etching it away, 2) direct electropolishing of part of the porous Si, or 3) thermal rearrangement of

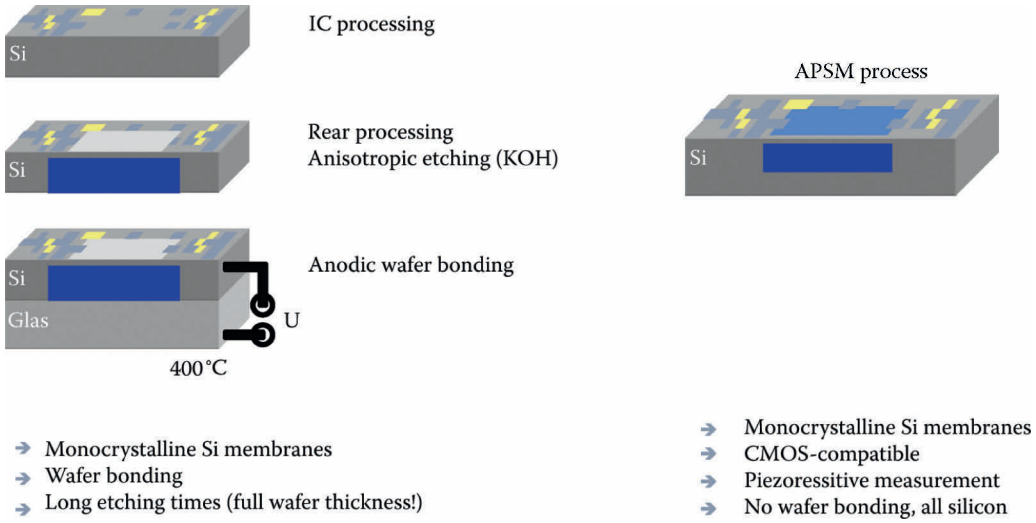


FIGURE 4.78 Comparison of a bulk Si micromachined absolute pressure sensor (left) with such a sensor made with the new Bosch porous Si-based APSM process (right). (From a Bosch presentation.)

part of the porous layer. These three different techniques all lead to a cavity underneath a suspended porous membrane, and (c) epitaxial growth of a silicon membrane layer over the remaining porous Si membrane. As we saw earlier, porous Si remains single-crystalline, providing a suitable substrate for epitaxial Si growth. The layer thicknesses of the porous layer over the cavity and the cavity layer may be 1–10 µm or more, depending on the application.

The degree of porosity of the porous Si layer in the APSM process deliberately varies in the lateral and vertical direction. The lateral doping of the semiconductor material for the edge region and central region of the porous diaphragm is selected in such a way that mesopores (pores on the order of 5–50 nm) are formed in the edge region and macropores (pores > 50 nm and up to a few micrometers) are formed in the center region. This may be accomplished, for example, with an n-type doping to produce macropores in the center region, whereas the edge region of the diaphragm is provided with p^+ doping to obtain mesoporous Si. This makes it possible for the gases produced in etching to escape more easily through regions of the porous layer, which have a comparatively higher porosity, when producing the cavity underneath the porous layer. In the edge region of the diaphragm with the lower porosity, a good layer quality is possible, for example, for the subsequent epitaxial process, whereas in the central region of the diaphragm the epitaxial quality is comparatively lower as a result of higher-porosity sections. However, in the pressure sensor application, this is unimportant because the properties of the pressure sensor are not impaired. The semiconductor substrate is also doped differently in the region of the porous diaphragm layer and in the region of what will become the cavity in such a way that mesopores are produced in the semiconductor material of the diaphragm layer and nanopores (pores from 2–5 nm) are produced as a "precursor structure" in the cavity region. The production of a small-pore precursor structure prevents the formation of comparatively large air bubbles, which makes it possible to adequately remove the gas bubbles through the porous diaphragm layer. In this procedure, use is also made of the fact that the "nanostructure" has a considerably higher internal surface area than the "mesostructure" of the diaphragm region, which is usable in a subsequent process step for a shorter oxidation time. We learned above that the very reactive porous Si etches or oxidizes very rapidly. This makes it possible to subsequently produce a completely oxidized porous Si, which is then selectively removed, e.g., in a vapor etching process. The cavity may also be produced by rearranging the nanoporous layer in a high-temperature process step. Different porosities subjected to heat treatment at high temperatures in a reducing atmosphere induce a structural rearrangement, forming in this way a membrane with a thickness ranging from few tenths of micrometers to 1 µm, as well as the cavity under the membrane.

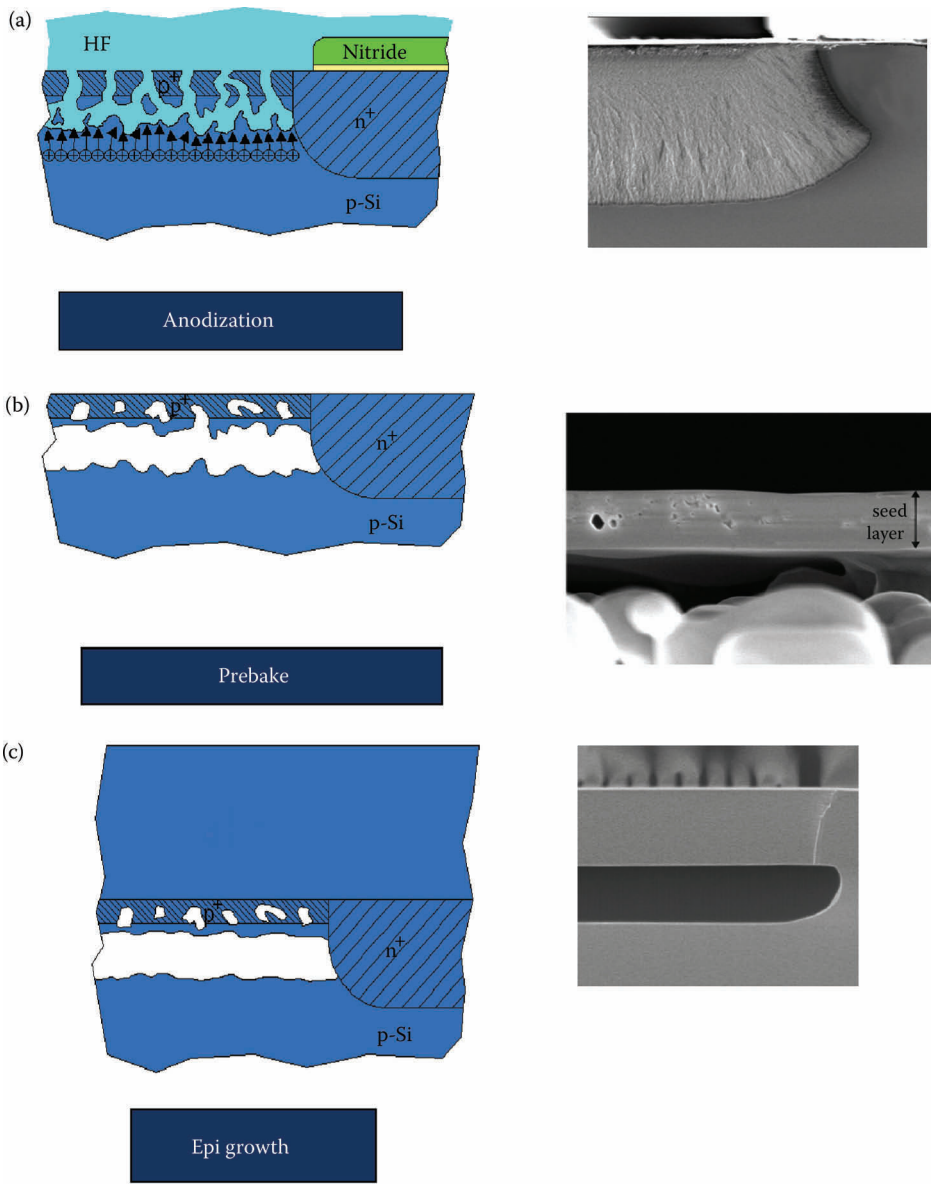


FIGURE 4.79 The Bosch porous Si-based APSM process. (From a Bosch presentation.) (a) Local anodic etching of layered porous Si with different porosities. (b) Thermal rearrangement of the porous Si. (c) Epitaxial growth of the Si membrane layer. (From a Bosch presentation.)

A third alternative procedure in forming the cavity structure involves directly dissolving the semiconductor material underneath the porous layer using electrolytic polishing, for example, at comparatively higher current densities and lower HF concentrations (see also page 287).

In contrast to conventional bulk micromachining, the new technology has the benefit of more design freedom because the shape/size of the monocrystalline silicon membrane only depends on the porous Si region. Except for the anodic etching, all process steps are part of a standard mixed signal IC production line. We may also compare the APSM process to a surface micromachined membrane pressure sensor. In surface micromachining, we learn in Chapter 7, to make a membrane sensor, a sacrificial layer is normally used, which is applied to the front of a carrier substrate before depositing the sensor diaphragm. The sacrificial layer is later removed from the front of the sensor through "dissolution openings" in the diaphragm, whereby an unsupported structure is obtained. The surface micromachining method is relatively complicated because of the need for separate sacrificial layers. Using APMS,

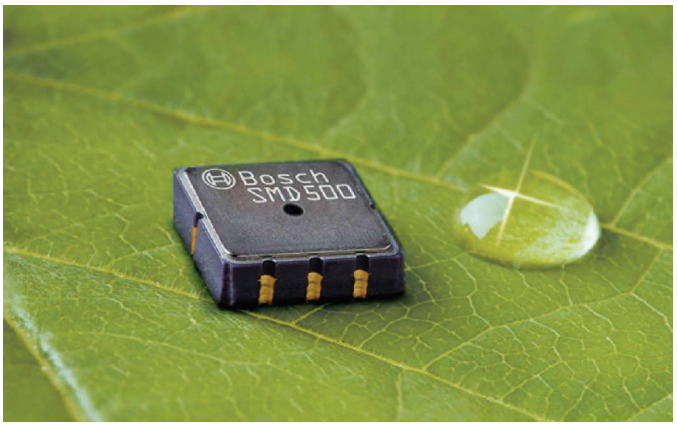


FIGURE 4.80 The Bosch SMD 500. The world's smallest digital pressure sensor! With a size of only 5 × 5 × 1.6 mm, the SMD 500 is the smallest pressure sensor for consumer electronics available. Manufactured using the new APSM technology (advanced porous Si membrane), it is much smaller compared with conventional bulk micromachined sensors. (From a Bosch presentation.)

it is possible to considerably simplify the manufacture of such component because no sacrificial layer needs to be additionally applied, and the sacrificial layer and the porous diaphragm are produced from the semiconductor material itself.

A piezoresistive pressure sensor (the SMD 500, the world's smallest digital pressure sensor) with integrated ASIC based on the new fabrication method is shown in Figure 4.80. The sensor that is shown here can measure pressure ranges from ±2.5 kPa (relative) up to 3.3 MPa (absolute). The sensor accuracy is 1.0% over the sensor lifetime, and its high robustness makes it ideal for applications like barometer, altimeter, weather forecast, health care, and fluid level measurements.

A good update on the chemistry of porous Si can be found at Sailor's web site: http://chem-faculty.ucsd.edu/sailor/research/porous_Si_intro.html.

Problems Associated with Wet Bulk Micromachining of Single-Crystal Silicon

Introduction

Despite the introduction of more controllable etch-stop techniques, bulk micromachining of Si remains a difficult industrial process to control. It is also not an applicable submicrometer technology because wet chemistry is not able to etch reliably on that scale. For submicrometer structure definition, dry etching is required (dry etching is also more environmentally safe). We will now look into some of the other problems associated with Si bulk micromachining, such as the extensive real estate consumption and difficulties in etching at convex corners, and detail current solutions to avoid, control, or alleviate them.

Problem: Extensive Real Estate Consumption

Introduction Bulk micromachining involves extensive real estate consumption. This quickly becomes a problem when making arrays of devices. Consider the diagram in Figure 4.81, illustrating two membranes created by etching through a <100> wafer from the back until an etch stop, e.g., a Si_3N_4 membrane, is reached. In creating two of these small membranes, a large amount of Si real estate is wasted, and the resulting device might become fragile.

Solution 1: Real Estate Gain by Etching from the Front One solution to limit the amount of Si real estate loss is to use thinner wafers, but this solution becomes impractical below 200 µm because such wafers often break during handling. A more elegant solution is to etch from the front rather than from the back. Anisotropic etchants will undercut a masking material by an amount dependent on the orientation of the wafer with respect to the mask. Such an etchant will etch any <100> silicon until a pyramidal pit is formed, as shown in Figure 4.82. These pits have sidewalls with a characteristic 54.7° angle with respect to the surface of a <100> silicon wafer because the delineated planes are {111} planes. This etch property makes it possible to form cantilever structures by etching from the front because the cantilevers will be undercut and eventually will be suspended over a pyramidal pit in the silicon. Once this pyramidal pit is completed, the etch rate of the {111} planes exposed is extremely slow and practically stops. Therefore, process sequences,

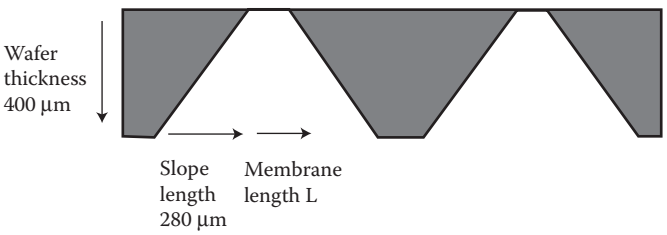


FIGURE 4.81 Two membranes formed in a <100>-oriented Si wafer.

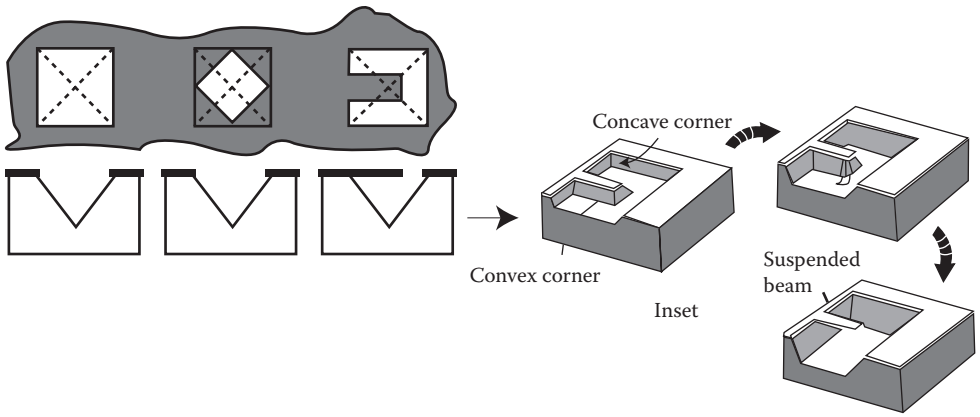


FIGURE 4.82 Three anisotropically etched pits etched from the front in a <100>-oriented silicon wafer. Inset, illustration of etching at convex corners for the formation of suspended beams.

which depend on achieving this type of a final structure, are very uniform across a wafer and very controllable. The top drawings in Figure 4.82 represent patterned holes in a masking material: a square, a diamond, and a square with a protruding tab. The drawings immediately below represent the etched pit in the silicon produced by anisotropic etching. Note in the first drawing that a square mask aligned with the <110> direction produces a four-sided pyramidal pit. In the second drawing, a similar mask shape oriented at 45° produces an etched pit that is oriented parallel to the pit etched in the first drawing independent of the mask orientation. In the second drawing, the corners of the diamond are undercut by the etchant as it produces the final etch pit. The third drawing illustrates that any protruding member is eventually undercut by the anisotropic etchant, leaving a cantilever structure suspended over an identical etch pit. The inset detailing the third drawing shows how undercutting at the convex corners of the cantilever, in pure KOH, is determined by {411} planes.[225]

Solution 2: Real Estate Gain by Using Silicon Fusion-Bonded Wafers Using silicon fusion-bonded (SFB) wafers rather than conventional wafers also makes it possible to fabricate much smaller microsensors and gain real estate. The process, introduced in the 1980s, is clarified in Figure 4.83 for the fabrication of a pressure sensor.[225] The bottom handle wafer has a standard thickness of 525 µm and is anisotropically etched with a square cavity pattern. Next, the etched handle wafer is fusion bonded to a top sensing wafer (the SFB process itself is detailed in Chapter 7, dealing with surface micromachining, and is also touched on in Volume III, Chapter 4, on packaging). The sensor wafer consists of a p-type substrate with an n-type epilayer corresponding to the required thickness of the pressure-transducing membrane. The sensing wafer is thinned all the way to the epilayer by electrochemical etching, and resistors are ion implanted. The handle wafer is ground and polished to the desired thickness. For gauge measurement, the anisotropically etched cavity is truncated by the polishing operation, exposing the back of the diaphragm (Figure 4.83a). For an absolute pressure sensor, the cavity is left enclosed (Figure 4.83b). With the same diaphragm dimensions and the same overall thickness of the chip, an SFB device, because of the inward sloping {111} walls, is almost 50% smaller than a conventional machined device (see Figure 4.84). For example, Lucas NovaSensor (Fremont, CA; now GE Novasensor, http://www.gesensing.com/novasensorproducts) manufactures a sensor that is 400 µm wide, 800 µm long, and 150 µm thick and fits inside the tip of a catheter.

Solution 3: Real Estate Gain by Using the Bosch Advanced Porous Silicon Membrane Process As we saw on page 293, the Bosch APSM process makes use of porous Si to form a cavity underneath the surface of a single-crystalline silicon membrane, which turns the traditional bulk micromachining process into a surface micromachining technology (see Figures 4.78–4.80). A piezoresistive pressure sensor made this way, the SMD 500, is the world's smallest digital pressure sensor (Figure 4.80).

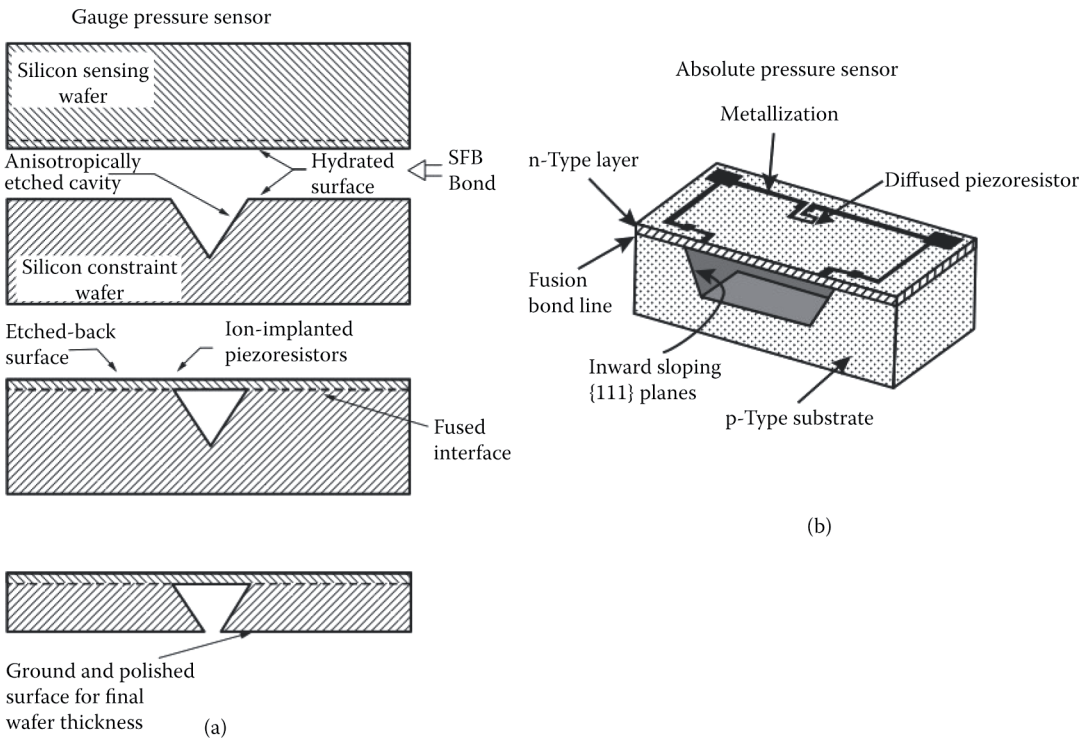


FIGURE 4.83 Fabrication process of SFB-bonded pressure sensors. (a) For gauge measurement, the anisotropically etched cavity is truncated by the polishing operation, exposing the back of the diaphragm. (b) For an absolute pressure sensor, the cavity is left enclosed. (From Bryzek, J., K. Petersen, J. R. Mallon, L. Christel, and F. Pourahmadi. 1990. *Silicon sensors and microstructures.* Fremont, CA: Novasensor. With permission.[226])

Problem: Underetching and Undercutting

Mask Underetching Underetching of a mask that contains no convex corners (that is, corners turning outside in) in principle stems from mask misalignment and/or from a finite etching of the {111} planes. Peeters measured the widening of {111}-walled V-grooves in a (100) Si wafer after etching in 7 M KOH at 80 ± 1° over 24 hours as 9 ± 0.5 µm.[126] The sidewall slopes of the V-groove are a well-defined 54.74°, and the etch rate R_{111} is related to the rate of V-groove widening R_v through:

$$R_{111} \frac{1}{2}\sin(54.74°)\, R_v \text{ or } R_{111} = 0.40R_v \qquad (4.13)$$

with R_{111} the etch rate of the (111) plane in nanometers per minute and R_v the groove widening, also in nanometers per minute. The V-groove widening experiment then results in an R_{111} of 2.55 ± 0.15 nm/min. In practice, this etch rate implies a mask underetching of only 0.9 µm for an etch depth of 360 µm. For a 1-mm-long V-groove and a 1° misalignment angle, a total underetching of 18 µm is theoretically expected, with 95% resulting from misalignment and only 5% caused by etching of the {111} sidewalls.[126] The total underetching will almost always be determined by misalignment rather than by etching of {111} walls.

Mask underetching with masks that include convex corners is much larger than the underetching

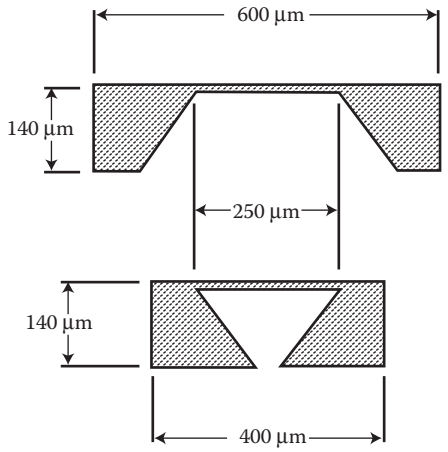


FIGURE 4.84 Comparison of conventional and SFB processes. The SFB process results in a chip that is at least 50% smaller than the conventional chip. (From Bryzek, J., K. Petersen, J. R. Mallon, L. Christel, and F. Pourahmadi. 1990. *Silicon sensors and microstructures.* Fremont, CA: Novasensor. With permission.[226])

just described because the etchant tends to circumscribe the mask opening with {111} walled cavities. This is usually called *undercutting* rather than underetching.

Undercutting—Convex Corner Attack It is advisable to avoid mask layouts with convex corners. However, mesa-type structures often are essential, and in that case there are two possible ways to reduce the undercutting. One is by chemical additives, reducing the undercut at the expense of a reduced ansiotropy ratio, and the other is by a special mask compensating the undercut at the expense of more lost real estate. When etching rectangular convex corners, deformation of the edges occurs because of undercutting. This is often an unwanted effect, especially in the fabrication of, e.g., acceleration sensors, where total symmetry and perfect 90° convex corners on the proof mass are mandatory for good device prediction and specification. The undercutting is a function of etch time and thus directly related to the desired etch depth. An undercut ratio is defined as the ratio of undercut to etch depth (δ/H).

Solutions: Additives and Corner Compensation Structures Saturating KOH solutions with isopropyl alcohol (IPA) reduces the convex corner undercutting; unfortunately, this happens at the cost of the anisotropy of the etchant. This additive also often causes the formation of pyramidical or cone-shaped hillocks.[126,141] Peeters claims that these hillocks are the result of carbonate contamination of the etchant, and he advises etching under inert atmosphere and stockpiling all etchant ingredients under an inert nitrogen atmosphere.[126]

Undercutting can also be reduced or even prevented by "corner compensation structures," which are added to the corners on the mask layout to be undercut during etching. Depending on the etching solution, different corner compensation schemes are used, among them square corner compensation (EDP or KOH) and rotated rectangle corner compensation (KOH) as illustrated in Figure 4.85.

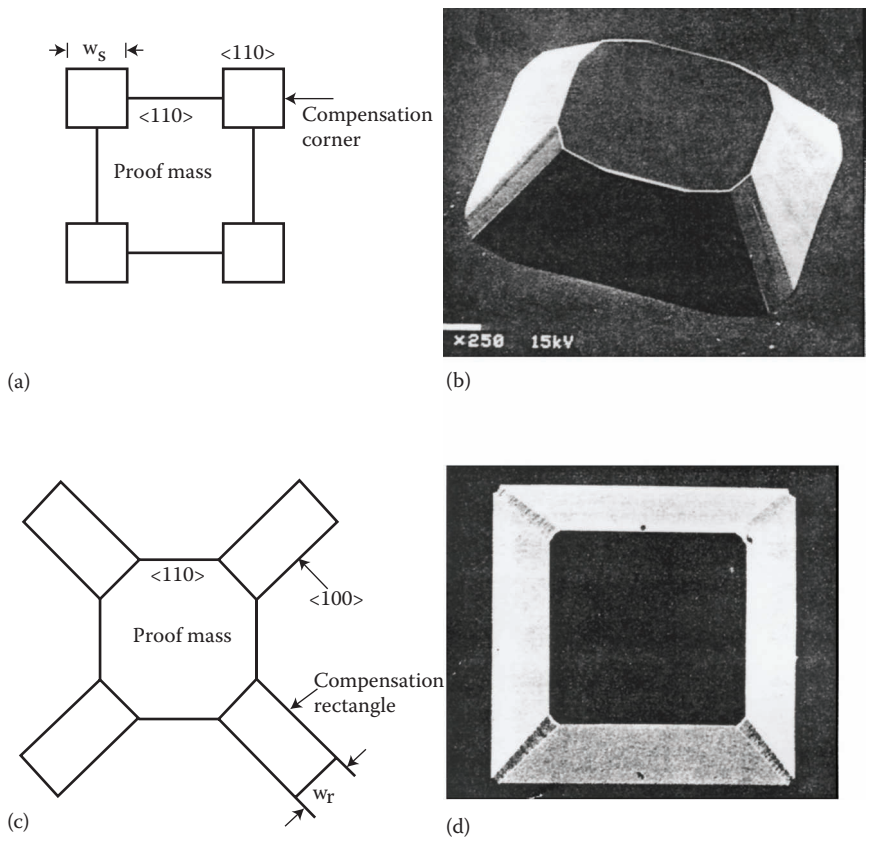


FIGURE 4.85 Formation of a proof mass by silicon bulk micromachining. (a and b) Square corner compensation method, using EDP as the etchant. (c and d) Rotated rectangle corner compensation method, using KOH as the etchant.

In square corner compensation, the square of SiO_2 in the mask, outlining the square proof mass feature for an accelerometer, is enhanced by adding an extra SiO_2 square to each corner (Figure 4.85a). Both the proof mass and the compensation squares are aligned with their sides parallel to the <110> direction. In this way, two concave corners are created at the convex corner to be protected. Thus, direct undercutting is prevented. The three "sacrificial" convex corners at the protective square are undercut laterally by the fast etching planes during the etch process. The dimension of the compensation square, w_s, depends on the depth of the required cavity; for example, for a 300-µm-deep cavity, a square with a side length of 300 µm is used. The resulting mesa structure after EDP or KOH etching is shown in Figure 4.85b. In rotated rectangle corner compensation, shown in Figure 4.85c, a properly scaled rectangle (w_r should be twice the depth of etching) is added to each of the mask corners. The four sides of the mesa square are still aligned along the <110> direction, but the compensation rectangles are rotated (45°) with their longer sides along the <100> directions. Using KOH as an etchant reveals the mesa shown in Figure 4.85d. A proof mass is frequently dislodged by simultaneously etching from the front and the back. Corner compensation requires a significant amount of real estate in the mask layout around the corners, making the design less compact, and the method is often only applicable for simple geometries.

Different groups using different corner compensation schemes all claim to have optimized spatial requirements. For an introduction to corner compensation, refer to Puers et al.[227] Sandmaier et al.[228] use <110>-oriented beams, <110>-oriented squares, and <010>-oriented bands for corner compensation during KOH etching. They found that spatial requirements for compensation structures could be reduced dramatically by combining several of these compensation structures. Figure 4.86 shows the mask layouts for some of the different compensation schemes they used. To understand the choice and dimensioning of these compensation structures, as well as those in Figure 4.85, we will first look at emerging planes at convex corners during KOH etching.

There is some disagreement in literature about the exact nature of planes that emerge at convex

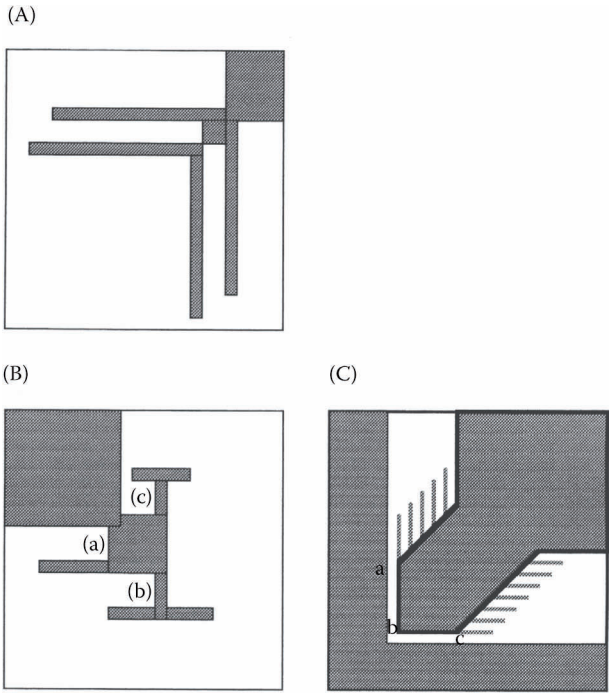


FIGURE 4.86 Mask layout for various convex corner compensation structures. (A) Splitting of the compensation beam creates concave corners, and by arranging two such double beams, a more symmetrical final structure is achieved. (B) Three convex corners of the compensation square are protected from undercutting by the added <110> beams. The side beam on corner b is made about 30% longer than the other two side beams. In this case, the corner is formed by the etch fronts starting at the corners a and b. (C) <010> bands that are added to convex corners in the <100> direction. (From Sandmaier, H., H. L. Offereins, K. Kuhl, and W. Lang. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91).* San Francisco. With permission.[228])

corners as etching progresses. Long et al.[229] in their work assumed {310} to be the dominant etch plane. Mayer et al.[225] found that the undercutting of convex corners in pure KOH etch is determined exclusively by {411} planes. However, the {411} planes of the convex underetching corner, as shown in Figure 4.87, are not entirely laid free; rugged surfaces, where only fractions of the main planes can be detected, overlap the {411} planes under a diagonal line shown as AB in Figure 4.87b. The ratio of {411} to {100} etching does not depend on temperature between 60 and 100°C. The value decreases with increasing KOH concentration from about 1.6 at 15% KOH to 1.3 above 40%, where the curve flattens out.[225] Ideally one avoids rugged surfaces and searches for well-defined planes bounding the convex corner. Figure

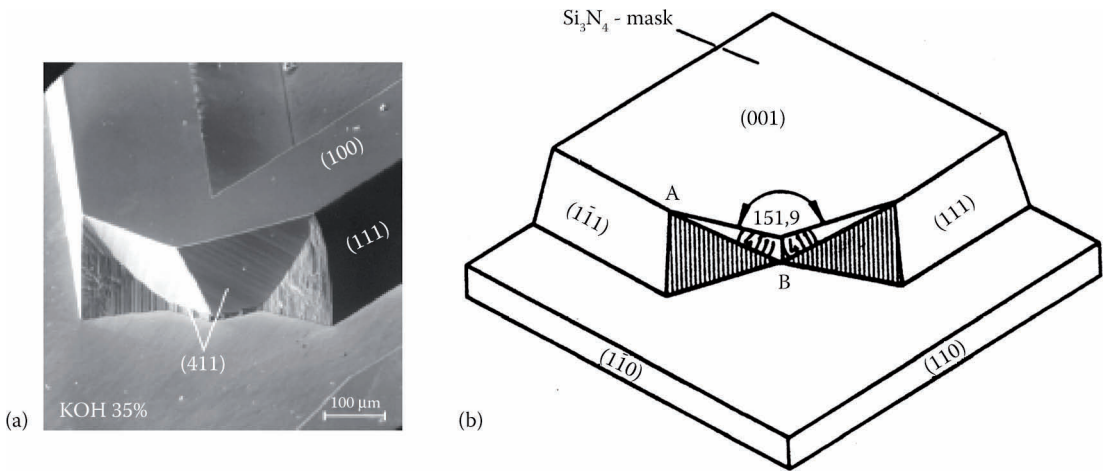


FIGURE 4.87 Planes occurring at convex corners during KOH etching of Si. (a) SEM picture. (b) Schematic. (From Mayer, G. K., H. L. Offereins, H. Sandmeier, and K. Kuhl. 1990. Fabrication of non-underetched convex corners in ansisotropic etching of (100)-silicon in aqueous KOH with respect to novel micromechanic elements. *J Electrochem Soc* 137:3947–51. With permission.[225])

4.88 shows how a <110> beam is added to the convex corner to be etched. Fast-etching {411} planes, starting at the two convex corners, laterally underetch a <110>-oriented beam (dashed lines in Figure 4.88). The longer this <110>-oriented beam is, the longer the convex corner will be protected from undercutting. It is essential that the beam disappear by the end of the etch to maintain a minimum of rugged surface at the convex edge. On the other hand, as is obvious from Figure 4.88, a complete disappearance of the beam leads to a beveling at the face of the convex corner. The dimensioning of the compensating <110> beam then works as follows: the length of the compensating beam is calculated primarily from the required etch depth (H) and the

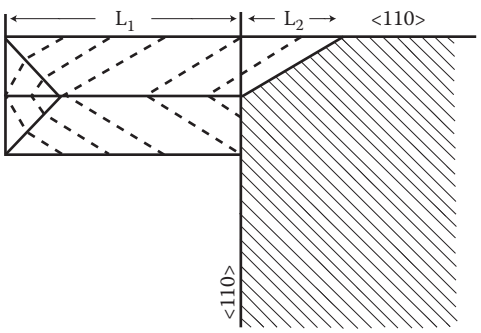


FIGURE 4.88 Dimensioning of the corner compensation structure with a <110>-oriented beam. (From Sandmaier, H., H. L. Offereins, K. Kuhl, and W. Lang. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91).* San Francisco. With permission.[228])

etch rate ratio $R\{411\}/R\{100\}$ ($\approx\delta/H$) at the concentration of the KOH solution used:

$$L = L_1 - L_2 = 2H\frac{R\{411\}}{R\{100\}} - \frac{B_{<110>}}{2\tan(30.9°)} \qquad (4.14)$$

with H the etch depth, $B_{<110>}$ the width of the <110>-oriented beam, and tan (30.9°) the geometry factor.[225] The factor 2 in the first term of this equation results because the etch rate of the {411} plane is determined normal to the plane and has to be converted to the <110> direction. The second term in Equation 4.15 takes into account that the <110> beam needs to disappear completely by the time the convex corner is reached. How long does etching go in {110} direction when it goes down in {100} direction by a certain depth?

The resulting beveling in Figure 4.88 can be further reduced by altering the etch compensation structures. This is done by decelerating the etch front, which largely determines the corner undercutting. One way to accomplish this is by creating more concave shapes right before the convex corner is reached. In Figure 4.86a, splitting of the compensation beam creates such concave corners, and by arranging two such double beams, a more symmetrical final structure is achieved. By using these split beams, the beveling at the corner is reduced by a factor of 1.4–2 and leads to bevel angles less than 45°.

Corner compensation with <110>-oriented squares (as shown in Figure 4.85a) features

considerably higher spatial requirements than the <110>-oriented beams. Because these squares are again undercut by {411} planes that are linked to the rugged surfaces described above, the squares do not easily lead to sharp {111}-defined corners. Dimensioning of the compensation square is done by using Equation 4.15 again, where L_1 is half the side length of the square; for $B_{<110>}$, the side length is used. All fast-etching planes have to reach the convex corner at the same time. As before, the spatial requirements of this compensation structure can be reduced if it is combined with <110>-oriented beams. Such a combination is shown in Figure 4.86b. The three convex corners of the compensation square are protected from undercutting by the added <110> beams. During the first etch step, the <110>-oriented side beams are undercut by the etchant. Only after the added beams have been etched does the square itself compensate the convex corner etching. The dimensioning of this combination structure is carried out in two steps. First, the <110>-oriented square is selected with a size that is permitted by the geometry of the device to be etched. From these dimensions, the etch depth corresponding to this size is calculated from Equation 4.15. For the remaining etch depth, the <110>-oriented beams are dimensioned like any other <110>-oriented beam. If the side beam on corner b is made about 30% longer than the other two side beams, the quality of the convex corner can be further improved. In this case, the corner is formed by the etch fronts starting at the corners a and b (Figure 4.86b).

A drawback to all the above-proposed compensation schemes is the impossibility of obtaining a sharp corner in both the top and the bottom of a convex edge as a result of the rugged surfaces associated with {411} planes. Buser et al.[230] introduced a compensation scenario where a convex corner was formed by two {111} planes, which were well defined all the way from the mask to the etch bottom. No rugged, undefined planes show in this case. The mask layout to create such an ideal convex corner uses <010> bands that are added to convex corners in the <100> direction (see Figure 4.86c). These bands will be underetched by vertical {100} planes from both sides. With suitable dimensioning of such a band, a vertically oriented membrane results, thinning and eventually freeing the convex edge shortly before the final intended etch depth is reached. In contrast to compensation structures undercut by {411} planes, posing problems with undefined rugged surfaces (see above), this compensation structure is mainly undercut by {100} planes. For the temperature range of 50–100°C and KOH concentrations ranging from 25–50 wt%, no undefined surfaces could be detected in the case of structures undercut by {100} planes.[228] The width of these <010>-oriented compensation beams, which determines the minimum dimension of the structures to etch, has to be twice the etching depth. These beams can either connect two opposite corners and protect both from undercutting simultaneously or be added to the individual convex corners (open beam). With an open-beam approach (see Figure 4.89), it is extremely important that the {100} planes reach the corner faster than the {411} planes. For that purpose the beams have to be wide enough to avoid complete underetching by {411} planes moving in from the front before they are completely underetched by {100} planes moving in from the side. For example, in a 33% KOH etchant a ratio between beam length and width of at least 1.6 is required. To make these compensation structures smaller while at the same time maintaining {100} undercutting to define the final convex corner, Sandmaier et al.[228] remarked that the shaping {100} planes do not need to be present at

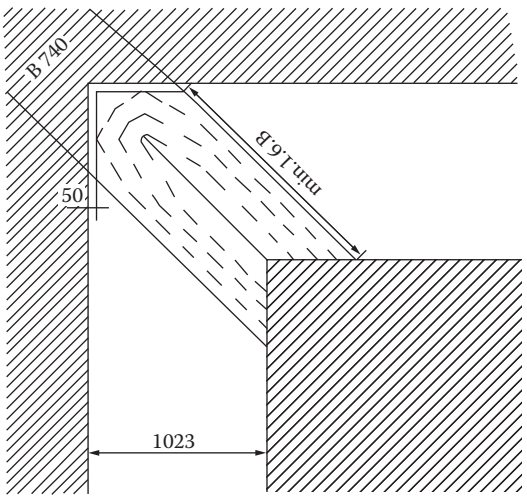


FIGURE 4.89 Beam structure open on one side. The beam is oriented in the <010> direction. The dimensions are in microns, and *B* is the width of the beam.

the beginning of the etching process. These authors implement delaying techniques by adding fan-like <110>-oriented side beams to a main <100>-oriented beam (see Figure 4.86c). As described above, these narrow beams are underetched by {411} planes and the rugged surfaces they entail until reaching the <100>-oriented beam. Then the {411} planes are decelerated in the concave corners between the side beams by the vertical {100} planes with slower etching characteristics. The length of the <110>-oriented side beams is calculated from:

$$L_{<110>} = \left(H - \frac{B_{<010>}}{2} \right) \frac{R\{411\}}{R\{100\}} \qquad (4.15)$$

with H being the etching depth at the deepest position of the device.

The width of the side beams does not influence the calculation of their required length. To avoid the rugged surfaces at the convex corner, the width of the side beams and the spaces between them should be kept as small as possible. For an etching depth of 500 µm, a beam width of 20 µm and a space width of 2 µm are optimal.[228] The biggest drawback of this compensation strategy is that the mask layout is rather complicated, and it takes a lot of time to generate the pattern to fabricate the masks themselves.

Modifications of the compensation scheme that uses <010> bands have been proposed to simplify the mask design. Zhang et al.[231] introduced a modified method that uses a <100> bar with a width greater than twice the etching depth, so that lateral etching would stop at <410> sidewalls, and improved results were obtained. Another novel structure was described by Enoksson[232] and involves a layout that consists of a <010> diagonal band connected to a concave corner on the opposite side of the groove (similar to the situation in Figure 4.89). The band has a slit in the middle of the concave corner, and according to Enoksson, this gives a near-perfect square corner.

Depending on the etchant, different planes are responsible for undercutting. From the above we learned that in pure KOH solutions, undercutting, according to Mayer et al.[225] and Sandmaier et al.,[228] mainly proceeds through {411} planes or {100} planes. That the {411} planes are the fastest undercutting planes was confirmed by Seidel.[233] At the wafer surface, the sectional line of a (411) and a (111) plane point in the <410> direction, forming an angle of 30.96° with the <110> direction, and it was in this direction that he found a maximum in the etch rate. In KOH and EDP etchants, Bean[131] identified the fast undercutting planes as {331} planes. For alkali/alcohol/water, Puers et al.[227] identified the fast underetching planes as {331} planes as well. Mayer et al.,[225] working with pure KOH, could not confirm the occurrence of such planes. Lee[112] indicated that in hydrazine/water, the fastest underetching planes are {211} planes. Abu-Zeid[234] reported that the main beveling planes are {212} planes in ethylenediamine-water solution (no added pyrocatechol). Wu et al.[235] found the main beveling planes at undercut corners to be {212} planes whether using KOH, hydrazine, or EPW solutions. In view of our earlier remarks on the sensitivity of etching rates of higher index rates to a wide variety of parameters (e.g., temperature, concentration, etching size, stirring, cation effect, alcohol addition, and complexing agent), these contradictory results are not too surprising. Along the same line, Wu et al.[236] and Puers et al.[227] have suggested triangles to compensate for underetching, but Mayer et al.[225] found them to lead to rugged surfaces at the convex corner. Combining a chemical etchant with more limited undercutting (IPA in KOH) with Sandmaier's reduced compensation structure schemes could further decrease the required size of the compensation features while retaining an acceptable anisotropy.

Corner compensation for <110>-oriented Si was explored by Ciarlo,[51] who comments that both corner compensation and corner rounding can be minimized by etching from both surfaces so as to minimize the etch time required to achieve the desired features. This requires accurate front-to-back alignment and double-sided polished wafers.

Using corner compensation offers access to completely new applications such as rectangular solids, orbiting V-grooves, truncated pyramids with low cross-sections on the wafer surface, bellow structures for decoupling mechanical stresses between micromechanical devices and their packaging, and so on.[228]

Some additional important references on corner compensation are Ted Hubbard's PhD thesis

(MEMS Design, Caltech 1994)[237] and references by Nikpour et al.[238] and Kim et al.[239]

Comparing Wet and Dry Etching of Si

For the construction of Si micromechanical devices, etching is a commonly used technique. Silicon can be etched in four principal ways: wet anisotropic, wet isotropic, dry anisotropic, and dry isotropic (see Figure 4.90). For wet anisotropic etching, aqueous KOH solutions are mostly used. Slow-etching (111) planes determine the shape of the etched structure. For dry etching, plasma etching machines are common tools. By varying the conditions, the shape of the channels can be varied from perfectly isotropic to exactly vertical. Positive or negative tapering is also possible.

In bulk micromachining, which requires more extreme topologies (more *z*-axis) than classical IC devices, wet etching of crystalline Si used to dominate the state of the art. The development of the newest fast etching, highly anisotropic DRIE equipment is changing this picture rapidly (see Chapter 3). Wet anisotropic etching of Si results in atomically smooth planes and atomically sharp edges, properties hard to achieve with dry etching. Wet etching also offers the advantage of low-cost batch fabrication—usually 20–25 wafers at a time. Some dry etching advantages include the elimination of handling of dangerous acids and solvents and the use of small amounts of chemicals that can provide both isotropic or anisotropic etch profiles, i.e., directional etching without relying on the crystal orientation of Si. Disadvantages of RIE etching are the fact that expensive equipment must be used, and the etching recipes are machine-dependent and also depend on the etching history of the machine. The walls of the etched structures are also rough compared with wet (electro)chemically etched structures.

Dry etching already is predominantly used in IC manufacture, and with higher-aspect-ratio devices and better materials selectivity being attained, the same is happening in micromachining. Besides selectivity and aspect ratio, problems with and concerns about the slow dry etch rate, mask erosion, and high sensitivity to operating parameters are being addressed. Etch rates as high as 6 μm/min have become possible, and the operation is much more automated. However, the equipment is still very specialized and expensive and requires very clean lines to bring the various gases into the vacuum chamber.

Environmental issues dictated a switch from wet to dry etching, but a recent concern in this context is the search for alternative chemistries because chlorofluorocarbon (CFC) production was banned. According to the Montreal Protocol, all ozone-depleting CFCs had to be eliminated by the year 2000. This included etchants such as $CFCl_3$, CCl_2F_2, CF_3Cl, C_2F_5Cl, and CF_3Br.[240] However, the most decisive reason for a switch to dry etching was the need for better CD control.

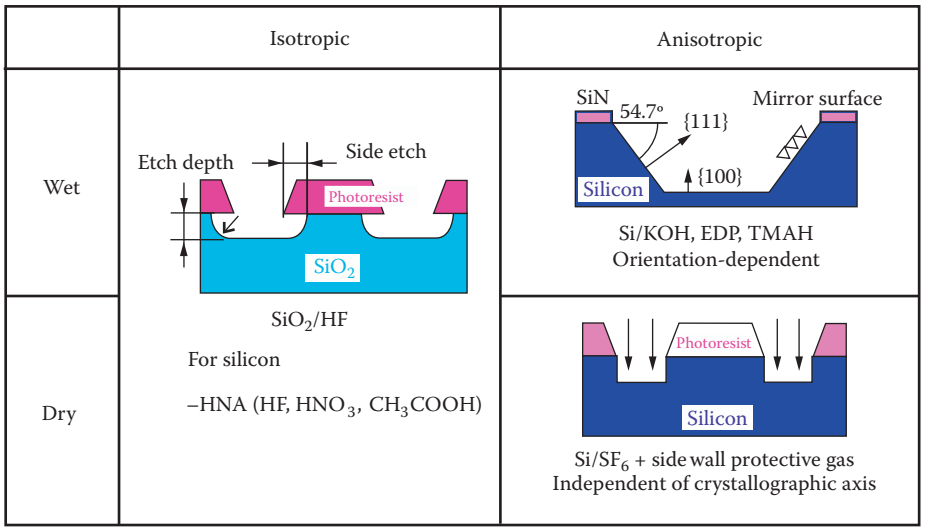


FIGURE 4.90 Different means for etching Si.

TABLE 4.13 Comparison of Wet versus Dry Etching Techniques

Parameter	Dry Etching	Wet Etching
Directionality	Can be highly directional with most materials (aspect ratio of 25 and above)	Only directional with single crystal materials (aspect ratio of 100 and above)
Production-line automation	Good ease of automation (e.g., cassette loading)	Poor
Environmental impact	Low but some gases are toxic and corrosive	High
Masking film adherence	Not as critical	Very critical
Cost chemicals	Low	High
Selectivity	Poor	Can be very good
Materials that can be etched	Only certain materials can be etched (e.g., not Fe, Ni, Co)	All
Radiation damage	Can be severe	None
Process scale-up	Difficult	Easy
Cleanliness	Good under the right operational conditions. Redeposition of nonvolatile compounds might occur.	Good to very good
CD control	Very good (<0.1 μm) Faithfully transfers lithographically defined photoresist patterns into underlying layers. High resolution. Less undercutting	Poor
Equipment cost	Need for specialized (expensive) equipment	Inexpensive
Submicron features	Applicable	Not applicable
Typical etch rate	Slow (0.1 μm/min)	Fast (1 μm/min, anis.)
Theory	Very complex, not well understood	Better understood (Volume I, Chapter 4)
Operating parameters	Many	Few
Control of etch rate	Good because of slow etching	Difficult

Surface micromachining (see Chapter 7) relies on both wet and dry etchants, with processes similar to the ones used in the IC industry. For example, wet etchants are used to etch away sacrificial layers (e.g., HF to etch CVD phosphosilicate glass), whereas dry etching helps to define polysilicon structural elements (e.g., in an NF_3-O_2 plasma). A combination of wet and dry techniques presently makes up the majority of Si micromachining work. Table 4.13 compares the characteristics of dry versus wet etching techniques.

Example 4.1: Dissolved Wafer Process

Figure 4.91 illustrates the dissolved wafer.[241] This process, used by Draper Laboratory for the fabrication of low-cost inertial sensors, involves a sandwich of a silicon sensor anodically bonded to a glass substrate.[242,243] The preparation of the silicon part requires only two masks and three processing steps. A recess is KOH-etched into a p-type (100) silicon wafer (step 1 with mask 1), followed by a high-temperature boron diffusion (step 2 with no mask). In step 1, reactive ion etching (RIE) may be used as well. Cho claims that by maintaining a high-temperature uniformity in the KOH etching bath (±0.1°C), the accuracy and absolute variation of the etch across the wafer, wafer-to-wafer and lot-to-lot, can be maintained to <0.1 μm using premixed, 45 wt% KOH.[241] Cho also uses low-defect oxidation techniques (e.g., nitrogen annealing and dry oxidation) to form

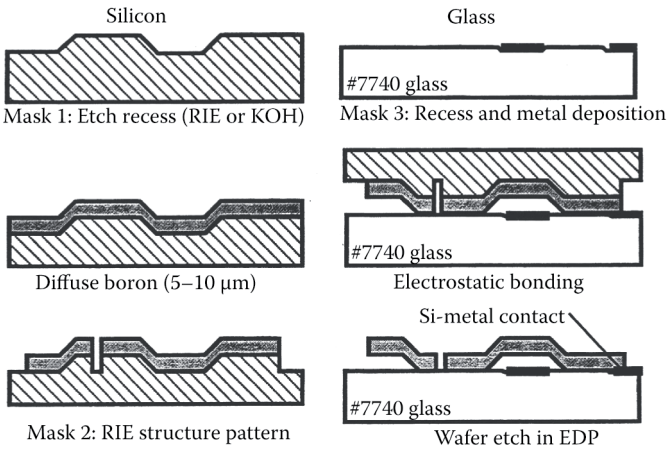


FIGURE 4.91 Dissolved wafer process. (From Greiff, P. 1995. *Micromachining and microfabrication process technology. Proceedings of the SPIE.* Austin, TX: SPIE. With permission.[242])

defect-free silicon surfaces. In the boron diffusion, the key is optimizing the oxygen content. In general, the optimal flow of oxygen is on the order of 3–5% of the nitrogen flow, in which case the doping uniformity is on the order of ±0.2 μm. Varying the KOH etch depth and the shallow boron diffusion time, a wide variety of operating ranges and sensitivities for sensors can be obtained. Next, the silicon is patterned for a RIE etch (step 3 with mask 2). Aspect ratios greater than 10 are accessible. Using some of the newest dry etching techniques, depths in excess of 500 μm at rates greater than 4 μm/min (with an SF_6 chemistry) are possible.[244] The glass substrate (#7740 Corning glass) preparation involves etching a recess, depositing, and, in a one-mask step, patterning a multimetal system of Ti/Pt/Au. The electrostatic bonding of glass to silicon takes place at 335°C with a potential of 1000 V applied between the two parts (electrostatic or anodic bonding is explained in detail in Volume III, Chapter 4). Commercial bonders have alignment accuracies on the order of <1 μm. The lightly doped silicon is dissolved in an EDP solution at 95°C. The keys to uniform EDP etching are temperature uniformity and suppression of etchant depletion through chemical aging or restricted flow (e.g., through bubbles). These effects can be minimized by techniques that optimize temperature control and reduce bubbling (e.g., proper wafer spacing, lower temperature, large bath). The structures are finally rinsed in deionized water and a hot methanol bath.

Draper Laboratory, although obtaining excellent device results with the dissolved wafer process, is now exploring an SOI process as an alternative. The latter yields an all-silicon device while preserving many of the dissolved wafer process advantages (see also Chapter 7 under "SOI Surface Micromachining").

Example 4.2: An Electrochemical Sensor Array Measuring pH, CO_2, and O_2 in a Dual Lumen Catheter

An electrochemical sensor array developed by the author is shown in Figure 4.92, packaged and ready for in vivo monitoring of blood pH, CO_2, and O_2. The linear electrochemical array fits inside a 20-gauge catheter (750 μm in diameter) without taking up so much space as to distort

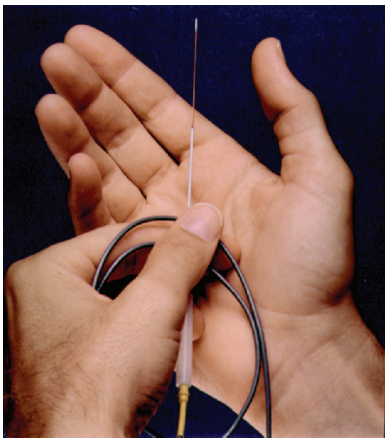


FIGURE 4.92 An in vivo pH, CO_2, and O_2 sensor based on a linear array of electrochemical cells.

the pressure signal monitored by a pressure sensor outside the catheter. A classical (macro) reference electrode, making ionic contact with the blood through the saline drip, was used for the pH signal, whereas the CO_2 and O_2 had their own internal reference electrodes. The high impedance of the small electrochemical probes makes a close integration of the electronics mandatory; otherwise, the high impedance connector leads, in a typical hospital setting, act as antennas for the surrounding electronic noise. As can be seen from the computer-aided design (CAD) picture in Figure 4.93, the thickness of the sensor comes from two silicon pieces: the top piece containing electrochemical cells and the bottom piece containing the active electronics. Each wafer is 250 μm thick. The individual electrochemical cells are etched anisotropically into the top

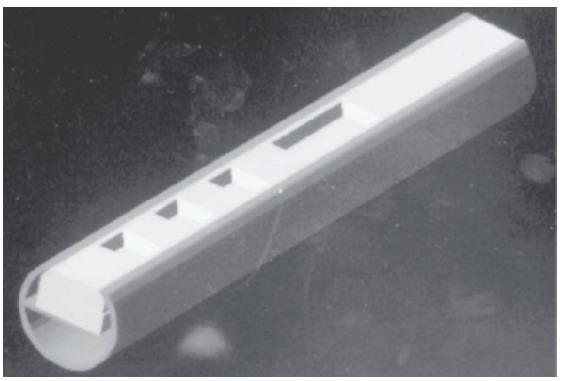


FIGURE 4.93 CAD of the electrochemical sensor array showing two pieces of silicon (each 250-μm thick) on top of each other, mounted in a dual-lumen catheter. The bottom part of the catheter is left open so pressure can be monitored and blood samples can be taken.

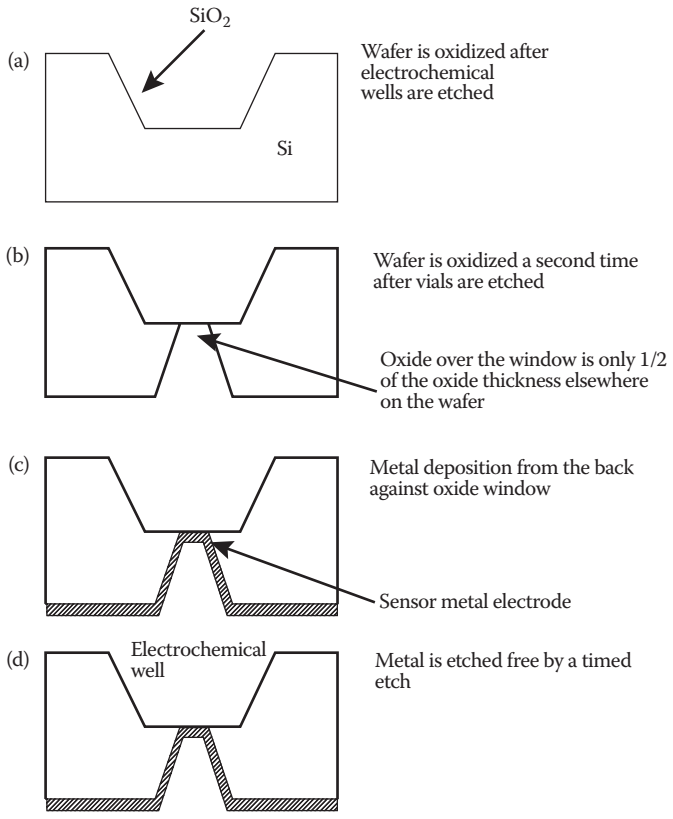


FIGURE 4.94 (a–d) Fabrication sequence for a generic electrochemical cell in Si. Depth of the electrochemical well, number of electrodes, and electrode materials can be varied.

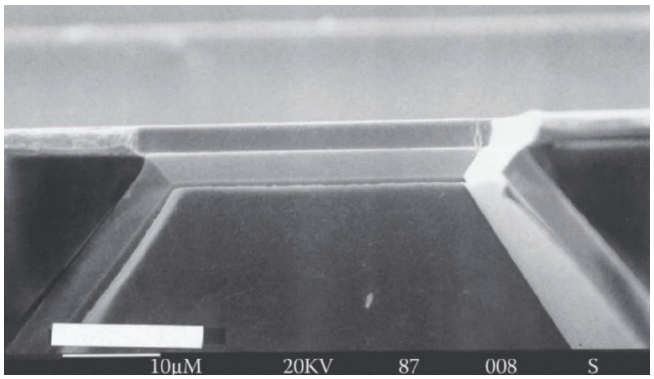


FIGURE 4.95 SEM micrograph illustrating process step (d) from Figure 4.94. A 30 × 30 μm Pt electrode is shown at the bottom of an electrochemical well. This Pt electrode is further contacted to the electronics from the back.

silicon wafer. The bottom piece is fabricated in a custom IC housing using standard IC processes. The process sequence to build a generic electrochemical cell in a 250-μm-thick silicon wafer with one or more electrodes at the bottom of each well is illustrated in Figure 4.94. Wells are etched from the front of the wafer, and after an oxidation step access cavities for the metal electrodes are etched from the back. The etching of the vias stops at the oxide-covered bottoms of the electrochemical wells (Figure 4.94a). Next, the wafers are oxidized a second time with the oxide thickness doubling everywhere except in the suspended window areas where no Si can feed further growth (Figure 4.94b). The desired electrode metal is subsequently deposited from the back of the wafer into the access cavities and against the oxide window (Figure 4.94c). Finally, a timed oxide etch removes the sacrificial oxide window from above the underlying metal, while preserving the thicker oxide layer in the other areas on the chip (Figure 4.94d).[245–247] An SEM micrograph demonstrating Figure 4.94d is shown in Figure 4.95. The electrodes in the electrochemical cells of the top wafer are further connected to the bottom wafer electronics by solder balls in the access vias of the top silicon wafer (see Figure 4.96). Separating the chemistry from the electronics in this way provides extra protection for the electronics from the electrolyte, as well as from the electronics, and chemical sensor manufacture can proceed independently. Depending on the type of sensor element, one or more electrodes are fabricated at the bottom of the electrochemical cells. For example, shown in Figure 4.97 is an almost completed (Severinghaus) CO_2 sensor with an Ag/AgCl electrode as the reference electrode (left in the figure) and an IrOx pH-sensitive electrode, both at the bottom of one electrochemical well. The metal electrodes are electrically isolated from each other by the SiO_2 passivation layer over the surface of the silicon wafer. To complete the CO_2 sensor, we silk-screen a hydrogel containing an electrolytic medium into the silicon sensor cavity and dip-coat the sensor into a silicone-polycarbonate rubber solution to form the gas-permeable membrane. For hydrogel inside the micromachined well, we use poly(2-hydroxyethyl)methacrylate or polyvinylalcohol.

The concept of putting the sensor chemistries and the electronics on opposite sides of a substrate is a very important design feature we decided on several years ago in view of the overwhelming problems encountered in building chemical sensors based on ion-sensitive field effect transistors or extended gate field effect transistors.[187]

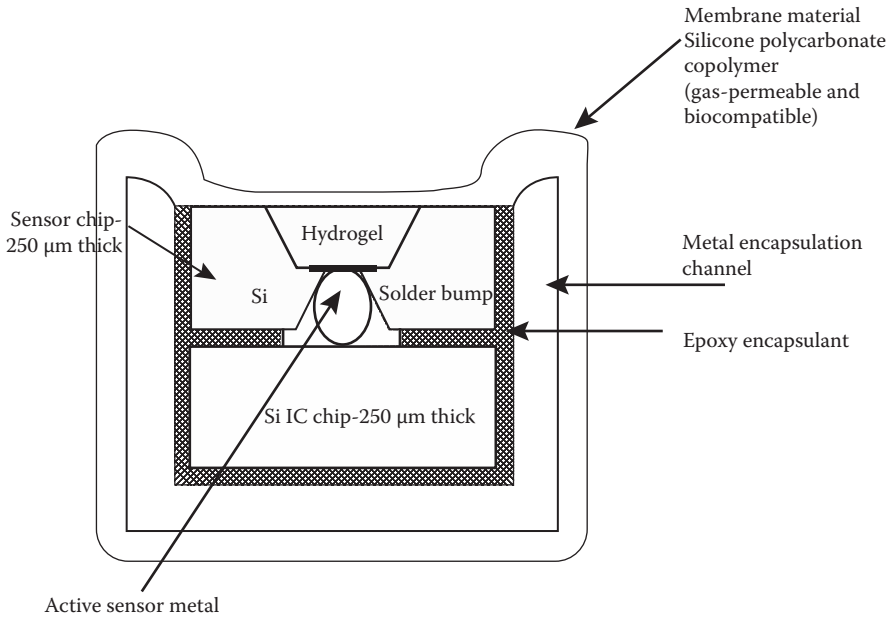


FIGURE 4.96 Schematic of the bonding scheme between sensor wafer and IC. The schematic is a cut-through of the catheter.

Example 4.3: Disposable Electrochemical Valves

One of the most difficult aspects of developing a microfluidic system is the miniaturization of valves. Most current MEMS valve technologies are still too complicated, too large, too power hungry, and too expensive for deployment in such applications as disposables (prohibitive cost) or implants (required power and size are too large). A more appropriate technology for a valve in a disposable such as a diagnostic panel for blood electrolytes or for an implantable drug delivery system may be a "sacrificial" or "use-once" valve. Such a valve could ensure device sterility, isolate reagents until they are required, and even entrap a vacuum to draw sample into the device (like a vacutainer blood-collection tube). Most current MEMS valves have moving parts, such as diaphragm valves that are micromachined in silicon, and are prone to malfunction through clogging. Valves without moving parts are obviously preferred, and some such valves have been incorporated in the electrokinetic and centrifugal fluidic platforms covered in Volume I, Chapter 6. These "valves" form a barrier for liquids but not for vapors. There is an even simpler valve, also without moving parts, which forms a barrier for both liquids and vapors. Creating a vapor barrier is essential if liquids are to be stored in small micromachined chambers for extended periods, e.g., for a diagnostic test incorporating different on-chip reagents. In the absence of such vapor barriers, liquid would be distributed over time driven by the gradients in vapor pressure above each individual chemical reservoir.

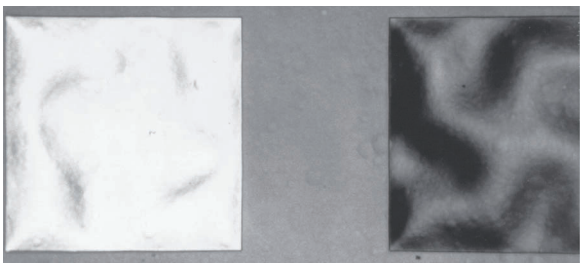


FIGURE 4.97 SEM micrograph of an Ag/AgCl (left) and IrO_x (right) electrode at the bottom of an anisotropically etched well in silicon. This electrochemical cell forms the basis for a Severinghaus CO_2 sensor.

In researching the structures described in Example 4.2 we discovered that a small current, applied between a counter electrode and a thin suspended metal electrode (see Figure 4.98), causes anodic dissolution of the metal or local electrolysis of water, depending on the nature of the valve metal. In both cases, the applied potential leads to bursting of the thin metal barriers. Thus, thin suspended metal membranes, such as silver or Au, can be burst open by passing a small current from the metal via a contacting electrolyte solution to a counter electrode. A Ag valve, for example, can be opened with an applied bias as low as 1–1.5 V. Although the "use-once"

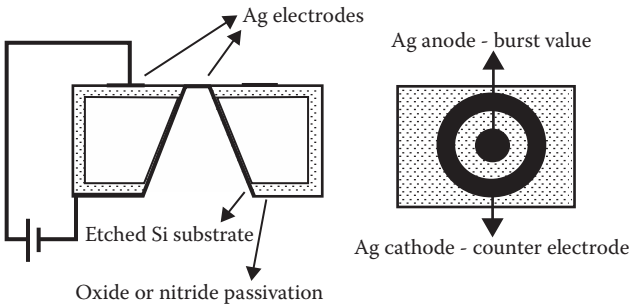


FIGURE 4.98 Disposable electrochemical valve: principle of operation.

microvalve in Figure 4.98 involves a Ag membrane suspended over an orifice, this patented electrochemical valve technology is generic, and a wide variety of metals may be used.[248] Small amounts of drugs can be kept in the Si micromachined chambers, and the chamber may be closed off using a polymeric laminate. The structure shown in Figure 4.98 has also been fabricated using a dry photoresist replacing the Si as structural material. Arrays of the element shown in Figure 4.98 can be made, and the individually addressable metal covers can then be opened electronically, releasing the drugs stored in the Si or polymer chambers. Different drugs can be released at different times, and by opening more "holes," the rate of drug delivery can be set. This is illustrated in Figure 4.99.[249]

One application for this type of valve is responsive drug delivery in a smart pill as sketched in Figure 4.100 (pharmacy-on-a-chip). In this type of implant, openings in a drug reservoir are regulated by a biological stimulus (e.g., the concentration of glucose sensed by a glucose biosensor); therefore, a patient receives only the amount of drug his/her body requires (e.g., the correct amount of insulin). A doctor may intervene via a telemetric link. This technology is currently being pursued by iGlyko (http://www.iglyko.com/iGlyko/Welcome_to_iGlyko.html).

Example 4.4: Self-Aligned Vertical Mirrors and V-Grooves for a Magnetic Micro-optical Matrix Switch

Hélin et al.[250] fabricated an elegant, monolithic magnetomechanical optical switch for dynamically reconfigurable dense wavelength division multiplexing networks. Arrays of optical switches allow for the rapid reconfiguration of optical

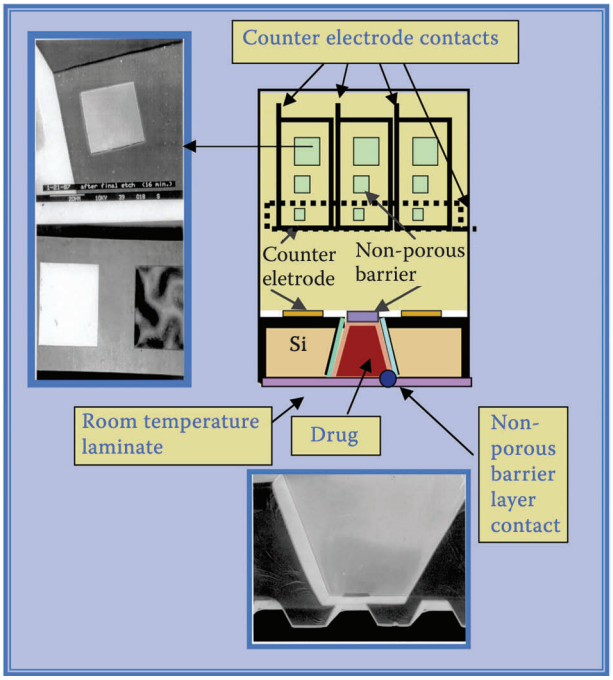


FIGURE 4.99 Metal electrodes are "blasted" open electrochemically. By making electrode arrays, the number of openings is selectable. The metal valve material may be Ag, Pt, Au, and so on. In the SEM picture on the left, we show a platinum electrode (top) and a set of an Ag/AgCl electrode and an IrO_x electrode (bottom). In the latter case the pH of the drug inside the reservoir can be monitored by measuring the voltage between the Ag/AgCl and IrO_x electrode.

networks by altering the light path in a system of intersecting fibers. A wavelength division multiplexer adds or deletes extra optical channels. A switch routing light from a single fiber into any of N output fibers is a $1 \times N$ switch. An $M \times N$ switch routes any one of M inputs to any one of N outputs. The $M \times N$ designation is the order of the optical switch. Commercially available optical switch arrays from companies such as JDS Uniphase Corp. (Milpitas, CA; http://www.jdsu.

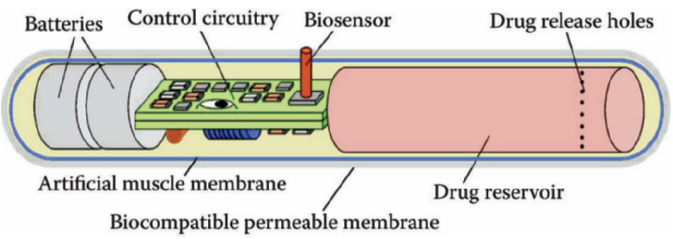


FIGURE 4.100 Model of responsive drug-delivery system (pharmacy-on-a-chip). Responsive drug-delivery pill-type configuration. A video on the operation of the device can be found at http://www.biomems.net/.

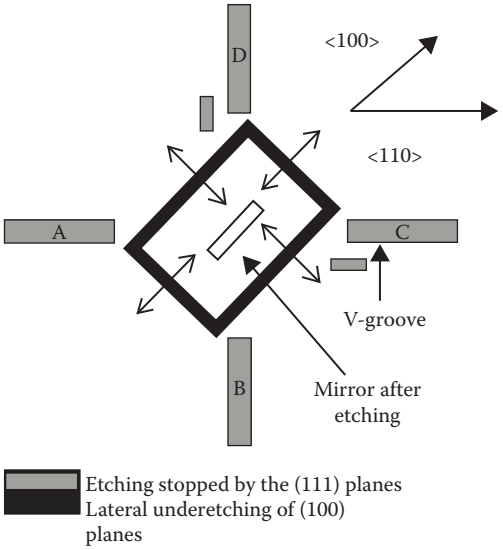


FIGURE 4.101 Principle of the self-aligned optical switch structure (arrows indicate the underetching).

com) and DiCon Fiberoptics, Inc. (Berkeley, CA; http://www.diconfiberoptics.com) are limited to 1×2 or 2×2.[251] Using MEMS, the prospect is to deliver 64×64 or even larger arrays in the case of Xros (bought by Nortel and then dissolved) (see also Volume III, Chapter 10 on IT applications of MEMS).

To make a low-cost, batch-machined switch, Hélin et al. use a simple, one-level mask process on a (100) silicon wafer, which allows for the simultaneous fabrication of (100) sidewalls for high quality mirrors and (111) V-grooves for self-alignment between V-grooves and mirrors. The etching principle, a nice illustration of the wet etching processes described in this chapter, is shown in Figure 4.101. The 45° angle between the <100> and <110> directions is used to self-align the vertical mirrors and V-grooves. As the

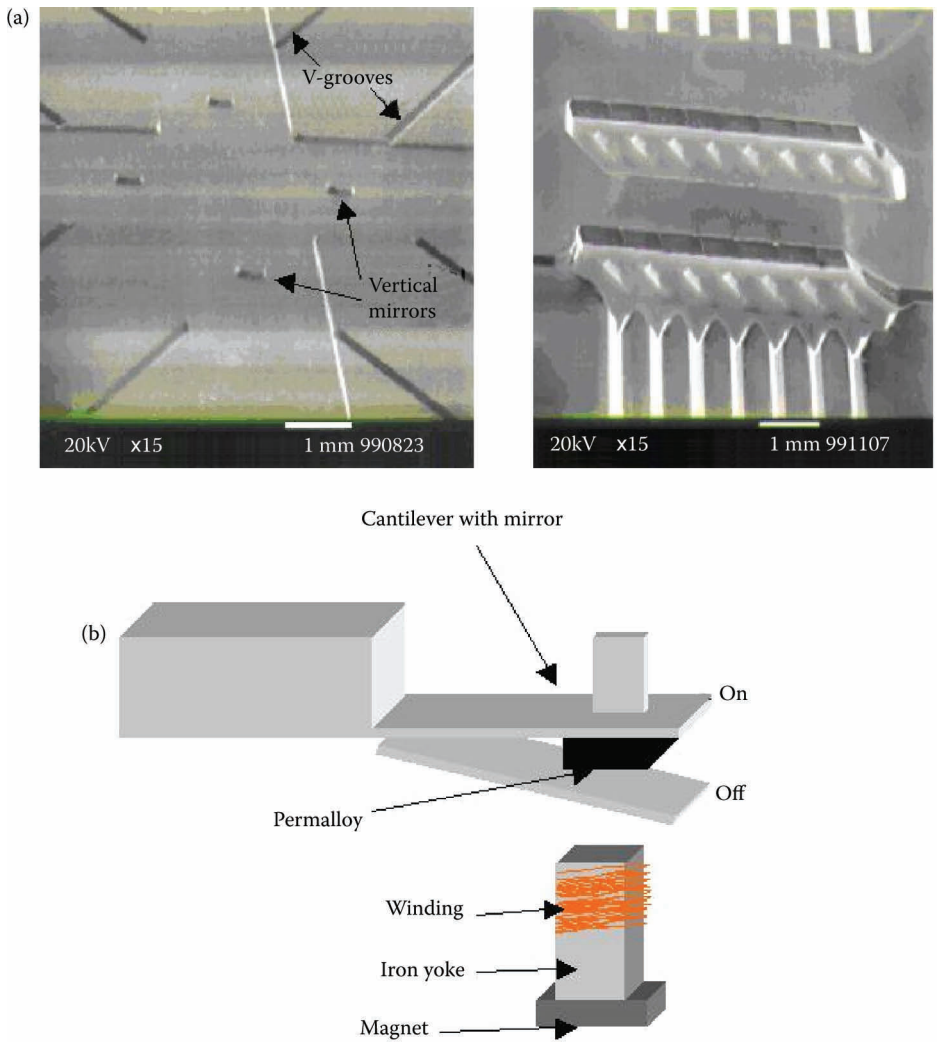


FIGURE 4.102 (a) Examples of optical switching structures for a self-aligned matrix switch; 2×2 bidirectional switch (left) and 1×8 (right). (b) Principle of switch operation with a self-latching system. (Courtesy of Dr. H. Fujita, University of Tokyo.)

bottom and sidewall planes are all from the same {100} family, the lateral underetch rate is equal to the vertical etch rate (see also Figure 4.102). Underetching in the <100> direction creates the upstanding mirror while at the same time V-grooves (A, B, C, and D) are formed in the <110> directions for optical fiber alignment. Contrary to DRIE-based processes, in which ripples etched in the vertical walls are inevitable, wet anisotropic etching fulfills the requirements of high surface quality mirrors, reducing optical losses. Because intersecting (111) planes stop the etching, the width of the V-groove in the mask layout fixes the position of the optical axis. The thickness of the mirror is determined by timing the etch. Cr/Au layers are deposited by vacuum evaporation to finish the mirror manufacture. The etching principle outlined in Figure 4.101 can be expanded to an $M \times N$ matrix switch as shown in Figure 4.102a for a 2×2 and 1×8 switch.

The front-side wafer etching process is preceded by a back-side etching of a cantilever beam, whose role is to support the mirror and allow it to move. The thickness of the cantilever support plate is determined by a timed etch, and the cantilever with mirror is actuated using electromagnetic actuation. The actuation principle is demonstrated in Figure 4.102b.

Acknowledgments

Special thanks to Mr. Chengwu Deng, graduate student UCI.

Questions

4.1: True or False. Give an explanation for your answer.
- (a) Sputtering technique yields a poor quality of SiO_2.
- (b) Etch selectivity of <110> over <111> for KOH is much higher than EDP.
- (c) EDP does not etch Si_3N_4.
- (d) KOH does not etch SiO_2.
- (e) Doping Si with high levels of boron provides a good etch stop for both EDP and KOH.
- (f) The lift-off technique is used mostly to pattern thick Si_3N_4 films.

4.2: A 0.6 μm film of silicon dioxide is to be etched with a buffered oxide etchant (BOE) with an etch rate of 750 Å/min. Process data show that the thickness may vary up to 10% and the etch rate may vary up to 15%.
- (a) Specify a time for the etch process.
- (b) Predict how much undercut will occur at the top of the film.

4.3: Using the isotropic etching diagram for silicon from H. Robbins and B. Schwartz, *J. Electrochem. Soc.*, 107, 108–11, 1960:
- (a) If HF: HNO_3: $HC_2H_3O_2$ = 70:10:20, what is the etch rate?
- (b) At this point, which component controls the etch rate?
- (c) If we use water (10%) instead of $HC_2H_3O_2$ (10%) as our diluent, the etch rate will be more sensitive to which components?
- (d) How can we significantly reduce the etch rate?

4.4: You are asked to build a 600 μm square, 25-μm thick diaphragm in an oxidized (100) silicon wafer. Given that the wafer thickness is 300 μm and that you will be using an anisotropic etching solution of KOH/water which attacks the (111) crystal plane 100 times slower than the (100) crystal direction, the etch rate of (100) silicon in this specific KOH/water solution is 10 μm/hr.
- (a) Find the etching time.
- (b) Find the dimensions of the mask opening you would use.

4.5: You are asked to make grooves 50 μm deep in an oxidized (100) silicon wafer. Use first a BOE etch to make an opening in the oxide layer and then an anisotropic etchant like KOH/water. What are the dimensions of the mask opening you would use to make this dimension insensitive to the amount of over-etch?

4.6: Write one paragraph on each of the Si etch stop techniques including electrochemical etch stop, p-n junction etch stop, and dopant-dependent etch stop, etc.

4.8: Explain in detail the various corner compensation schemes to reduce undercutting while etching anisotropically.

4.9: Which are the three orientations of silicon wafers most commonly employed in micromachining? Arrange them in order of most common usage and briefly describe their benefits and drawbacks.

4.10: Derive the relation that connects the size of an etched cavity with the size of the mask opening in the case of anisotropic etching of [100] silicon.

4.11: In a 4″ Si (100) wafer, with a thickness of 525 µm, 2 different etch features must be created. What should the mask opening be to create:

(a) A square orifice of width 200 µm that extends through the wafer?

(b) A V-groove that extends all the way through the wafer?

4.12: Compare the etching characteristics of (110)- and (100)-oriented silicon wafers. Which would you choose for the following applications and why?

(a) A diaphragm-based pressure sensor

(b) Closely spaced packed high-aspect-ratio comb actuators

4.13: Draw a mask with a convex and a concave corner and describe what happens when anisotropically etching with such a mask. What is the effect etch stop layer has on undercutting of a mask?

4.14: We are building a MEMS device using a <100>-oriented silicon substrate. The 4-inch wafer is p-type and 500 µm thick. There is a 200-nm-thick PECVD SiO_2 thin film deposited onto the surface. We use photolithography to pattern a 750-µm square onto the wafer backside, and while the photoresist is still in place, employ a 33% KOH bath at 70°C to etch the device (see also the figure below*).

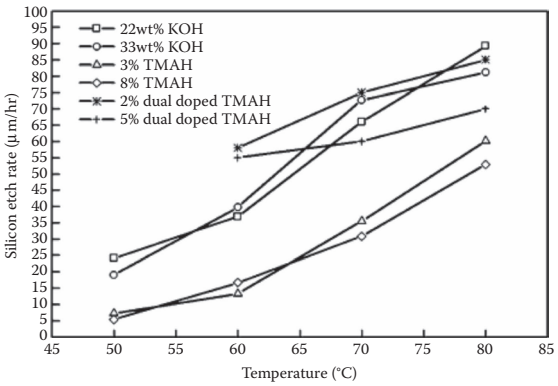
(a) What is the final size of the internal square projected onto the SiO_2 on the opposite side of the wafer?

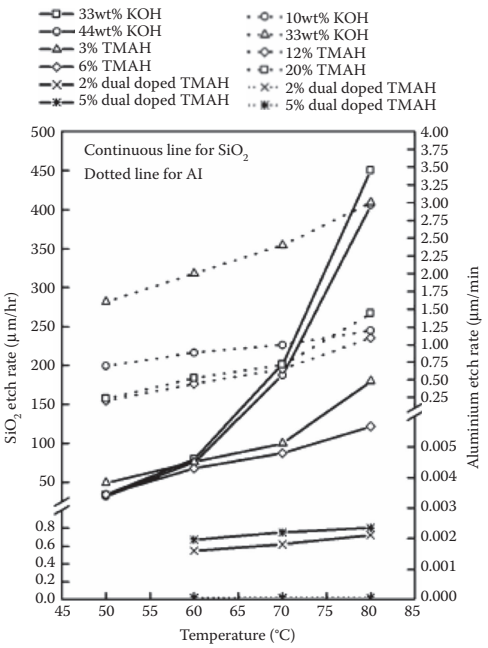
(b) What would have been the minimum photolithography square feature that will etch all the way through the wafer?

(c) How long will it take to etch through the SiO_2 layer (see also the figure below)?

(d) Once the SiO_2 layer is removed, how long will it take to etch through the Si layer?

(e) The etch rate of SiO_2 in buffered hydrofluoric acid (BHF) is 100 nm/min. How long will it take to etch the remaining oxide layers away from the device?





4.16: What wafer orientation would you select if you had to etch under a micro-bridge with a shallow V-groove crossing it at a 90° angle?

4.17: Compare wet and dry etching of Si.

4.18: Prepare a table with the most relevant design parameters for making a very sensitive

* Figures appearing under Question 4.14 reprinted from *Microelectronics Journal*, Volume 37, pp. 519–525, 2006. With kind permission from Elsevier.

piezoresistive Si cantilever to measure bending induced by differential surface stress.

4.19: With what you have learned so far, how would you make a very thin Si membrane in a Si frame and pattern a square Au electrode on that membrane inside the cavity? Where does your process differ from typical IC manufacture?

4.20: How would you make, using lift-off technology, an array of Pt electrodes at the bottom of a deep cavity and contact the Pt electrodes with leads extending over the cavity's edge?

4.21: Mark the wrong statements:

- The factors that affect the etching process are temperature, transport of reactant, reaction rate, diffusion rate, and so on.
- Acetic acid is more polar than water, which helps in achieving proper wetting of slightly hydrophobic Si wafers.
- Photon-pumped etching uses photogenerated electron-hole airs to supply oxidation sites for the etching process.
- Boron doped regions can be used as an etch stop layer.
- KOH etching is used for MOS and CMOS processing.

4.22: In the context of using piezoresistivity to sense strain, what are the advantages/disadvantages of using silicon vs. silicon nitride for pressure sensitive diaphragms? How does residual stress affect the sensitivity?

References

1. Harris, T. W. 1976. *Chemical milling.* Oxford: Clarendon Press.
2. Durant, W. 1957. *The reformation: a history of European civilization from Wyclif to Calvin, 1300–1564*. New York: MJF Books.
3. Robbins, H., and B. Schwartz. 1959. Chemical etching of silicon-I: the system, HF, HNO_3, and H_2O. *J Electrochem Soc* 106:505–08.
4. Robbins, H., and B. Schwartz. 1960. Chemical etching of silicon-II: the system HF, HNO_3, H_2O, and $HC_2C_3O_2$. *J Electrochem Soc* 107:108–11.
5. Schwartz, B., and H. Robbins. 1961. Chemical etching of silicon-III: a temperature study in the acid system. *J Electrochem Soc* 108:365–72.
6. Schwartz, B., and H. Robbins. 1976. Chemical etching of silicon-IV: etching technology. *J Electrochem Soc* 123:1903–9.
7. Uhlir, A. 1956. Electrolytic shaping of germanium and silicon. *Bell Syst Tech J* 35:333–47.
8. Hallas, C. E. 1971. Electropolishing silicon. *Solid State Technol* 14:30–32.
9. Turner, D. R. 1958. Electropolishing silicon in hydrofluoric acid solutions. *J Electrochem Soc* 105:402–08.
10. Stoller, A. I., and N. E. Wolff. 1966. *Proceedings: second international symposium on microelectronics.* Munich, Germany.
11. Stoller, A. I. 1970. The etching of deep vertical–walled patterns in silicon. *RCA Rev* 31: 271–75.
12. Forster, J. H., and J. B. Singleton. 1966. Beam–lead sealed junction integrated circuits. *Bell Lab Rec* 44:313–17.
13. Kenney, D. M. 1967. Methods of isolating chips of a wafer of semiconductor material. US Patent 3,332,137.
14. Lepselter, M. P. 1966. Beam lead technology. *Bell Sys Tech J* 45:233–54.
15. Lepselter, M. P. 1967. Integrated circuit device and method. 1967. US Patent 3,335,338.
16. Waggener, H. A. 1970. Electrochemically controlled thinning of silicon. *Bell Sys Tech J* 49:473–75.
17. Kragness, R. C., and H. A. Waggener. 1973. Precision etching of semiconductors. US Patent 3,765,969.
18. Waggener, H. A., R. C. Kragness, and A. L. Tyler. 1967. *Technical digest: IEEE international electron devices meeting 68.* Washington, DC: IEEE.
19. Waggener, H. A., R. C. Krageness, and A. L. Tyler. 1967. Two–way etch. *Electronics* 40:274.
20. Bean, K. E., and W. R. Runyan. 1977. Dielectric isolation: comprehensive, current and future. *J Electrochem Soc* 124:5C–12C.
21. Rodgers, T. J., W. R. Hiltpold, B. Frederick, J. J. Barnes, F. B. Jenné, and J. D. Trotter. 1977. VMOS memory technology. *IEEE J Solid-State Circuits* SC-12:515–23.
22. Rodgers, T. J., W. R. Hiltpold, J. W. Zimmer, G. Marr, and J. D. Trotter. 1976. VMOS ROM. *IEEE J Solid-State Circuits* SC-11:614–22.
23. Ammar, E. S., and T. J. Rodgers. 1980. UMOS transistors on (110) silicon. *IEEE Trans Electron Devices* ED-27:907–14.
24. Smith, C. S. 1954. Piezoresistance effect in germanium and silicon. *Phys Rev* 94:42–49.
25. Pfann, W. G. 1961. Improvement of semiconducting devices by elastic strain. *Solid State Electron* 3:261–67.
26. McGeough, J. A. 1974. *Principles of electrochemical machining.* New York: Wiley.
27. Romankiw, L. T., and T. A. Palumbo. 1987. *Proceedings: symposium on electrodeposition technology, theory and practice.* Eds. L. T. Romankiw and D. R. Turner. Pennington, NJ: The Electrochemical Society, Inc.
28. Stevens, G. W. W. 1968. *Microphotography: photography and photofabrication at extreme resolution.* New York: Wiley.
29. Humphrey, E. F., and D. H. Tarumoto, eds. 1965. *Fluidics.* Boston: Fluidic Amplifier Associates.
30. Allen, D. M. 1986. *The principles and practice of photochemical machining and photoetching.* Bristol and Boston: Adam Hilger.
31. Trotter, D. M. 1991. Photochromic and photosensitive glass. *Sci Am* 124:124–29.
32. Stookey, S. D. 2000. *Explorations in glass: an autobiography.* Hoboken, NJ: Wiley.
33. Vollmer, J., H. Hein, W. Menz, and F. Walter. 1994. Bistable fluidic elements in liga technique for flow control in fluidic microactuators. *Sensors Actuators A* A43:330–34.
34. Schomburg, W. K., J. Vollmer, B. Bustgens, J. Fahrenberg, H. Hein, and W. Menz. 1994. Microfluidic components in LIGA technique. *J Micromech Microeng* 4:186–91.
35. Ehrfeld W. 1994. Notes from handouts. Banff, Canada.

36. Phillips, R. E. 1985. Electrochemical grinding: What is it? How does it work? *SME Technical Paper* MR-85:383.
37. Boothroyd, G., and W. A. Knight. 1989. *Fundamentals of machining and machine tools.* New York: Marcel Dekker.
38. Sommer, C. 2000. *Non-traditional machining handbook.* Houston: Advance Publishing.
39. Datta, M. 1995. Interface. *The Electrochemical Society.* 4:32.
40. Datta, M., L. T. Romankiw, D. R. Vigliotti, and R. J. von Gutfeld. 1989. Jet and laser-jet electrochemical micromachining of nickel and steel. *J Electrochem Soc* 136:2251.
41. Bard, A. J. 1994. *Integrated chemical systems.* New York: Wiley.
42. Amemiya, S., A. J. Bard, F.-R. F. Fan, M. V. Mirkin, and P. R. Unwin. 2008. Scanning electrochemical microscopy. *Annu. Rev. Anal. Chem*. 1:95–131.
43. Landolt, D. 1994. *Proceedings of symposium on electrochemical microfabrication.* Eds. M. Datta,K. Sheppard, and J. O. Dukovic. Pennington, NJ: The Electrochemical Society.
44. Datta, M. J. 1995. Fabrication of an array of precision nozzles by through-mask electrochemical micromachining. *Electrochem Soc* 142:3801–05.
45. Madore, C., and D. Landolt. 1997. Electrochemical micromachining of controlled topographies on titanium for biological applications. *J Micromech Microeng* 7:270–75.
46. den Braber, E. T., J. E. den Ruijte, L. A. Ginsel, A. F. von Recum, and J. A. Jansen. 1996. Quantitative analysis of fibroblast morphology on microgrooved surfaces with various groove and ridge dimensions. *Biomaterials* 17:2037–44.
47. Gonsalvez, M. 2003. *Atomistic modelling of anisotropic etching of crystalline silicon.* PhD diss., Helsinki University of Technology.
48. Kendall, D. L. 1979. Vertical etching of silicon at very high aspect ratios. *Ann Rev Mater Sci* 9:373–403.
49. Tuckerman, D. B., and R. F. W. Pease. 1981. High-performance heat sinking for VLSI. *IEEE Electron Device Lett* EDL-2:126–29.
50. Barth, P. W., P. J. Shlichta, and J. B. Angel. 1985. *Third international conference on solid-state sensors and actuators.* Philadelphia.
51. Ciarlo, D. R. 1987. *Proceedings: IEEE micro robots and teleoperators workshop*. Hyannis, MA: IEEE.
52. Seidel, H., L. Csepregi, A. Heuberger, and H. Baumgartel. 1990. Anisotropic etching of crystalline silicon in alkaline solutions–part I: orientation dependence and behavior of passivation layers. *J Electrochem Soc* 137:3612–26.
53. Elwenspoek, M., H. Gardeniers, M. de Boer, and A. Prak. 1994. Micromechanics. Report No. 122830. Twente, the Netherlands: University of Twente.
54. Herr, E., and H. Baltes. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
55. Elwenspoek, M. 1993. On the mechanism of anisotropic etching of silicon. *J Electrochem Soc* 140:2075–80.
56. Elwenspoek, M., U. Lindberg, H. Kok, and L. Smith. 1994. *IEEE international workshop on micro electro mechanical systems (MEMS '94)*. Oiso, Japan: IEEE.
57. Kendall, D. L. 1975. On etching very narrow grooves in silicon. *Appl Phys Lett* 26:195–98.
58. Schumacher, A., H.-J. Wagner, and M. Alavi. 1994. Mit Laser und Kalilauge. *Technische Rundschau* 86:20–23.
59. Seidel, H., and L. Csepregi. 1988. Advanced methods for the micromachining of sensors. *Technical digest of the 7th sensor symposium*. Tokyo, Japan. Reprinted in MBB Jahresbericht 1989, Forschung und Entwicklung - Technisch-wissenschaftliche Veröffentlichungen 1989, S. 41.
60. Alavi, M., S. Buttgenbach, A. Schumacher, and H. J. Wagner. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
61. Alavi, M., S. Buttgenbach, A. Schumacher, and H. J. Wagner. 1992. Fabrication of microchannels by laser machining and anisotropic etching of silicon. *Sensors Actuators A* A32:299–302.
62. Kaminsky, G. 1985. Micromachining of silicon mechanical structures. *J Vac Sci Technol* B3:1015–24.
63. Stoller, A. I., R. F. Speers, and S. Opresko. 1970. A new technique for etch thinning silicon wafers. *RCA Rev* 31:265–70.
64. Kern, W., and C. A. Deckert. 1978. *Thin film processes.* Eds. J. L. Vossen, and W. Kern. Orlando, FL: Academic Press.
65. Yang, K. H. 1984. An etch for delineation of defects in silicon. *J Electrochem Soc* 131: 1140–45.
66. Chu, T. L., and J. R. Gavaler. 1965. Dissolution of silicon and junction delineation in silicon by the CrO_3-HF-H_2O system. *Electrochim Acta* 10:1141–48.
67. Archer, V. D. 1982. Methods for defect evaluation of thin <100> oriented silicon in epitaxial layers using a wet chemical etch. *J Electrochem Soc* 129:2074–76.
68. Schimmel, D. G., and M. J. Elkind. 1973. An examination of the chemical staining of silicon. *J Electrochem Soc* 125:152–55.
69. Secco d'Aragona, F. 1972. Dislocation etch for (100) planes in silicon. *J Electrochem Soc* 119:948–51.
70. Tuck, B. 1975. Review: the chemical polishing of semiconductors. *J Mater Sci* 10:321–39.
71. Kooij, E. S., K. Butter, and J. J. Kelly. 1999. Silicon etching in HF/HNO solution: charge 3 balance for the oxidation reaction. *Electrochem Solid-State Lett* 2:178–80.
72. Kovacs, G. T. A., N. I. Maluf, and K. E. Petersen. 1998. *Proceedings of the IEEE*. Washington, DC: IEEE.
73. Wise, K. D., M. G. Robinson, and W. J. Hillegas. 1981. Solid state processes to produce hemispherical components for inertial fusion targets. *J Vac Sci Technol* 18:1179–82.
74. Tjerkstra, R. W. 1999. *Isotropic etching of silicon in fluoride containing solutions as a tool for micromachining.* PhD diss., University of Twente, Twente, the Netherlands.
75. Petersen, K. E. 1982. Silicon as a mechanical material. *Proceedings of the IEEE*. 70:420–57.
76. Kuiken, H. K., J. J. Kelly, and P. H. L. Notten. 1986. Etching profiles at resist edges. *J Electrochem Soc* 133:1217–32.
77. Kern, W. 1978. Chemical etching of silicon, germanium, gallium arsenide, and gallium phosphide. *RCA Rev* 39:278–308.
78. Hu, S. M., and D. R. Kerr. Observation of etching of n-type silicon in aqueous HF solutions. *J Electrochem Soc* 114:414.
79. Hoffmeister, W. 1969. Determination of the etch rate of silicon in buffered HF using a ^{31}Si tracer method. *Int J Appl Radiation Isotopes* 2:139.
80. Muraoka, H., T. Ohashi, and T. Sumitomo. 1973. Controlled preferential etching technology. *Semiconductor Silicon*. Pennington, NJ: The Electrochemical Society, pp. 327–38.
81. Seidel, H. 1989. Mikromechanik. Ed. A. Heuberger. Heidelberg, Germany: Springer Verlag.
82. Bogenschutz, A. F., K.-H. Locherer, W. Mussinger, and W. Krusemark. 1967. Chemical etching of semiconductors in HNO_3-HF-CH_3COOH. *J Electrochem Soc* 114:970–73.
83. Dash, W. C. 1956. *J Appl Phys* 27:1193.
84. Sirtl, E., and A. Adler. 1961. *Z Metallkd* 52:529.
85. Jenkins, M. W. 1977. *J Electrochem Soc* 124:757.

86. Schimmel, D. G. 1979. Defect etch for <100> silicon evaluation. *J Electrochem Soc* 126:479–83.
87. Theunissen, M. J., J. A. Apples, and W. H. C. G. Verkuylen. 1970. Applications of preferential electrochemical etching of silicon to semiconductor device technology. *J Electrochem Soc* 117:959–65.
88. Meek, R. L. 1971. Electrochemically thinned N/N+ epitaxial silicon: method and applications. *J Electrochem Soc* 118:1240–46.
89. van Dijk, H. J. A., and J. de Jonge. 1970. Preparation of thin silicon crystals by electrochemical thinning of epitaxially grown structures. *J Electrochem Soc* 117:553–54.
90. Shengliang, Z., Z. Zongmin, and L. Enke. 1987. *Fourth international conference on solid-state sensors and actuators (Transducers '87)*. Tokyo, Japan.
91. Lehmann, V. 1993. The physics of macropore formation in low doped n-type silicon. *J Electrochem Soc* 140:2836–43.
92. Barla, K., G. Bomchil, R. Herino, and J. C. Pfister. 1984. X-ray topographic characterization of porous silicon layers. *J Cryst Growth* 68:721–26.
93. Arita, Y. 1978. Fomation and oxidation of porous silicon by anodic reaction. *J Cryst Growth* 45:383–92.
94. Higashi, G. S., Y. J. Chabal, G. W. Trucks, and K. Raghavachari. 1990. Ideal hydrogen termination of the Si (111) surface. *Appl Phys Lett* 56:656–68.
95. Allongue, P., V. Kieling, and G. Gerischer. 1995. Eching mechanism and atomic structure of H-Si (111) surfaces prepared in NH4F. *Electrochim Acta* 40:1353–60.
96. Gerischer, H., and M. Lubke. 1987. The electrochemical behaviour of n-type silicon (111)-surfaces in fluoride containing aqueous electrolytes. *Ber Bunsenges Phys Chem* 91:394–98.
97. Kooij, E. S., and D. Vanmaekelbergh. 1997. Catalysis and pore initiation in the anodic dissolution of silicon in HF. *J Electrochem Soc* 144:1296–301.
98. da Fonseca, C., F. Ozanam, and J.-N. Chazalviel. 1996. In situ infrared characterization of the interfacial oxide during the anodic dissolution of a silicon electrode in fluoride electrolytes. *Surface Sci* 365:1–14.
99. Matsumura, M., and S. R. Morrison. 1983. Anodic properties of n-Si and n-Ge electrodes in HF solution under illumination and in the dark. *J Electroanal Chem* 147:157–66.
100. Gerischer, H., and M. Lubke. 1988. Electrolytic growth and dissolution of oxide layers on silicon in aqueous solutions of fluorides. *Ber Bunsenges Phys Chem* 92:573–77.
101. Serre, C., S. Barret, and R. Herino. 1994. Characterization of the electropolishing layer during anodic etching of p-type silicon in aqueous HF solutions. *J Electrochem Soc* 141:2049–53.
102. Matsumura, M., and S. R. Morrison. 1983. Photoanodic properties of an n-type silicon electrode in aqueous solution containing fluorides. *J Electroanal Chem* 144:113–20.
103. Eddowes, M. J. 1990. Anodic dissolution of p- and n-type silicon: kinetic study of the chemical reaction. *J Electroanal Chem* 280:297–311.
104. Allongue, P., V. Costa-Kieling, and H. Gerischer. 1993. Etching of silicon in NaOH solutions: parts I and II. *J Electrochem Soc* 140:1009–18 (Part I); 1018–26 (Part II).
105. Palik, E. D., O. J. Glembocki, and J. I. Heard. 1987. Study of bias-dependent etching of Si in aqueous KOH. *J Electrochem Soc* 134:404–09.
106. Linde, H., and L. Austin. 1992. Wet silicon etching with aqueous amine gallates. *J Electrochem Soc* 139:1170–74.
107. Price, J. B. 1973. Anisotropic etching of silicon with KOH-H_2O- isopropyl alcohol. Eds. H. R. Huff, and. R. R. Burgess, *Semiconductor Silicon*. Princeton, NJ: The Electrochemical Society, pp. 339–53.
108. Finne, R. M., and D. L. Klein. 1967, A water-amine-complexing agent system for etching silicon. *J Electrochem Soc* 114:965–70.
109. Reisman, A., M. Berkenbilt, S. A. Chan, F. B. Kaufman, and D. C. Green. 1979. The controlled etching of silicon in catalyzed ethylene-diamine-pyrocathechol-water solutions. *Electrochem Soc* 126:1406–14.
110. Seidel, H., L. Csepregi, A. Heuberger, and H. Baumgartel. 1990. Anisotropic etching of crystalline silicon in alkaline solutions—part II: influence of dopants. *J Electrochem Soc* 137:3626–32.
111. Kendall, D. L., and G. R. de Guel. 1985. *Micromachining and micropackaging of transducers*. Ed. C. D. Fung. New York: Elsevier.
112. Lee, D. B. 1969. Anisotropic etching of silicon. *J Appl Phys* 40:4569–74.
113. Noworolski, J. M., E. Klaassen, J. Logan, K. Petersen, and N. Maluf. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
114. Waggener, H. A., and J. V. Dalton. 1972. Control of silicon etch rates in hot alkaline solutions by externally applied potentials. *J Electrochem Soc* 119:236C.
115. Weirauch, D. F. 1975. Correlation of the anisotropic etching of single-crystal silicon spheres and wafers. *J Appl Phys* 46:1478–83.
116. Clemens, D. P. 1973. Anisotropic etching of silicon on sapphire. *J Electrochem Soc* 407, 120:C240–C240.
117. Bean, K. E., R. L. Yeakley, and T. K. Powell. 1974. Orientation dependent etching and deposition of silicon. *J Electrochem Soc* 121:87C.
118. Declercq, M. J., J. P. DeMoor, and J. P. Lambert. 1975. A comparative study of three anisotropic etchants for silicon. *Electrochem Soc Abstracts* 75-2:446.
119. Wu, X. P., Q. H. Wu, and W. H. Ko. 1986. A study on deep etching of silicon using ethylene-diamine-pyrocathechol-water. *Sensors Actuators* 9:333–43.
120. Declercq, M. J., L. Gerzberg, and J. D. Meindl. 1975. Optimization of the hydrazine-water solution for anisostropic etching of silicon in integrated circuit technology. *J Electrochem Soc* 122:545–52.
121. Mehregany, M., and S. D. Senturia. 1988. Anisotropic etching of silicon in hydrazine. *Sensors Actuators* 13:375–90.
122. Asano, M., T. Cho, and H. Muraoko. 1976. Applications of choline in semiconductor technology. *Electrochem Soc Extend Abstr* 354:911–13.
123. Tabata, O., R. Asahi, and S. Sugiyama. 1990. Anisotropic etching with quaternary ammonium hydroxide solutions. *Technical Digest: Ninth Sensor Symposium*. Tokyo, Japan.
124. Pugacz-Muraszkiewicz, I. J., and B. R. Hammond. 1977. Application of silicates to the detection of flaws in glassy passivation films deposited on silicon substrates. *J Vac Sci Technol* 14:49–53.
125. Williams, K. R., and R. S. Muller. 1996. Etch rates for micromachining processes. *J Electrochem Soc* 137:3612–32.
126. Peeters, E. 1994. Process development for 3D silicon microstructures, with application to mechanical sensor design. Ph.D. diss., Catholic University of Louvain, Belgium.
127. Lambrechts, M., and W. Sansen. 1992. *Biosensors: microelectrochemical devices.* Philadelphia: Institute of Physics Publishing.

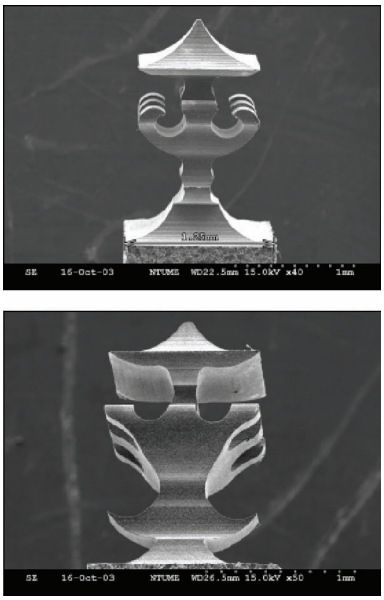
128. Palik, E. D., V. M. Bermudez, and O. J. Glembocki. 1985. Ellipsometric study of the etch-stop mechanism in heavily doped silicon. *J Electrochem Soc* 132:135–41.
129. Clark, L. D., and D. J. Edell. 19897. *Proceedings: IEEE micro robots and teleoperators workshop*. Hyannis, MA.
130. Schnakenberg, U., W. Benecke, and B. Lochel. 1990. NH_4OH-based etchants for silicon micromachining. *Sensors Actuators A* A23:1031–35.
131. Bean, K. 1978. Anisotropic etching of silicon. *IEEE Trans Electron Devices* ED–25:1185–93.
132. Barth, P. 1984. Si in biomedical applications. Microelectronics—photonics, materials, sensors and technology.
133. Linde, H., and L. Austin. 1992. Wet silicon etching with aqueous amine gallates. *J Electrochem Soc* 139:1170–74.
134. Tabata, O., R. Asahi, H. Funabashi, K. Shimaoka, and S. Sugiyama. 1992. Anisotropic etching of silicon in TMAH solutions. *Sensors Actuators A* A34:51–57.
135. Tabata, O. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
136. Reay, R. J., E. H. Klaassen, and G. T. A. Kovacs. 1994. Thermally and electrically isolated single-crystal silicon structures in CMOS technology. *IEEE Electron Device Lett* 15:309–401.
137. Steinsland, E., M. Nese, A. Hanneborg, R. W. Bernstein, H. Sandmo, and G. Kittilsland. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
138. Wise, K. D. 1985. *Micromachining and micropackaging of transducers.* Eds. C. D. Fung, P. W. Cheung, W. H. Ko, and D. G. Fleming. New York: Elsevier.
139. Seidel, H. 1987. The mechanism of anisotropic silicon etching and its relevance for micromachining. *Fourth international conference on solid-state sensors and actuators (Transducers '87)*. Tokyo, Japan.
140. Ternez, L. 1988. Ph.D. diss., Uppsala University.
141. Gravesen, P. 1986. *Silicon sensors. Status report for the industrial engineering thesis.* Lyngby, Denmark: DTH.
142. Ohwada, K., Y. Negoro, U. Konaka, and T. Oguchi. 1995. *Proceedings: IEEE micro electro mechanical systems conference.* Amsterdam, the Netherlands.
143. Baum, T., and D. J. Schiffrin. 1997. AFM study of surface finish improvement by ultrasound in the anisotropic etching of Si <100> in KOH for micromachining applications. *J Micromech Microeng* 7:338–42.
144. Bressers, P. M. M. C., J. J. Kelly, J. G. E. Gardeniers, and M. Elwenspoek. 1996. Surface morphology of p-type (100) silicon etched in aqueous alkaline solution. *J Electrochem Soc* 143:1744–50.
145. Klaassen, E. H., R. J. Reay, C. Storment, J. Audy, P. Henry, P. A. P. Brokaw, and G. T. A. Kovacs. 1996. *Proceedings: solid-state sensors and actuators workshop*. Hilton Head Island, SC.
146. Puers, R. 1991. *Proceedings: themadag sensoren*. Rotterdam, the Netherlands.
147. Buttgenbach, S. 1991. *Mikromechanik*. Stuttgart, Germany: Teubner Studienbucher.
148. Samaun, S., K. D. Wise, and J. B. Angell. 1973. An IC piezoresistive pressure sensor for biomedical instrumentation. *IEEE Trans Biomed Engr* 20:101–09.
149. Nunn, T., and J. Angell. 1975. *Workshop on indwelling pressure transducers and systems.* Cleveland, OH.
150. Greenwood, J. C. 1969. Ethylene diamine-cathechol-water mixture shows preferential etching of p-n junction. *J Electrochem Soc* 116:1325–26.
151. Bogh, A. 1971. Ethylene diamine-pyrocatechol-water mixture shows etching anomaly in boron-doped silicon. *J Electrochem Soc* 118:401–2.
152. Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press.
153. Heuberger, A. 1989. *Mikromechanik*. Heidelberg, Germany: Springer Verlag.
154. Rehrig, D. L. 1990. In search of precise epi thickness measurements. *Semicond Int* 13:90–95.
155. Palik, E. D., J. W. Faust, H. F. Gray, and R. F. Green. 1982. Study of the etch-stop mechanism in silicon. *J Electrochem Soc* 129:2051–59.
156. Jackson, T. N., M. A. Tischler, and K. D. Wise. 1981. An electrochemical P-N junction etch-stop for the formation of silicon microstructures. *IEEE Electron Device Lett* EDL–2:44–45.
157. Faust, J. W., and E. D. Palik. 1983. Study of the orientation dependent etching and initial anodization of Si in aqueous KOH. *J Electrochem Soc* 130:1413–20.
158. Kim, S. C., and K. Wise. 1983. Temperature sensitivity in silicon piezoresistive pressure transducers. *IEEE Trans Electron Devices* ED–30:802–10.
159. Kloeck, B., S. D. Collins, N. F. de Rooij, and R. L. Smith. 1989. Study of electrochemical etch-stop for high-precision thickness control of silicon membranes. *IEEE Trans Electron Devices* 36:663–69.
160. McNeil, V. M., S. S. Wang, K.-Y. Ng, and M. A. Schmidt. 1990. *Technical digest: 1990 solid state sensor and actuator workshop*. Hilton Head Island, SC.
161. Palik, E. D., O. J. Glembocki, and R. E. Stahlbush. 1988. Fabrication and characterization of Si membranes. *J Electrochem Soc* 135:3126–34.
162. Hirata, M., K. Suzuki, and H. Tanigawa. 1988. Silicon diaphragm pressure sensors fabricated by anodic oxidation etch-stop. *Sensors Actuators* 13:63–70.
163. Theunissen, M. J. 1972. Etch channel formation during anodic dissolution of n-type silicon in aqueous hydrofluoric acid. *J Electrochem Soc* 119:351–60.
164. Wen, C. P., and K. P. Weller. 1972. Preferential electrochemical etching of p+ silicon in an aqueous $HF–H_2SO_4$ electrolyte. *J Electrochem Soc* 119:547–48.
165. van Dijk, H. J. A. 1972. Method of manufacturing a semiconductor device and semiconductor device manufactured by said method. US Patent 3,640,807.
166. Saro, P. M., and A. W. van Herwaarden. 1986. Silicon cantilever beams fabricated by electrochemically controlled etching for sensor applications. *J Electrochem Soc* 133:1722–29.
167. Mlcak, R., H. L. Tuller, P. Greiff, and J. Sohn. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL: IEEE.
168. Yoshida, T., T. Kudo, and K. Ikeda. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92)*. Travemunde, Germany: IEEE.
169. Shockley, W. 1963. Method of making thin slices of semiconductive material. US Patent 3,096,262.
170. Peeters, E., D. Lapadatu, W. Sansen, and B. Puers. 1993. *Seventh international conference on solid-state sensors and actuators (Transducers '93)*. Yokohama, Japan.
171. Peeters, E., D. Lapadatu, R. Puers, and W. Sansen. 1994. Phet, an electrodeless photovoltaic electrochemical etch-stop technique. *J. Microelectromech. Syst.* 3(3):113–23.
172. Kelly, J. J., and H.G.G.Philipsen. 2005. Anisotropy in the wet-etching of semiconductors. *Curr Opin Solid State Mat Sci* 9:84–90.

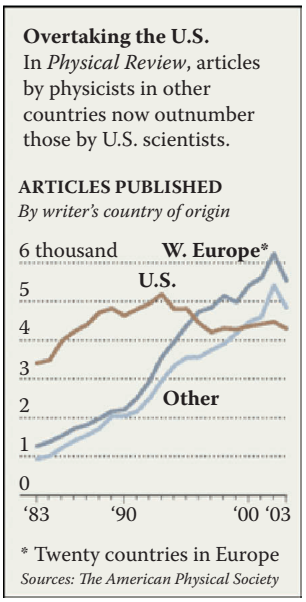
173. Baum, T., and D. J. Schiffrin. 1998. Mechanistic aspects of anisotropic dissolution of materials: etching of single-cystal in alkaline solutions. *J Chem Soc Far Trans* 94:691–94.
174. Hines, M. A., Y. J. Chabal, T. D. Harris, and A. L. Harris. 1994. Measuring the structure of etched silicon surfaces with Raman spectroscopy. *Chem Phys* 101:8055.
175. Palik, E. D., H. F. Gray, and P. B. Klein. 1983. A Raman study of etching silicon in aqueous KOH. *J Electrochem Soc* 130:956–59.
176. Raley, N. F., F. Sugiyama, and T. Van Duzer. 1984. (100) Silicon etch-rate dependence on boron concentration in ethylenediamine-pyrocatechol-water solutions. *J Electrochem Soc* 131:161–71.
177. Wu, X. P., Q. H. Wu, and W. H. Ko. 1985. *Third international conference on solid-state sensors and actuators (Transducers '85)*. Philadelphia.
178. Abu-Zeid, M. M., D. L. Kendall, G. R. de Guel, and R. Galeazzi. 1985. Abstract 275. *JECS* 400.
179. Seidel, H. 1990. *Technical digest: IEEE solid-state sensor and actuator workshop.* Hilton Head Island, SC.
180. Steinsland, E., M. Nese, A. Hanneborg, R. W. Bernstein, H. Sandmo, and G. Kittilsland. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
181. Ghandi, S. K. 1968. *The theory and practice of micro-electronics.* New York: Wiley.
182. Glembocki, O. J., R. E. Stahlbush, and M. Tomkiewicz. 1985. Bias-dependent etching of silicon in aqueous KOH. *J Electrochem Soc* 132:145–51.
183. Elwenspoek, M., and J. P. van der Weerden. 1987. Kinetic roughening and step free energy in the solid-on-solid model and on naphtalene crystals. *J Phys A Math Gen* 20:669–78.
184. Bennema, P. 1984. Spiral growth and surface roughening: developments since Burton, Cabrera, and Frank. *J Cryst Growth* 69: 182–97).
185. Jackson, K. A. 1966. Crystal growth. *International conference for crystal growth*. Boston, June 20–24, 1966.
186. Madou, M. J., and S. R. Morrison. 1989. *Chemical sensing with solid state devices.* New York: Academic Press.
187. Madou, M. J., K. W. Frese, and S. R. Morrison. 1980. Photoelectrochemical corrosion of semiconductors for solar cells. *SPIE* 248:88–95.
188. Madou, M. J., B. H. Loo, K. W. Frese, and S. R. Morrison. Bulk and surface characterization of the silicon electrode. *Surf Sci* 108:135–52.
189. Levy-Clement, C., A. Lagoubi, R. Tenne, and M. Neumann-Spallart. 1992. Photoelectrochemical etching of silicon. *Electrochim Acta* 37:877–88.
190. Lewerenz, H. J., J. Stumper, and L. M. Peter. 1988. Deconvolution of charge injection steps in quantum yield multiplication on silicon. *Phys Rev Lett* 61:1989–92.
191. Brantley, W. A. 1973. Calculated elastic constants for stress problems associated with semiconductor devices. *J Appl Phys* 44:534–35.
192. Hesketh, P. J., C. Ju, S. Gowda, E. Zanoria, and S. Danyluk. 1993. Surface free energy model of silicon anisotropic etching. *J Electrochem Soc* 140:1080–85.
193. Sundaram, K. B., and H.-W. Chang. 1993. Electrochemical etching of silicon by hydrazine. *J Electrochem Soc* 140:1592–97.
194. Morrison, S. R. 1980. *Electrochemistry on semiconductors and oxidized metal electrodes.* New York: Plenum Press.
195. Gaburro, Z., N. Daldosso, and L. Pavesi, eds. 2005. *Encyclopedia of condensed matter physics: porous silicon.* Amsterdam: Elsevier Ltd.
196. Canham, L. T. 1990. Silicon quantum wire array fabrication by electrochemical and chemical dissolution of wafers. *Appl Phys Lett* 57:1046–48.
197. Lehmann, V., and U. Gosele. 1991. Porous silicon formation: a quantum wire effect. *Appl Phys Lett* 58:856–58.
198. Richter, A., P. Steiner, F. Kozlowski, and W. Lang. 1991. Current-induced light emission from a porous silicon device. *IEEE Electron Device Lett* 12:691–92.
199. Barrett, E. P., L. G. Joyner, and P. P. Halenda. 1951. The determination of pore volume and area distributions in pure substances. *J Am Chem Soc* 73:373.
200. Brunauer, S., P. H. Emmett, and E. Teller. 1938. Adsorption of gases in multimolecular layers. *J Am Chem Soc* 60: 309–19.
201. Bomchil, G., R. Herino, and K. Barla. 1986. *Energy beam-solid interactions and transient thermal processes.* Eds. V. T. Nguyen, and A. Cullis. Les Ulis, France: Les Editions de Physique.
202. Watanabe, Y., Y. Arita, T. Yokoyama, and Y. Igarashi. 1975. Formation and properties of porous silicon and its applications. *J Electrochem Soc* 122:1351–55.
203. Tserepi, A., et al. 2003. Dry etching of porous silicon in high density plasmas. *Physica Status Solidi (A)* 197:163–67.
204. Smith, R. L., and S. D. Collins. 1990. Thick films of silicon nitride. *Sensors Actuators A* A23:830–34.
205. Unagami, T. 1980. Formation mechanism of porous silicon layer by anodization in HF solution.*J Electrochem Soc* 127:476–83.
206. Tjerkstra, R. W. 1999. Isotropic etching of silicon in fluoride containing solutions as a tool for micromachining. PhD diss., University of Twente, Twente, the Netherlands.
207. Jung, K. H., S. Shih, and D. L. Kwong. 1993. Developments in luminescent porous Si. *J Electrochem Soc* 140:3046–64.
208. Zhang, X. G. 1991. Mechanism of pore formation on n-type silicon. *J Electrochem Soc* 138:3750–56.
209. Lehmann, V. 1996. *Ninth annual international workshop on micro electro mechanical systems (MEMS '96)*. San Diego.
210. Grüning, U., V. Lehmann, S. Ottow, and K. Busch. 1996. Macroporous silicon with a complete two-dimensional photonic band gap centered at 5 mm. *Appl Phys Lett* 68:747–49.
211. Thönissen, M., and M. G. Burger, eds. 1997. *Properties of porous silicon*. Washington, DC: IEEE.
212. Parkhutik, V. P., L. K. Glinenko, and V. A. Labunov. 1983. Kinetics and mechanism of porous layer growth during n-type silicon anodization in HF solution. *Surf Technol* 20:265–77.
213. Lehmann, V., and H. Foll. 1990. Formation mechanism and properties of electrochemically etched trenches in n-type silicon. *J* Electrochem *Soc* 137:653–59.
214. Smith, L., S. F. Chuang, and S. D. Collins. 1988. *J Electron Mater* 17:533.
215. Christophersen, M., J. Carstensen, A. Feuerhake, and H. Föll. 2000. Crystal orientation and electrolyte dependence for macropore nucleation and stable growth on p-type-silicon. *Mater Sci Eng B* 69-70:194 .
216. Lehmann, V. 2006. *Porous semiconductors-science and technology 5th international conference.* Sitges, Spain.
217. Cahill, S. S., W. Chu, and K. Ikeda. 1993. *Seventh international conference on solid-state sensors and actuators (Transducers '93)*. Yokohama, Japan.
218. Föll, H., S. Langa, J. Carstensen, M. Christophersen, and I. M. Tiginyanu. 2003. Pores in III-V semiconductors. *Adv Mater* 15:183–98.

219. Searson, P. C., S. M. Prokes, and O. J. Glembocki. 1993. Luminescence at the porous silicon/electrolyte interface. *J Electrochem Soc* 140:3327–31.
220. Doan, V. V., and M. J. Sailor. 1992. Luminescent color image generation on porous Si. *Science* 256:1791–92.
221. Lin, V. S., K. Motesharei, K. S. Dancil, M. J. Sailor, and M. R. Ghadiri. 1997. A porous silicon-based optical interferometric biosensor. *Science* 278:840–43.
222. Lammel, G., and P. Renaud. 2000. Two mask tunable optical filter of porous silicon as microspectrometer. *Eurosensors XIV.* Eds. R. de Reus, and S. Bouwstra. Copenhagen: MIC-Mikroelektronik Center.
223. Canham, L. T. 1995. Bioactive silicon structure fabrication through nanoetching techniques. *Adv Mater* 7:1033.
224. Armbruster, S., F. Schafer, G. Lammel, H. Artmann, C. Schelling, H. Benzel, S. Finkbeiner, F. Larmer, R. Ruther, and O. Paul. 2003 A novel micromachining process for the fabrication of monocrystalline Si-membranes using porous silicon. *The 12th International Conference on Transducers, Solid-State Sensors, Actuators and Microsystems.*1:246–49.
225. Mayer, G. K., H. L. Offereins, H. Sandmeier, and K. Kuhl. 1990. Fabrication of non-underetched convex corners in ansisotropic etching of (100)-silicon in aqueous KOH with respect to novel micromechanic elements. *J Electrochem Soc* 137:3947–51.
226. Bryzek, J., K. Petersen, J. R. Mallon, L. Christel, and F. Pourahmadi. 1990. *Silicon sensors and microstructures.* Fremont, CA: Novasensor.
227. Puers, B., and W. Sansen. 1990. Compensation structures for convex corner micromachining in silicon. *Sensors Actuators A* A23:1036–41.
228. Sandmaier, H., H. L. Offereins, K. Kuhl, and W. Lang. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
229. Long, M. K., J. W. Burdick, and E. K. Antonsson. 1999. *MSM '99: modeling and simulation of microsystems, semiconductors, sensors and actuators.* San Juan, Puerto Rico.
230. Buser, R. A., and N. F. de Rooij. 1988. Monolithishes Kraftsensorfeld. *VDI-Berichte Nr*. 677.
231. Zhang, Q., L. L. Liu, and Z. Li. 1996. A new approach to convex corner compensation for anisotropic etching of (100) Si in KOH. *Sensors Actuators A* 56:251.
232. Enoksson, P. 1997. New structure for corner compensation in anisotropic KOH etching. *J Micromech Microeng* 7:141.
233. Seidel, H. 1986. PhD thesis. FU Berlin, Germany.
234. Abu-Zeid, M. M. 1984. Corner undercutting in anisotropically etched isolation contours. *J Electrochem Soc* 131:2138–42.
235. Wu, X.-P., and W. H. Ko. 1989. Compensating corner undercutting in anisotropic etching of (100) silicon. *Sensors Actuators* 18:207–15.
236. Wu, X., and W. H. Ko. 1987. *Fourth international conference on solid-state sensors and actuators (Transducers '87)*. Tokyo, Japan.
237. Hubbard, Ted J. 1994. *MEMS Design*. PhD thesis. Caltech.
238. Nikpour, B., L. M. Landsberger, T. J. Hubbard, M. Kahrizi, and A. Iftimie. 1998. Concave corner compensation between vertical (010)-(001) planes anisotropically etched in Si(100). *Sensors Actuators A* 66:299.
239. Kim, B., and D. C. Dong-II. 1998. Aqueous KOH etching of silicon (110): etch characteristics and compensation methods for convex corners. *J Electrochem Soc* 145:2499.
240. Mocella, M. T. 1991. The CFC-ozone issue in dry etch process development. *Solid State Technol* April:63–64.
241. Cho, S. T. 1995. Micromachining and microfabrication process technology II. *Proceedings of the SPIE*. Austin, TX: SPIE.
242. Greiff, P. 1995. Micromachining and microfabrication process technology. *Proceedings of the SPIE.* Austin, TX: SPIE.
243. Weinberg, M., J. Bernstein, J. Borenstein, J. Campbell, J. Cousens, B. Cunningham, R. Fields, P. Greiff, B. Hugh, L. Niles, and J. Sohn. 1996. *Micromachining and microfabrication process technology II*. Austin, TX: SPIE.
244. Craven, D., K. Yu, and T. Pandhumsoporn. 1995. *Micromachining and microfabrication process technology. Proceedings of the SPIE.* Austin, TX: SPIE.
245. Joseph, J., M. Madou, T. Otagawa, P. Hesketh, and A. Saaman. 1989. *Catheter-based sensing and imaging technology. Proceedings of the SPIE.* Los Angeles: SPIE.
246. Madou, M. J., and T. Otagawa. 1989. Microelectrochemical sensor and sensor array. US Patent 4,874,500.
247. Holland, C. E., E. R. Westerberg, M. J. Madou, and T. Otagawa. 1988. Etching method for producing an electrochemical cell in a crystalline substrate. US Patent 4,764,864.
248. Madou, M. J., and M. Tierney. 1994. Micro-electrochemical valves and methods. US Patent 5,368,704.
249. Seetharaman, S., H. Ke-Qin, and M. Madou. 2000. Microactuators toward microvalves for responsive controlled drug delivery. *Sensors Actuators* B67:149–60.
250. Hélin, P., T. Bourouina, H. Fujita, H. Maekoba, O. Cugat, and G. Reyne. 2000. *Nano et micro technologies.* Ed. D. Hauden. Paris: Hermes Science Publications.
251. Maluf, N. 2000. *An introduction to microelectromechanical systems engineering*. Boston: Artech House.

5

Thermal Energy-Based Removing





Pagoda (1.25 mm × 1.75 mm; material is Al-6061) produced by microelectrical discharge machining wire cutting (μ-EDM-WC) at the National Taiwan University (Prof. Y. S. Liao); see also Figure 5.10.

Outline

Introduction
Historical Note
Electrical Discharge Machining
Laser Beam Machining
Electron Beam Machining
Plasma Beam Machining
Questions
References

Introduction

In thermal removing processes, thermal energy, provided by a heat source, melts and/or vaporizes the volume of the material to be removed. Among thermal removal methods, electrical discharge machining (EDM) is the oldest and most widely used. Electron beam machining (EBM) and laser beam machining (LBM) are newer thermal techniques also widely accepted in industry today. Plasma arc cutting using a plasma arc torch is mostly used for cutting relatively thick materials in the range of 3–75 mm and is less pertinent to most miniaturization science applications. In thermal removal processes, a heat-affected zone (HAZ), sometimes called a *recast layer*, is always left on the work piece. In EBM, LBM, and arc machining, deposition and removal methods are available.

Additive thermal techniques involving EBM, LBM, and arc machining are covered in Chapter 9.

Historical Note

Among thermal methods, electrical discharge machining (EDM), using electrical sparks for cutting, is one of the oldest. In an "uncontrolled" fashion, EDM has existed ever since the first thunderstorm. The first widespread "controlled" application for shaping some of the toughest metals was developed during World War II. Although the EDM principle was discovered during the early 1940s, it was not until the early 1950s that the first practical EDM machine was designed. In the United States, during the 1960s, EDM machining was used to create features in the 75-µm range; even more precise EDM, using rotating spindles, was pioneered in Japan in the late 1980s (5-µm size range). There are two major types of EDM: wire EDM and sinker EDM. Sinker EDM is the older of the two. Wire EDM was introduced as a machining process in 1969. EDM, until 10 years ago, found limited application in micromachining[1]; given its very good accuracy, the method is now being investigated for a wide variety of miniaturization problems. Electron beam machining (EBM) and laser beam machining (LBM), two other examples of thermal techniques, were developed in the 1960s and 1980s, respectively. In EBM, high velocity electrons are used instead of sparks, and in LBM intense photon pulses are used. Some authors group EBM, LBM, and ion-beam machining together as high-energy beam machining. Ion-beam machining is a mechanical removing technique and as such is covered in Chapter 6 on mechanical energy-based removing techniques. We only touch briefly on plasma arc cutting because the method has only limited utility in miniaturization science.

Electrical Discharge Machining Introduction

Electrical discharge machining, more commonly known as EDM or spark machining, has traditionally been used to machine unusual designs in hard, brittle metals. Metal is removed by high-frequency electrical sparks, which are generated by pulsing a high voltage between the cathode tool and a work piece anode. The work piece and the tool are both submerged in a dielectric fluid as shown in Figure 5.1. Electrodes are usually made from graphite, copper, or silver. We cover two major types of EDM: wire EDM and sinker EDM.

Silicon-based machining often does not achieve the demands of applications like microsurgery, biotechnology, fluidics, or high temperature environments. As a consequence, EDM methods (including µ-EDM) are providing more and more solutions for application, such as fuel-injection nozzles for automobiles (regulations on environmental laws are forcing manufacturers to improve designs and making them smaller and more compact) and miniaturization of medical tools.

Sinker Electrical Discharge Machining

Sinker or cavity-type electrical discharge machining (EDM) is a thermal mass-reducing process that uses

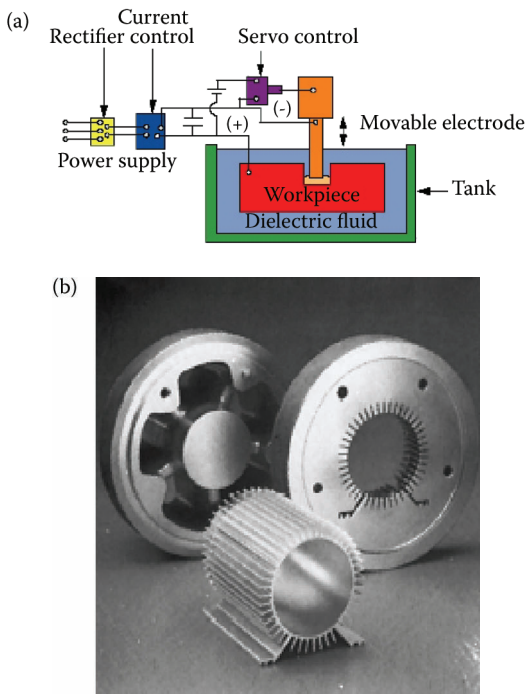


FIGURE 5.1 (a) Discharge machining. Schematic illustration of the electrical discharge machining (EDM) process. Process is based on erosion of metals by spark discharge. The cavity is formed by the shape of the electrode. (b) Examples of cavities produced by the EDM process using shaped electrodes. Two round parts (rear) are the set of dies for extruding the aluminum piece shown in front.

Table 5.1 Electrical Discharge Machining Characteristics

Typical Use	Hard, Machining of Brittle Metals, Tool Making
Tool	Carbon, zinc, brass, copper, silver-tungsten, or copper-tungsten
Dielectric medium	Distilled water (DI), petroleum oils, silicones, triethylene, glycol water mixtures
Aspect ratio of holes	As high as 100:1
Surface finish	1–3 μm but even 0.25 μm has been reported
Gap size/voltage	25 μm/80 V
Removal rate	0.001–0.1 cm^3/h
Work piece	Conductor

a shaped conductive tool (the cathode) to remove electrically conductive material (anode). It does this by means of rapid, controlled, repetitive spark discharges. A dielectric fluid is used to flush the removed particles, to regulate the discharge, and to keep the tool and work piece cool. With a gap between tool and work piece of about 25 μm and a voltage of, for example, 80 V, intense sparking occurs across the dielectric filled gap, melting and vaporizing metal from both pieces (ideally much more from the work piece than from the tool).

Resolidified small hollow spheres are washed away by the recirculating dielectric, for example, oil or deionized water (DI water). A servomechanism adjusts the work piece–tool gap, and DC current pulses produce sparks at a rate of up to 500,000/s. Each spark generates a localized high temperature on the order of 12,000°C in its immediate vicinity (the power concentration is 10^5–10^7 W/mm^2). The tool wear ratio, a measure of work piece material removal to tool material removal, varies from about 3:1 for metallic electrodes to about 70:1 for the better carbon electrodes. The tool wear can be reduced to 1% or less with an adequate selection of tool and work piece materials and appropriate generator settings.[2] Volumetric metal removal rates, ranging from 0.001–0.1 cm^3/h, are determined by thermal characteristics, such as conductivity and melting temperature of tool and work piece rather than hardness, and are small compared with more conventional machining methods. Metal removal rates increase as spark intensity (current) increases, but so does the surface roughness. Increasing the spark frequency, while keeping the other parameters constant, results in a decrease in surface roughness because the energy available is shared between more sparks, and smaller-sized surface craters are produced in the work piece. The frequency range of EDM machines is from 180 Hz, for rough cuts, to several hundred kilohertz, for fine finishing. Although sparks are effective for material removal, arcs result in bad surface quality. Arcs are created when the standby time is not long enough for the deionization of the dielectric to take place. Both short circuit and open circuit result in a stoppage of material removal. Thus, the goal of pulse control is to always create sparks and avoid arcs by optimizing the pulse width (on) and standby time (off). With carefully controlled conditions, it is possible to achieve tolerances as fine as 0.005 mm, but 0.02 mm is more typical. The accuracy of the process is closely related to the spark-gap width: the smaller the gap, the higher the accuracy, but a smaller gap results in a lower working voltage and a lower removal rate. A typical surface finish is in the 1–3-μm range. Two layers are of concern in EDM products. One is a layer of melted and resolidified material usually 0.0025–0.05-mm thick, known as *recast*, which is left on the surface produced by EDM. This layer tends to be very hard and brittle and may have to be removed if high levels of fatigue resistance are required (e.g., with shot peening*). A second layer is the so-called *heat-affected zone* (HAZ) that can be removed through electrochemical abrasion. In Table 5.1 some key EDM features are summarized.

With EDM, holes as small as 0.3 mm in materials 20 mm or more in thickness can be readily achieved. With efficient flushing, holes with an aspect ratio as high as 100:1 have been produced.[3]

* Shot peening is a process used as a decorative finish and/or to modify the mechanical properties of metals. It entails impacting a surface with shot (round metallic particles) with a force sufficient to create dimples and with enough shot so that those dimples overlap.

In general, EDM is one of the most widely used nonmechanical, noncontact machining processes in the industry today. Applications of this process include extrusion dies, fuel injector nozzles, aircraft engine turbine blades, and machining of difficult-to-machine materials like tool steel, tungsten carbide, and cemented carbides. EDM provides very good accuracy and repeatability; however, the tradeoff is a very low material removal rate (0.001–0.1 cm^3/h) for conductive and semiconductive materials only accompanied by thermal damage of the machined surface as a result of high heat generated during the discharge pulses (recast layer of 20–50 μm). Finally, the severe electrode wear affects machining accuracy, and the process poses some fume and skin hazards to operators.

An excellent starting point on the Internet for studying EDM in more detail is EDM Talk (http://www.edmmachining.com), the original EDM industry bulletin board and online forum. EDM equipment includes Fanuc's Robocut (http://www.fanuc.co.jp/en/product/robocut/index.html), Panasonic's MG-ED82W(http://www.panasonicfa.com/?id=MG-ED82W), and Agie's Agie® Mondo Star™ 50 CNC EDM (http://www.agie.com). Also check SmallTec International (http://www.smalltec.com) for sinker EDM and wire-cutting EDM (EDM-WC). A typical EDM machine may cost $350,000, and a typical service supplier is National Jet Company (Lavelle, MD), performing state-of-the-art micro-EDM (http://www.najet.com). The *EDM Handbook* by E. Bud Guitrau constitutes excellent further reading.[3]

A recent variant, EDM-WC, promises to make yet more impact on the future of microelectromechanical systems and is covered next.

Electrode Discharge Machining: Wire Cutting

One of the most interesting adaptations of EDM is the development of electrodischarge machining wire cutting (EDM-WC), also called *wire electrode discharge grinding* (WEDG), which can be used for cutting complex two- (2D) and three-dimensional (3D) shapes from electrically conducting materials.[4] EDM-WC, like EDM, is a thermal mass-reducing process that uses a continuously moving wire to remove material by means of rapid controlled repetitive spark discharges. In EDM-WC, the cutting electrode is a continuously moving thin copper or brass wire, 0.15–0.3 mm in diameter, or, for high-precision cutting, stronger molybdenum-steel wire (0.15–0.03 mm). The wire is pulled from a supply reel past two fixed sapphire or diamond guides, one located above and one below the work piece, as shown in Figure 5.2, and collected on a take-up reel. This continuously delivers fresh wire to the work area. As much as 50 hours of machining can be performed with one reel of wire, which is then discarded. Deionized water is used as dielectric and as a flushing fluid. Flushing is essential for achieving high accuracy and preventing short-circuiting between the wire and part. The main advantage of EDM-WC is that it can mill through metals four or five times faster than EDM. Moreover, no expensive, shaped EDM electrodes are required. Linear cutting rates are still relatively slow, especially for thick substrates: from 38–115 mm/h in 25-mm-thick steel.

Numerically controlled EDM-WC has revolutionized die making, particularly for plastic molders. EDM-WC is now common in tool-and-die shops. Shape accuracy in EDM-WC in a working environment with temperature variations of about 3°C is about 4 μm. If temperature control is within ±1°C,

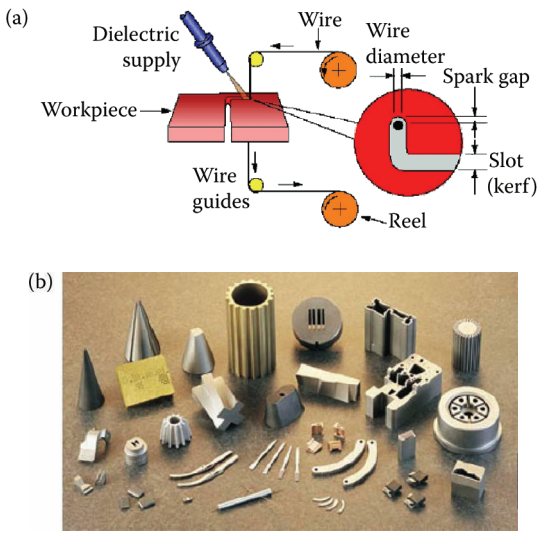


FIGURE 5.2 Electrical discharge machining, wire cutting (EDM-WC). (a) Schematic illustration of the wire EDM process. As much as 50 hours of machining can be performed with one reel of wire, which is then discarded. (b) Typical EDM-WC products.

the obtainable accuracy is closer to 1 µm.[5] No burrs are generated, and because no cutting forces are present, EDM-WC is ideal for delicate parts. No tooling is required, so delivery times are short. Pieces more than 16-in. thick can be machined. Tools and parts are machined after heat treatment; thus, dimensional accuracy is held and not affected by heat treat distortion.

Service and equipment providers in the EDM-WC arena include Small Tec International (http://www.smalltec.com), Reliable, EDM Inc. (http://www.reliableedm.com), and XACT (http://www.xactedm.com/index.html). Oki is a provider of brass wire for EDM-WC (OB wire) (http://www.okidensen.co.jp/english/e_home.htm). Equipment may be bought from companies such AGIE (http://www.agie.com), Fanuc (http://www.fanuc.co.jp/en/product/index.htm), and Panasonic (http://www.panasonicfa.com/?id=Micro%20EDM). For further reading also consult the *Wire EDM Handbook*.[6]

Microelectrical Discharge Machining

When referring to microelectrical discharge machining (µ-EDM) one refers either to working with a small EDM machine (see Figure 5.3 for a hand-held EDM at Panasonic) or to working with smaller than usual electrodes (in sinker EDM) or with thinner wires (in EDM-WC).

In die-sinking µ-EDM, an electrode with microfeatures is used to produce its mirror image in the work piece. The Matsushita Research Institute Tokyo, Inc. developed such a microsinker EDM machine.

FIGURE 5.3 Handheld µ-EDM at Panasonic.

This machine incorporates submicron movement resolution and uses a microtool (microelectrode).

The smallest microtools may be produced by the wire electrode discharge grinding method (see above) (http://www.panasonicfa.com/?id=Micro%20EDM). µ-EDM machining can produce a hole of 5–300 µm in diameter, with a depth three to five times the diameter, and at a precision of ± 0.5 µm on its circularity. Under the right conditions, the machine may result in a surface roughness of as little as 0.1 µm R_{max}. The latter is achieved by minimizing the electrical discharge energy to an extremely small amount. These microholes can be used as nozzles for ink-jet printers, flow control orifices, and pinholes for x-ray measurements. There are also more applications other than simply drilling microholes, such as microshafts, gears, springs, molds, dies, precision optics, and magnetic heads for digital video cameras.

In µ-EDM drilling, a microelectrode is used to "drill" microholes; a circular electrode is rotated at high speed (to help flushing of the liquid), and holes of 5 µm in diameter are reported. In µ-EDM milling, a simple-shaped electrode, rode or tube, is used to produce 3D cavities by adopting a contour motion.

The µ-EDM-WC technique can be used to make circular rods by rotating the work piece against the cutting wire as shown in Figure 5.4a, resulting in diameters as small as 5 µm and a surface roughness of 0.1 µm. The rods can have cross-sections such as triangular, rectangular, slit-type, and polygonal as well. The way to make a triangular shape is illustrated in Figure 5.4b.

Kuo et al.[7] used µ-EDM-WC to make micropipes with an inner diameter of 23 µm, an outer diameter of 186 µm, and a length of 3 mm. These micropipes came with a variety of micronozzle exit shapes and different sizes at the ends; slit openings of 100 × 10 µm, triangular openings with a base of 50 µm, and circular openings with a radius of 6 µm were demonstrated.[8] The manufacturing process of those micropipes is illustrated in Figure 5.5 (steps a–d). µ-EDM-WC is first used to machine a stainless steel rod (a); then electroforming is used to deposit a Ni coat (b); the Ni deposit is subsequently sized by µ-EDM-WC (c); and finally the stainless steel core is removed (d). To separate the core from the deposit

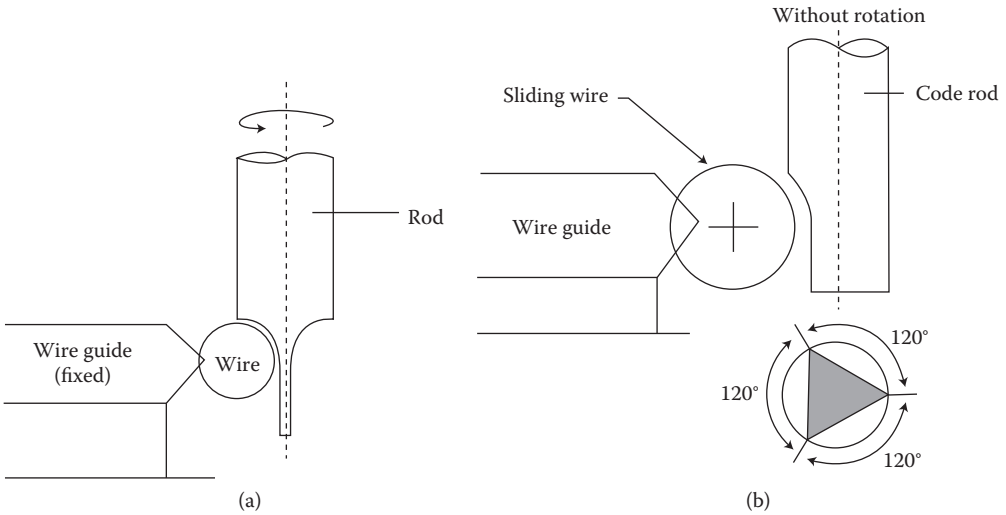


FIGURE 5.4 Microelectrical discharge machining, wire cutting (μ-EDM-WC) for a round rod profile (a) and for a triangular rod profile (b). (From Kuo, C.-L., T. Masuzawa, and M. Fujino. 1991. *Proceedings: IEEE micro electro mechanical systems (MEMS '91).* Nara, Japan; and Kuo, C.-L., T. Masuzawa, and M. Fujino. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92).* Travemunde, Germany. With permission.[7,8])

after plating, it is necessary to reduce the deposit tension stress as much as possible. A sulfamate nickel-plating solution with a plating temperature of 50°C and a pH of 4 was selected for this experiment because it can be operated with high current density and results in low residual stress. A schematic side view of a typical micropipe with a circular exit is shown in Figure 5.6. None of the lithography techniques discussed in the preceding chapters can produce this type of intricate pipe as easily as μ-EDM-WC. It is of great value for a micromachinist to recognize at a glance which shapes are better produced by a lithography-based method and which are better produced by a truly 3D lathe-based method. Microtools produced by EDM-WC may be used as sinker EDM electrodes to enable μ-EDM.

The use of microelectrode arrays enables one to use μ-EDM in batch mode as pioneered by Takahata.[9] Takahata used the LIGA process (Chapter 10) to make the microelectrode array shown in Figure 5.6. Using the setup shown here, 400 holes through 50-μm-thick stainless steel were made in

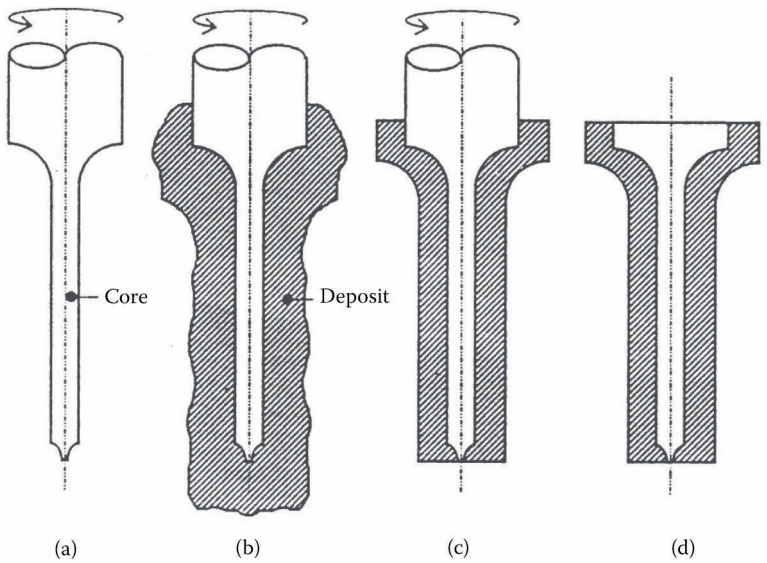


FIGURE 5.5 Micronozzle fabrication process with microelectrical discharge machining, wire cutting (μ-EDM-WC). For fabrication steps (a–d), see text.

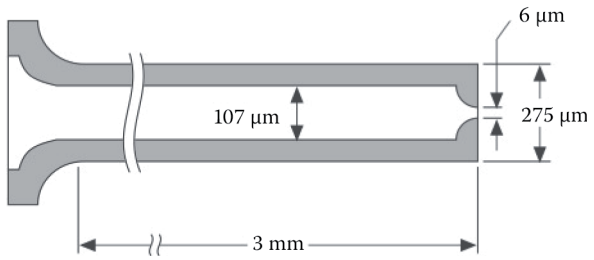


FIGURE 5.6 Schematic side view of micronozzle (see also Figure 5.5d). Schematic side view of a circular pipe with 6-mm exit hole. (From Kuo, C.-L., T. Masuzawa, and M. Fujino. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92).* Travemunde, Germany. With permission.[8])

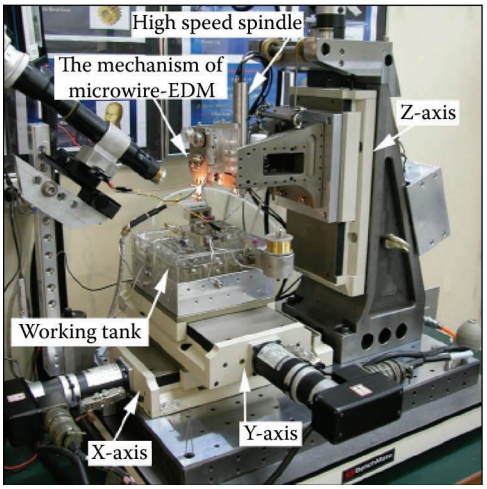


FIGURE 5.8 High precision four-axes multipurpose CNC machine for making microparts developed by Y. S. Liao et al. at the National Taiwan University. (Ehmann, K. F., D. Bourell, M. L. Culpepper, T. J. Hodgson, T. R. Kurfess, M. Madou, and K. Rajurkar. 2007. *Micromanufacturing: international research and development.* New York: Springer.[10])

5 minutes (20–30 times faster than serial machining with a single electrode). Other structures made with this hybrid LIGA-EDM method are also shown in Figure 5.7.

A desk-top high-precision four-axes computer numerical controlled (CNC) machine for making microparts was developed by Liao et al.[10] at the National Taiwan University. This multifunction high-precision machine (accuracy of 1 μm) does on-machine assembly and on-machine metrology (Figure 5.8). Parts in the size range of 0.1–2 mm can be produced, and all manufacturing processes are completed on the same machine without moving the work piece.

This hybrid desk-top machine is capable of performing the following four functions. 1) Milling with a low-speed (0–3,000 rpm), a middle-speed (1,000–10,000 rpm), and a high-speed (2,000–80,000 rpm) spindle. 2) μ-EDM with microelectrodes with a diameter as small as 8 μm and very high high-aspect ratios. These microelectrodes are made using the on-board μ-EDM-WC with a 0.02 mm ϕ brass wire (see next). 3) μ-EDM-WC that features microwire tension control and wire vibration suppression (Figure 5.9) and a novel mechanism that permits vertical, horizontal, and slanted wire cutting. 4) On-line measuring of the flatness, circularity, concentricity, dimensions, and geometric tolerances of the work piece.

The vertical, horizontal, and slanted cutting with the μ-EDM-WC tool has successfully fabricated complex features and parts. An example is the impressive Chinese pagoda (1.25 mm × 1.75 mm) shown in the

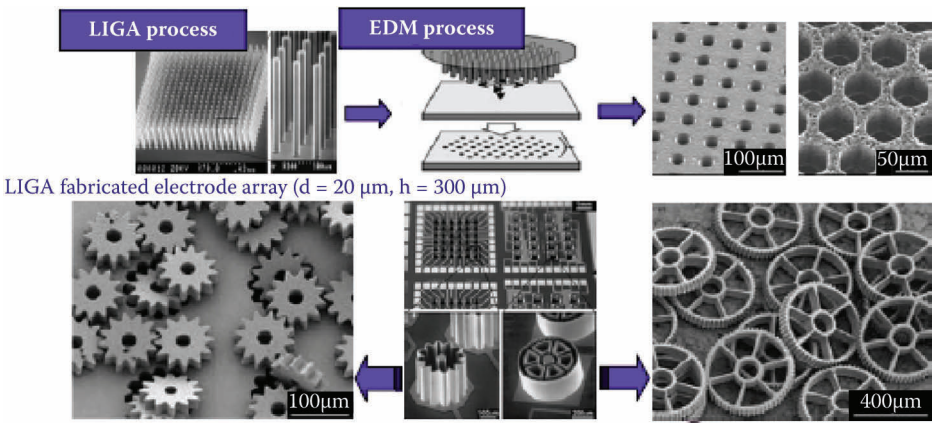


FIGURE 5.7 Microelectrode array fabricated using LIGA. This array toll enables the use μ-EDM in batch mode. (Takahata, K., and Y. B. Gianchandani. 2002. Batch mode micro-EDM for high-density and high-throughput micromachining. *IEEE/ASME J Microelectromechanical Systems* 11:102–110.[9]) Example products include copper electrode arrays and WC-Cu super hard gears.

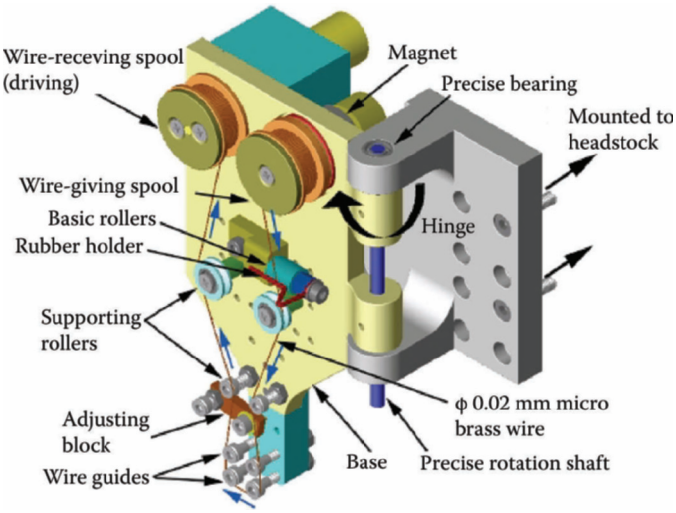


FIGURE 5.9 The μ-EDM-WC tool features microwire tension control and wire vibration suppression. The tool allows vertical, horizontal, and slanted wire cutting. μ-EDM-WC at the National Taiwan University (Prof. Y. S. Liao).

heading of this chapter. The vertical and horizontal μ-EDM-WC cuts to make this pagoda are shown in Figure 5.10.

High-speed micromilling with a single crystal diamond milling cutter is demonstrated in Figure 5.11. The depth of the cut is 0.03 mm; the rotational speed is 80,000 rpm; and the feed rate is 20 mm/min.

We conclude this section with a table comparing EDM with ECM, a chemical etching technique covered in Chapter 4, using conventional mechanical machining (CNC machining; see Chapter 6) as a reference point. In Table 5.2 we list metal removal rates (MRRs), tolerance, surface finish and damage depth, and required power.

The metal removal rate by ECM is much higher than that of the EDM, with a metal removal rate 0.3 that of CNC, whereas EDM is only a small fraction of the CNC material removal rate. Power requirements for ECM are comparatively high. The tolerance obtained by EDM and ECM is within the range of CNC machining, which means satisfactory dimensional accuracy can be maintained. All the processes obtain satisfactory surface finishes. Depth of surface damage is very small for ECM, whereas it is very high in the case of EDM. For this reason, ECM can be used for making dies and punches. Capital cost for ECM is very high when compared with conventional CNC machining, and

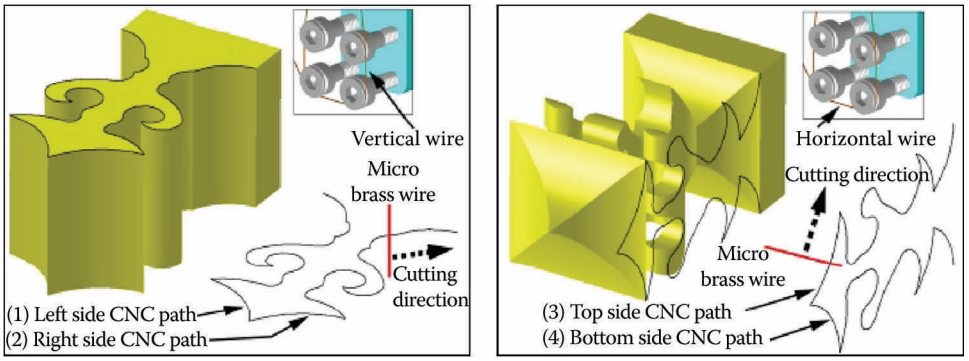


FIGURE 5.10 Vertical and horizontal μ-EDM-WC to make the pagoda in Al-6061. μ-EDM-WC at the National Taiwan University (Prof. Y. S. Liao).

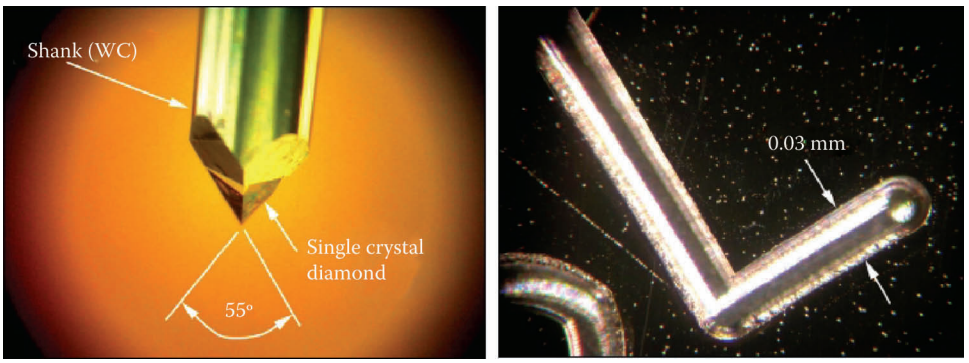


FIGURE 5.11 High-speed micromilling with a single diamond crystal on a WC shank.

Table 5.2 Machining Characteristics of EDM and ECM

Process	MRR (mm^3/min)	Tolerance (μm)	Surface Finish (μm)	Damage Depth (μm)	Power (W)
ECM	15,000	50	0.1–2.5	5	100,000
EDM	800	15	0.2–1.2	125	2,700
CNC	50,000	50	0.5–5	25	3,000

MRR, metal removal rate; tolerance, tolerance maintained; surface finish, surface finish required; damage depth, depth of surface damage; ECM, electrochemical machining; EDM, electrical discharge machining; CNC, computer numerical control machining.

EDM has also a higher tooling cost than the other machining processes.

Laser Beam Machining

Introduction

The basic operating principle of the laser was put forth by Charles Townes and Arthur Schalow from Bell Telephone Laboratories in 1958, and the first laser, based on a pink ruby crystal, was demonstrated in 1960 by Theodore Maiman at Hughes Research Laboratories. Lasers are defined as narrow beams (a divergence angle might be about 1 mrad, i.e., 1 m spread over 1 km of travel length) of monochromatic and coherent light in the visible, infrared (IR), or ultraviolet (UV) ranges of the spectrum. The key components of a laser unit are the active medium (a gas, liquid, or solid in which the beam is generated), the power supply (which delivers energy to the laser medium in the form needed to stimulate its emission), and the optical cavity (where the beam characteristics are defined). Lasers can be run in continuous mode or in pulsed mode to enhance the irradiance. The word *laser*, we saw in Volume I, Chapter 5, stands for Light Amplification by Stimulated Emission of Radiation. In Volume I, Chapter 5 we explained the lasing action; here we detail laser use in subtractive manufacturing. In 1960, the laser was said to be a "solution waiting for a problem," but today lasers are used in a wide range of applications, and some laser types have even become commodity products (e.g., laser pointers). In Table 5.3 we show the operating mode, wavelength, power, and example applications of some popular lasers (the list of potential applications is far from complete).

The CO_2 laser is an example of a gas laser based on a mixture of helium (83%) and nitrogen (16%) with carbon dioxide (6%) as the lasing material. The CO_2 laser produces a collimated coherent beam in the IR region (10.6 μm) characteristic of the active material (CO_2). The wavelength is strongly absorbed by glasses or polymers; thus, either mirrors or ZnSe lenses with excellent IR transparency are used to handle the beam.

The neodymium-doped yttrium aluminum garnet (Nd:YAG) laser is a solid-state laser producing a collimated coherent beam in the near-IR region of wavelength 1.06 μm, and can be pulsed or continuous. In pulsed mode, peak power pulses of up to 20 kW are now available. The shorter wavelength allows the use of optical glasses to control the beam path.

Excimer (EXCIted diMER) lasers, invented in 1975, are a family of pulsed lasers with a fast electrical discharge in a high-pressure mixture of a rare gas (krypton, argon, and xenon) and a halogen gas (fluorine and hydrogen chloride) as the source of emission. The particular combination of rare gas and halogen determines the output wavelength. Excimer lasers cover a range of wavelengths in the UV range from 157 nm (F_2) to 353 nm (XeF) with 193 nm (ArF), 248 nm (KrF), and 308 nm (XeCl) being particularly useful intermediate wavelengths.* Peak powers typically range from 3 MW (F_2) to 50 MW (KrF or XeCl) in pulses lasting a few tens of nanoseconds.

Machining with laser beams, first introduced in the early 1970s, is now used routinely in many industries. Laser machining, with long or continuous wave (CW), short, and ultrashort pulses, includes the following applications:

- Heat treatment
- Welding (see Volume III, Chapter 4 on packaging)
- Ablation or cutting of plastics, glasses, ceramics, semiconductors, and metals (this chapter)

* Recently, the 308-nm xenon chloride laser has gained much attention as a very effective treatment modality in dermatological disorders.

Table 5.3 Laser Types with Characteristics and Applications

	Wavelength	Maximumm Power Output (W)	Operation Mode	Applications
Semiconductor lasers, also injection or junction laser				
Single diodes	Infrared to visible (0.33–40 μm)	0.6 W (CW) 100 W (P)	CW and pulsed	Compact discs, laser printers, bar code scanners, optical communications
Diode laser bars	Infrared to visible	Up to 100 W	CW and pulsed	Pumping light source for solid state lasers
Solid-state lasers				
Ti-sapphire	690–960		Tunable	
Nd:YAG	1.064 μm	1000 W (CW) 2 10^8 W (P)	CW and pulsed	Welding, hole piercing, cutting
Ruby	694.3 nm	Several MW	Pulsed	Measurement instrumentation, pulsed holography
Gas/liquid lasers				
CO_2 (gas)	9.6 μm, 10.6 μm	1 W–40 kW (CW) 100 MW(P)	CW and pulsed	Heat treatment, scribing, welding, cutting
Excimer laser (gas)	193 nm, 248 nm, 308 nm (and others)	1 kW–100 MW	Pulse mode, pulse length 10–100 ns	Micromachining such as flex circuit via drilling and wire stripping, laser chemistry, medicine
HeNe (gas)	632.8 nm (most prominent), 1.15 μm, 3.39 μm	0.0005–0.05 (CW)	CW	Measurement instrumentation, holography, line-of-sight communications
Argon ion laser (gas)	515–488 nm (several)	0.005–20 (CW)	CW	Printing technology, pumping laser for dye laser stimulation, surgery
Dye laser (liquid)	Continuous between infrared and UV (different dyes): 0.38–1.0 μm	0.01 (CW) and 1 10^6 (P)	CW and pulsed	Measurement instrumentation, spectroscopy, medicine, pollution detection

Source: Based on http://www.ilt.fraunhofer.de/eng/100048.html.

- Material deposition (see Chapter 9)
- Etching with chemical assist, i.e., laser-assisted chemical etching (LACE; this chapter)
- Laser-enhanced jet plating and etching (see Chapter 9 and this chapter, respectively)
- Lithography (see Chapters 1 and 2)
- Surgery (see Volume III, Chapter 10)
- Photopolymerization (e.g., μ-stereolithography described in Chapter 9)

From this list, it is clear that lasers can be used for removal and for additive processes. We only consider ablative or subtractive laser processes here, as illustrated in Figure 5.12; additive laser processes are reviewed in Chapter 9. Interestingly, the laser may also serve as its own process monitor. This multiuse, multirole, and in situ capability is not offered with any of the other advanced material processing techniques such as molecular beam epitaxy, chemical vapor deposition, deep reactive ion etching, and so on.[11]

Earlier uses of laser beam machining (LBM) were mostly scientific in nature, and industrial applications involved macroprocesses, like sheet metal cutting and welding for the automobile industry (Figure 5.13). Today, industrial lasers are also used in a wide variety of microfabrication tasks. R. Srinivasan and V. Mayne-Banton of IBM Research first reported micromachining with laser ablation in 1982.[12] Initial studies by Srinivasan and coworkers showed that pulsed UV laser radiation at 193 nm (an ArF excimer laser) can be used to etch organic polymers several micrometers in depth. To illustrate how fast laser machining has improved and how little yet is known of the fundamental underlying processes in this field, consider that it was only recently recognized that femtosecond laser pulses interact with matter in a manner that is totally different from traditional lasers with longer pulses. During machining with femtosecond lasers, much less heat dissipates in the substrate, and, as

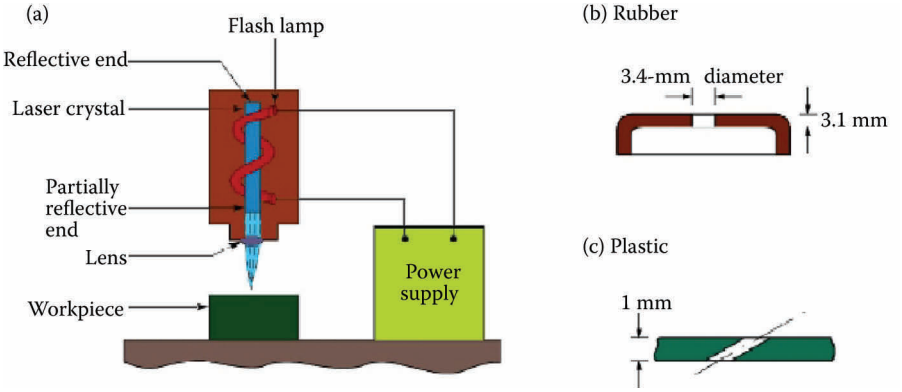


FIGURE 5.12 (a) Schematic illustration of the laser-beam machining process. (b and c) Examples of holes produced in nonmetallic parts by LBM.

a consequence, ultrafast lasers present the potential for micromachined products of superior quality (http://www.clark-mxr.com/industrial/handbook/index.htm).

Although the number of lasers delivered for strict micromachining applications has been low compared with others so far, it is predicted that the next 10 years will see a phenomenal market growth for this application.

Laser Parameters for Laser Material Removing—Laser Ablation

Introduction

Laser ablation is the process of removal of matter from a solid by means of an energy-induced transient disequilibrium in the lattice. The characteristics of the released atoms, molecules, clusters, and fragments (the dry aerosol) depend on the efficiency of the energy coupling to the sample structure, i.e., the material-specific absorbance of a certain wavelength, the velocity of energy delivery (laser pulse width), and the laser characteristics (beam energy profile, energy density or fluency, and the wavelength). Table 5.4 summarizes all the different laser parameters that can be controlled and their effect on materials processing.[13] More specifically for micromachining purposes, the wavelength, spot size [i.e., the minimum diameter of the focused laser beam, d_0 (Equation 5.9)], average laser beam intensity, depth of focus, laser pulse length, and shot-to-shot repeatability (stability and reliability in Table 5.4) are the six most important parameters to control. Additional parameters we will cover in the section below, not listed in Table 5.4, concern laser machining in a jet of water and laser-assisted chemical etching (LACE).

Like other thermal removing techniques, laser machining leaves a heat-affected zone (HAZ) on the work piece where molten material resolidified in situ or where material was sufficiently heated and cooled rapidly enough to result in embrittlement. This change in material properties can alter subsequent laser ablation and affect material performance.

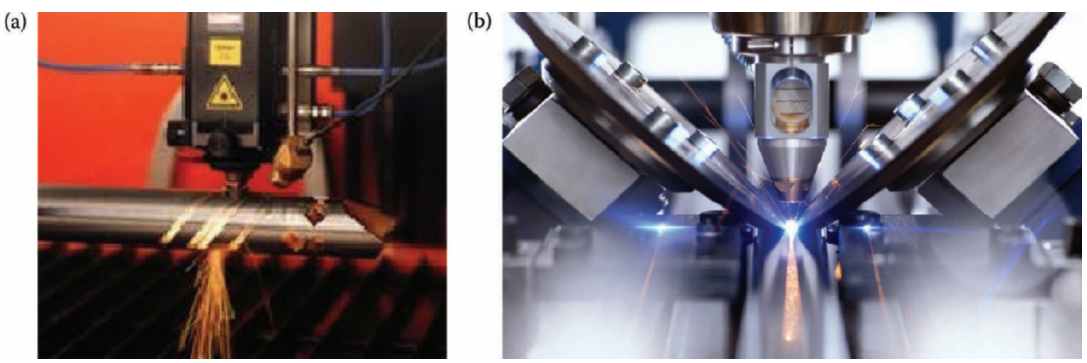


FIGURE 5.13 Laser manufacturing: (a) cutting; (b) welding.

Table 5.4 Laser Parameter and Related Processing Parameters

Laser Parameter	Influence on Material Processing
Power (average)	Temperature (steady state) Process throughput
Wavelength (μm)	Optical absorption, reflection, transmission, resolution, and photochemical effects
Spectral line width (nm)	Temporal coherence Chromatic aberration
Beam size (mm)	Focal spot size Depth of focus Intensity
Lasing modes	Intensity distribution Spatial uniformity Speckle Spatial coherence Modulation transfer function
Peak power (W)	Peak temperature Damage/induced stress Nonlinear effects
Pulse width (s)	Interaction time Transient processes
Stability (%)	Process latitude
Efficiency (%)	Cost
Reliability	Cost

Source: Liu, Y. S. 1989. *Laser microfabrication: thin film processes and lithography.* Eds. J. Y. Tsao, and D. J. Ehrlich. New York: Academic Press. With permission.[13]

The size of the HAZ is a function of the laser pulse duration and the thermal conductivity and specific heat of the work piece. The HAZ depends on the distance the heat is conducted within the material and varies with material and laser wavelength. The better the material conduction (thermal diffusivity), the greater the extent of the HAZ is. Associated with the HAZ are a recast layer and the formation of burrs. During the ablation process, the expulsion of material in the plasma jet creates a compressive force on the molten pool of material under the laser spot. This will cause a portion of the liquefied material to be forced out of the ablation zone, and it will deposit onto the surrounding region. Debris is also a related problem of the ablation process. Particulates redeposit onto the surface as solid fallout of higher boiling point second phases from the plume and as condensation of the primary material. Taken together, all the preceding effects reduce the quality of the ablated region.

Based on pulse length, there are three major laser machining operation modes: short, ultrashort, and long or CW. *Short* means that the pulse is longer than 10 ps. *Ultrafast* or *ultrashort* means that the laser pulse has a duration somewhat less than 10 ps, usually some fraction of a picosecond. *Long* or *continuous wave* lasers are lasers that are operated in a CW (pulse duration > 0.25 s) rather than a pulsed mode. Each operation mode has its own micromachining applications. Short pulses of coherent light replace electrons from the electron beam vacuum technology (EBM; see below) as the cutting tool by vaporizing the substrate (subtractive process). No vacuum is required, but the removal rate is much slower than in EBM. For "hard to drill" materials and when holes are smaller than 150 μm, mechanical drilling becomes less cost effective, and laser removing techniques become an attractive alternative. With ultrashort pulses, so much energy is deposited in the material at such a fast rate that the material is forced into a plasma state; the material goes from a solid to a gas phase without first going through a melt phase. This plasma then expands away from the material as a highly energetic gas, taking almost all the heat away with it. This newest operation mode avoids a lot of the problems associated with the HAZ and promises to open up a new universe of very accurate micromachined parts unavailable with long laser pulsing. Long-pulse or CW lasers can be used to melt the metal for welding (additive process).

Laser Wavelength

Laser wavelength has been identified as an important parameter influencing the efficiency of the energy transfer to the body of the irradiated sample (ablation characteristics). The selection of a wavelength with a minimum absorption depth helps ensure a high-energy deposition in a small volume for rapid and complete ablation. Several attempts have been made to describe the laser ablation phenomenon theoretically. It appears to be agreed now that, for very short wavelength radiation (~200 nm), direct photochemical bond breaking plays an important role in the ablation process. At that wavelength, photons have enough energy, and a one- or two-photon absorption by materials might cause direct bond cleavage in the solid, leading to rapid decomposition of the material into highly volatile molecular fragments. However, at longer wavelengths (>300 nm), the single-photon

energy is not sufficient for bond breaking; two-photon or multiphoton avalanche effect processes must be considered (see below). As mentioned above, a shorter absorption depth in the material to be machined means a higher laser energy density in metal.

The decay of the intensity of an electromagnetic wave with E along the *x*-axis and propagating in the *z*-direction in an absorbing material we derived in Volume I, Chapter 5 as:

$$I(z) = I_m e^{-\alpha z} \qquad \text{(Volume I, Equation 5.110)}$$

where $\alpha = \frac{4\pi n''}{\lambda}$ (with n'' the imaginary part of refractive index). Thus, the penetration depth d is given as [when d = z, then $I(z)/I_m = e$]:

$$d = 1/\alpha = \lambda / 4\pi n'' \qquad (5.1)$$

The surface condition also has some effect on the laser energy coupling. Rougher surface helps absorbing more, and special surface treatments can enhance the absorption.

Example 5.1

For copper the complex refractive index at 5893 Å is given as n = 0.62 + i2.57 and its reflectivity R = 0.73. For a λ = 100 Å one obtains a d = 6.2 Å and at a λ = 10μm, d = 62 Å.

Laser Spot Size

The laser beam spot size is another relevant factor influencing the efficiency and spatial resolution of laser ablation. Assuming cylindrical drilling, the ablated mass *M*, is given by:

$$M = \rho\pi L r^2 \qquad (5.2)$$

where ρ = the mass density
L = the ablation depth
r = the crater radius

This means that a variation of the spot size *r*, at constant ablation efficiency, leads to a quadratic variation of the signal intensity as a direct consequence of the ablated mass involved.

If the laser beam diameter is not of a controlled diameter, the ablation region may be larger than desired and result in excessive slope of the sidewalls.

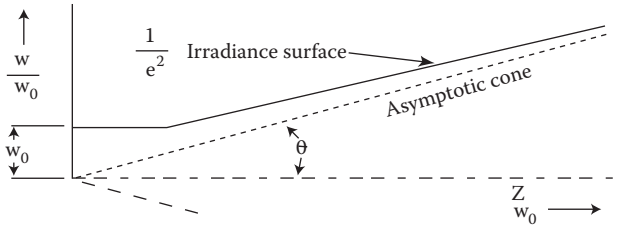


FIGURE 5.14 Growth in beam diameter as a function of distance from the beam waist for a Gaussian laser beam profile.

Diffraction causes light waves to spread transversely as they propagate; therefore, it is impossible to have a perfectly collimated beam. The spreading of a laser beam is in accord with the predictions of diffraction theory. If the objective is to create the smallest possible feature, it is important to be able to focus the laser beam to the smallest spot size. The latter depends on several factors, but crudely one can say that the smallest spot that one can create is about half the wavelength of the light used; if the wavelength of light is 1 μm, then the smallest spot one can create is about 0.5 μm (see Volume I, Equation 5.1 with NA = 1).

In a more rigorous description, we start from the irradiance distribution of a Gaussian TEM_{00}* laser beam (Figure 5.14):

$$I(r) = I_0 e^{-\frac{2r^2}{w(z)^2}} = \frac{2P}{\pi w(z)^2} e^{\frac{2r^2}{w(z)^2}} \qquad (5.3)$$

where r = the radial distance across the beam
P = total power in the beam in watts
w(z) = the radius or one-half the diameter of the $1/e^2$ irradiance contour after the wave has propagated a distance z

Gaussian beams propagate as Gaussians, and the expression for the spot size $w(z)$—the radius of the $1/e^2$ contour after the wave has propagated a distance z—is derived as:

$$w(z) = w_0\sqrt{1+\left(\frac{\lambda z}{\pi w_0^2}\right)^2} = w_0\sqrt{1+\left(\frac{z}{z_R}\right)^2} \qquad (5.4)$$

where λ = wavelength of light
w_0 = the waist radius of the $1/e^2$ irradiance contour at the plane where the wavefront is

* For the significance of the TEM_{00} laser beam designation, see Figure 5.19.

flat with all rays there moving in precisely parallel directions

z_R = *Rayleigh range* or *confocal parameter*, i.e., the distance over which the beam remains about the same size (or "collimated")

In Figure 5.14 we show this growth in beam diameter as a function of distance from the beam waist for a Gaussian laser beam profile. The point at z = 0 marks the location of a beam waist, or a place where the wavefront is flat.

The Raleigh range (z_R), also called the confocal parameter, defined as the distance over which the beam radius spreads by a factor of √2, is given by (see Figure 5.15):

$$z_R = \frac{\pi w_0^2}{\lambda} \tag{5.5}$$

Far from the waist ($z \gg z_R$), the laser beam asymptotically approaches the value:

$$w(z) = \frac{\lambda z}{\pi w_0} \tag{5.6}$$

From Figure 5.14 we see that a Gaussian beam does not diverge linearly; near the laser, the divergence θ angle is extremely small; far from the laser, half the divergence angle is given as θ ≈ tan (θ); thus, the $1/e^2$ irradiance contour asymptotically approaches a cone of angular radius:

$$\theta \approx \tan(\theta) = \frac{w(z)}{z} = \frac{\lambda}{\pi w_0} = 0.637\frac{\lambda}{w_0} = 1.27\frac{\lambda}{d_0} \tag{5.7}$$

This laser beam divergence is illustrated in Table 5.5, where we show the waist radius size w_0 of the beam and the collimation distances for two lasers, one at λ = 10.6 μm and one at λ = 0.633 μm.

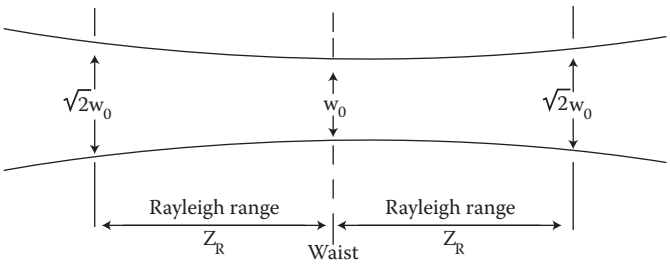


FIGURE 5.15 The Raleigh range (z_R), also called the confocal parameter.

TABLE 5.5 Waist Spot Size w_0 and Collimation Distances for Two Lasers at λ = 10.6 μm and λ = 0.633 μm

Waist Spot Size w_0	Collimation Distance λ = 10.6 μm	Collimation Distance λ = 0.633 μm
0.225 cm	0.003 km	0.045 km
2.25 cm	0.3 km	5 km
22.5 cm	30 km	500 km

We learn that tightly focused laser beams expand quickly and weakly focused beams expand less quickly, but still expand. As a result, it is very difficult to shoot down a missile with a laser. We also recognize that longer wavelengths expand faster than shorter ones.

Now let us focus this Gaussian laser beam using a lens (Figure 5.16) and calculate what its electric field will be one focal length *f* behind the lens. Gaussian beams focus as Gaussians.

If a "perfect" lens (no spherical aberration) is used to focus a collimated laser beam, the minimum spot size radius or the focused waist (w_0) is limited by diffraction only and is given by:

$$w_0 = \frac{\lambda f}{\pi w_{lens}} \tag{5.8}$$

where f = the focal length of the lens.

This is the same result as for a diffracting beam with z = f (Equation 5.6). Light being focused behaves just like diverging light, only reversed, and the divergence angle θ, given by Equation 5.7, is equally valid. With $d_0 = 1/e^2$ the diameter of the focus (= $2w_0$) and with the diameter of the lens $D_{lens} = 2w_{lens}$ (or the diameter of the laser beam at the lens—whatever is the smallest), we obtain:

$$d_0 = \frac{4\lambda f}{\pi D_{lens}} = \frac{1.27\lambda f}{D_{lens}} \tag{5.9}$$

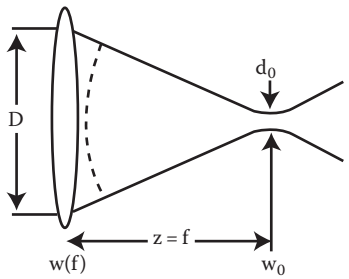


FIGURE 5.16 A Gaussian laser beam is focused using a lens.

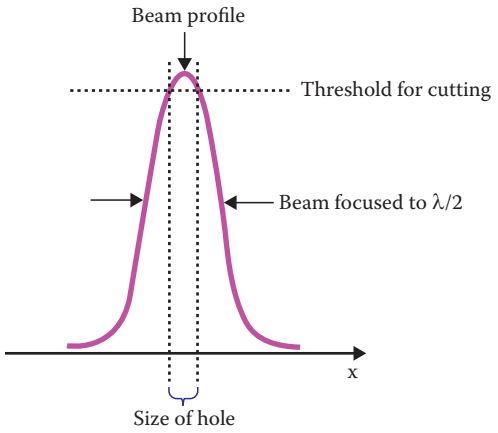


FIGURE 5.17 Size of hole cut with a laser can be substantially smaller than the beam size.

Thus, the principal way of increasing the resolution in laser machining, as in photolithography, is by reducing the wavelength, and the smallest focal spot will be achieved with a large-diameter beam entering a lens with a short focal length.

It is worth noting that ultrafast laser pulsing can create substantially smaller features than that of the central wavelength of the laser pulse itself. Because high-order nonlinear-optical processes are responsible for cutting, there is effectively a threshold for cutting (see Figure 5.17). This means that careful control of the intensity can yield a very small hole. By focusing the ultrafast laser pulse into a small spot with a Gaussian intensity profile, one can adjust the intensity of the laser spot on the surface of the material so that just the peak of the beam is above the ablation threshold; as a consequence, only material in that very limited area of the laser spot will be removed. That limited area can be as small as one-tenth the size of the spot itself; therefore, the unique combination of multiphoton absorption and saturated avalanche ionization provided by ultrafast laser pulses makes it possible to machine materials on dimensions much smaller than 1 μm: holes as small as 100 nm have been made.

Depth of Focus or Depth of Field

Twice the Raleigh range or 2 z_R is called the "depth of focus" because this is the total distance over which the beam remains relatively parallel, or "in focus" (see Figure 5.18). Or also, the depth of focus or depth of field (DOF) is the distance between the

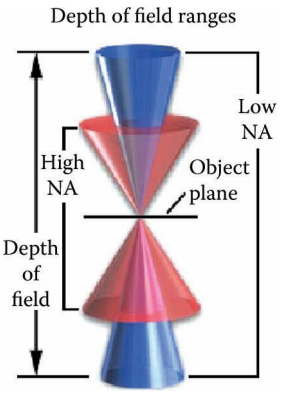


FIGURE 5.18 Depth of field (DOF) ranges.

values where the beam is √2 times larger than it is at the beam waist (see again Figure 5.18). Based on Equations 5.5, 5.8, and 5.9, we then obtain:

$$\mathrm{DOF} = 2z_R = \frac{2\pi w_0^2}{\lambda} = \left(\frac{8\lambda}{\pi}\right)\frac{f^2}{D_{lens}^2} \qquad (5.10)$$

The numerical aperture (NA) of a lens is given as $n\sin\theta$, and for small angles, one can derive that NA = D/2f with D the diameter of the lens and n the refractive index which in air = 1 (see Volume I, Chapter 5). Thus, we can recalculate the DOF also as:

$$\mathrm{DOF} = 1.27\lambda/\mathrm{NA}^2 \qquad (5.11)$$

Material processing with a very short depth of focus requires a very flat surface. If the surface has a corrugated topology, a servo loop connected with an interferometric auto-ranging device must be used.[11]

We want to contrast the results for resolution and depth of focus of laser beams obtained here with the resolution and depth of focus we calculated for other light sources in Volume I, Chapter 5 and in this volume, Chapter 1. In Volume I, Chapter 5 we saw that the Rayleigh criterion defines two incoherent light sources, separated by a distance d, as resolved when:

$$d = 1.22\frac{f\lambda}{D} \qquad \text{(Volume I, Equation 5.12)}$$

With an ideal lens system of focal length f, the Rayleigh criterion further yields a minimum spatial resolution as:

$$d = \frac{1.22\lambda}{2\mathrm{NA}} \qquad \text{(Volume I, Equation 5.13)}$$

TABLE 5.6 Comparing Standard Lens Parameters for Coherent and Incoherent Light

	Incoherent	Coherent (Gaussian Laser Beam)
Lens diameter	D	$D_{lens} = 2w_{lens}$
Focal spot radius size	$d = 1.22\frac{f\lambda}{D}$	$d_0 = \frac{4\lambda f}{\pi D_{lens}} = \frac{1.27\lambda f}{D_{lens}}$
Depth of focus	$DOF = 1.22\lambda\left(\frac{2f}{D}\right)^2 = 1.22\lambda/NA^2$	$DOF = 1.27\lambda\left(\frac{2f}{D_{lens}}\right)^2 = 1.27\lambda/NA^2$

In this expression d is the smallest object that the lens can resolve and also the radius of the smallest spot to which a collimated beam of light can be focused. From Equation 1.17 in Chapter 1 the DOF for an incoherent light source can be derived as:

$$DOF = 1.22\lambda\left(\frac{2f}{D}\right)^2 = 1.22\lambda/NA^2 \qquad (1.17)$$

We contrast these earlier results with those obtained for laser beams in Table 5.6. Whereas resolution is slightly worse for coherent light, its DOF is slightly better.

Laser beams can have any pattern, not just a Gaussian. The beam shape can even change with distance. Some beam shapes do not change with distance. These laser beam shapes are referred to as transverse electromagnetic (TEM) modes. Lasers can operate in various transverse modes as shown in Figure 5.19. Laser specifications will usually refer to the TEM_{00} mode. The 00 mode is the Gaussian beam. Higher-order modes involve multiplication of a Gaussian by a "Hermite polynomial." The TEM_{00} results in a single beam. The long narrow bore of a typical laser forces this mode of oscillation, and Equation 5.9 is only valid for this mode.

Laser Power

The laser power is also a critical parameter because it determines the damage impact on the irradiated solid. However, it must be considered in combination with the wavelength (see above). To clarify this, a distinction must be made between the optical penetration depth (Equation 5.1, wavelength-dependent) and the damage penetration depth (influenced by the power density as well). As we saw in Equation 5.1, the optical penetration depth is defined as the distance below the irradiated surface over which the laser energy is deposited. The fraction of this distance over which the irradiation exceeds the ablation threshold is defined as the damage penetration depth.

The laser intensity (W/cm^2) is given by the peak power (W) divided by the focal spot area (cm^2), and the peak power (W) is given as the pulse energy (J) divided by pulse duration (s). The intensity levels of a highly collimated, monochromatic, and coherent light beam at the work piece, which can be held in open air, may reach 10^6 W/mm^2, enough to melt any metal. The laser fluency or energy density (J/cm^2) equals the laser pulse energy (J) divided by focal spot area (cm^2). A typical 50-μJ output of a KrF laser

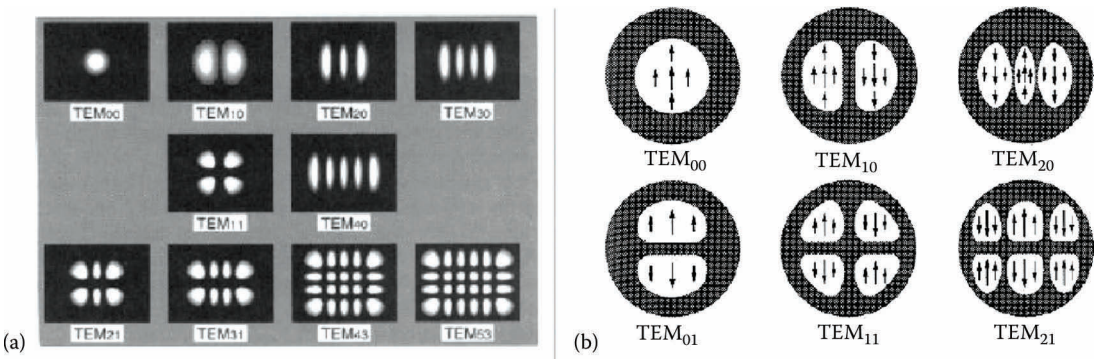


FIGURE 5.19 Some transverse electromagnetic (TEM) modes: (a) irradiance; (b) electric field.

focused by an optics system with NA of 0.5 will produce a focal spot of 15–20-µm diameter and a corresponding fluence of 10–20 J/cm^2.

The average on-axis laser beam intensity is related to the total laser power, *P* in watts, by:

$$I(r) = \frac{2P}{\pi w_0^2} \tag{5.12}$$

where P = the total power in watts

w_0 = the Gaussian beam radius—the radius at which the intensity has decreased to $1/e^2$ or 0.135 of its value on the axis

The average beam energy is of no use if it cannot be properly and efficiently delivered to the ablation region; thus, beam quality is not only measured by the brightness (energy) but also by its focusability and homogeneity.

Now we will have a look at a fifth important laser parameter influencing micromachining quality, that is, laser pulse length.

Laser Pulse Length

Long Pulses A further parameter with pulsed laser systems is the laser repetition rate, which expresses the number of pulses that are triggered per unit time. The laser pulse width determines the kinetics of the photon-induced disequilibrium on irradiated surfaces (for pulses longer than 100 ns the energy can be delocalized in thermal modes). It influences the role of the laser microplasma, too. The most fundamental feature of laser/material interaction in the long pulse regime (e.g., pulse duration 8 ns, energy 0.5 mJ) is that the heat deposited by the laser in the material diffuses away during the pulse duration; that is, the laser pulse duration is longer than the heat diffusion time. This may be desirable for laser welding, but for most micromachining jobs, heat diffusion into the surrounding material is undesirable and detrimental to the quality of the machining (http://www.clark-mxr.com). Here are a couple of reasons why one should avoid heat diffusion for precise micromachining:

- Heat diffusion reduces the efficiency of the micromachining process because it takes energy away from the work spot—energy that would otherwise go into removing work piece material. The higher the heat conductivity of the material, the more the machining efficiency is reduced.
- Heat diffusion reduces the temperature at the machining spot, pinning the working temperature to not much above the melting point of the material, which is removed by melting. Boiling of the substrate ejects globs of molten material away from the work zone. Those globs fall back onto the surface, contaminating the sample. The droplets can be rather large and retain a fair amount of residual heat so that they bind strongly to the sample surface. Removal of these contaminants is often difficult or impossible without damaging the work piece. An air jet is sometimes used to physically project the melt phase away from the work zone. This results in a cleaner-looking cut but contaminates the sample downstream.
- Heat diffusion reduces the accuracy of the micromachining operation. Heat diffuses away from the focal spot and melts an area that is much larger than the laser spot size. Therefore, it is difficult to do very fine machining; whereas the minimum laser spot size might be in the range of 1 µm or less, in many materials it is not possible to create features with dimensions much smaller than 10-µm diameter.
- Heat diffusion affects a large zone around the machining spot, a zone referred to as the *heat-affected zone* or HAZ. The heating (and subsequent cooling) waves propagating through the HAZ cause mechanical stress and may create microcracks (or in some cases, macrocracks) in the surrounding material. These defects are "frozen" in the structure when the material cools, and in subsequent routine use these cracks may propagate deep into the bulk of the material and cause premature device failure. A closely associated phenomenon is the formation of a recast layer of material around the machined feature. This resolidified material often has a physical and/or chemical structure that is very different from the unmelted material. This recast layer may be mechanically weaker and must often be removed.

- Heat diffusion is sometimes associated with the formation of surface shock waves. These shock waves can damage nearby device structures or delaminate multilayer materials. Whereas the amplitude of the shock waves varies with the material being processed, it is generally true that the more energy deposited in the micromachining process, the stronger the associated shock waves.

The various undesirable effects associated with long laser pulse etching are illustrated in Figure 5.20a and b.

Short Pulses We are particularly interested in femtosecond lasers with pulses more than 1 million times shorter than the several nanosecond-duration pulses used in traditional industrial laser micromachining systems (see above). Femtosecond lasers deliver an incredible amount of peak power, routinely up to 5–10 GW (this is more than the average power delivered by a large nuclear plant) (http://www.clark-mxr.com/industrial/handbook/introduction.htm). The laser intensity easily reaches the hundreds of terawatts per square centimeter at the work spot itself. No material can withstand the ablation forces at work at these power densities. This means that, with ultrafast laser pulses, very hard materials, such as diamond, as well as materials with extremely high melting points, such as molybdenum and rhenium, can be machined. The most fundamental feature of laser-matter interaction in the very fast pulse regime is that the heat deposited by the laser into the material does not have time to move away from the work spot during the time of the laser pulse. The duration of the laser pulse is shorter than the heat diffusion time. This regime has numerous advantages as listed below (http://www.clark-mxr.com/industrial/handbook/introduction.htm):

- Because the energy does not have the time to diffuse away, the efficiency of the machining process is high. Laser energy piles up at the level of the working spot, whose temperature increases instantly past the melting point of the material and keeps on climbing into what is called the *plasma* regime.
- After the ultrafast laser pulse creates the plasma at the surface of the work piece, the pressures created by the forces within it cause the material to expand outward from the surface in a highly energetic plume or gas. The internal forces that previously held the material together are vastly insufficient to contain this expansion of highly ionized atoms and electrons from the surface. Consequently, there are no droplets that condense onto the surrounding material. Additionally, because there is no melt phase, there is no splattering of material onto the surrounding surface.

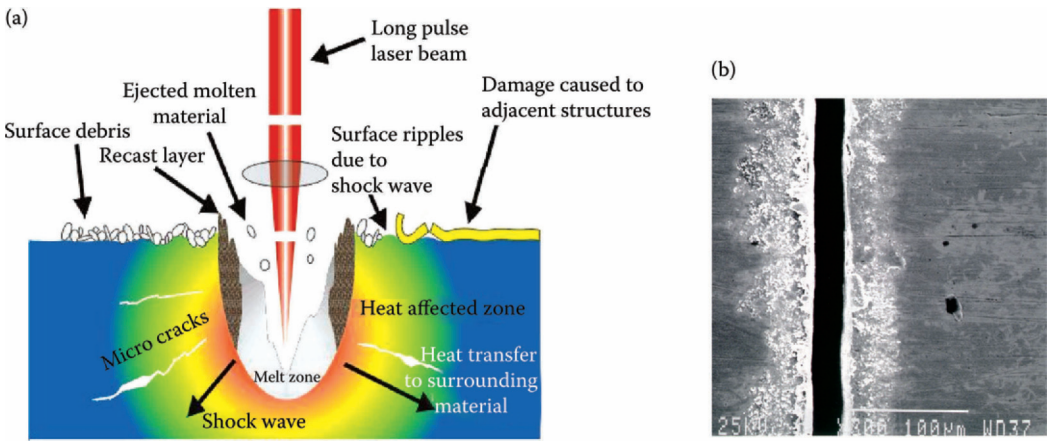


FIGURE 5.20 (a) Various undesirable effects associated with laser pulses that are too long, for example, pulse duration 8 ns, energy 0.5 mJ (http://www.clark-mxr.com). (b) Example of a 25-μm (1 mil) channel machined in 1-mm (40 mils) thick INVAR with a nanosecond laser. INVAR is often called for in the design of machinery that must be extremely stable. This sample was machined using a "long" pulse laser. The laser pulse parameters are pulse duration 8 ns, energy 0.5 mJ. A recast layer can be clearly seen near the edges of the channel. Large debris is also seen in the vicinity of the cut.

- Heating of the surrounding area is significantly reduced, and, consequently, all the negatives associated with a HAZ are no longer present—no melt zone, no microcracks, no shock wave that can delaminate multilayer materials, no stress that can damage adjacent structures, and no recast layer.

The many advantages associated with working in the femtosecond pulse range are illustrated in Figure 5.21a and b.

With ultrashort laser pulsing, it is possible to accurately control the depth of the cut by counting the number of pulses. Deep cuts (hundreds of microns) have been made this way, mostly by using an excimer laser, although it is now known that clean ablative etching, once the hallmark of excimer lasers, can also be achieved with pulsed laser sources at other wavelengths, including CO_2, copper vapor, and Nd:YAG and its harmonics. The latter lasers can be as effective for high-quality ablative micromachining provided laser photons are strongly absorbed in submicron depths at the surface of the material on time scales that are less than the time it takes for heat to diffuse away from the irradiated region. Lasers providing ultrafast pulse durations of less than 10 ps have, in recent years, demonstrated remarkably clean micromachining of many materials.

Ultrashort laser pulses have opened up many new possibilities in laser-matter interaction and materials processing. The extremely short pulse width makes it easy to achieve very high peak laser intensity with low pulse energies. The laser intensity can reach 10^{14}–10^{15} W/cm^2 with a pulse <1 mJ when a subpicosecond pulse is focused to a spot size of a few tens of micrometers. The extremely high intensity and short laser/matter interaction times lead to a highly nonequilibrium plasma state in the material under irradiation, where electrons are driven to much higher temperatures than ions. The short duration means that the hydrodynamic motion of the matter under laser irradiation can be ignored, and there is essentially no fluid dynamics to consider during the laser/matter interaction. The catch with ultrafast laser ablation is that throughput is low because the rate of removal of material is dependent on average power. The technology that makes these ultrafast laser pulses does not produce high average power. Additionally, the technology is very expensive (a system specifically tailored to an application can cost in the six- to seven-figure range), and the covered areas are very small.

In the past two decades lasers have become a precise, fast, wearless, flexible tool used for cutting, drilling, marking, and welding. Many of today's production processes are unimaginable without the laser.

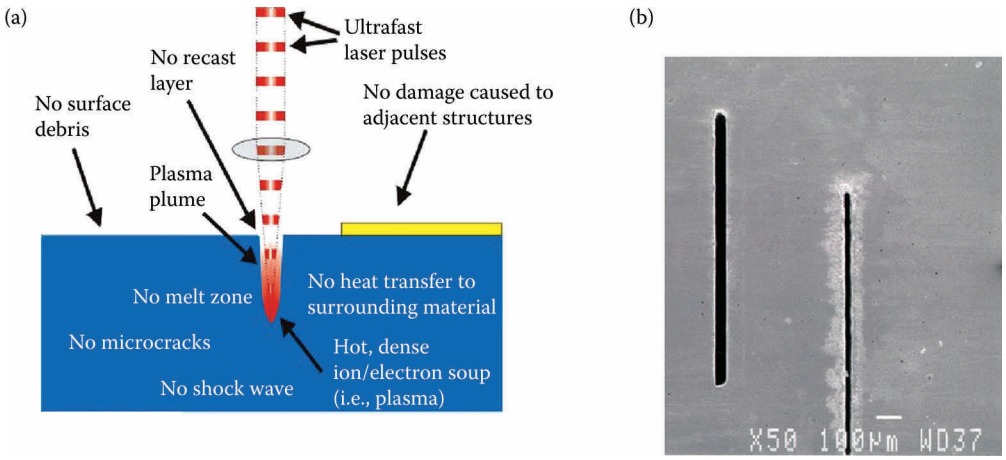


FIGURE 5.21 (a) Advantages associated with working with very short laser pulses (<10 ps). (b) A channel machined in 1-mm-thick INVAR, nickel/iron alloy under the same experimental conditions as the long pulse channel in Figure 5.20b, but with ultrafast pulses. The channel on the left was machined with 200-fs pulses, 0.5 mJ energy per pulse. It is obvious that the channel machined with femtosecond pulses is cleaner than the sample machined with nanosecond pulses (channel on the right). Note also the absence of a recast layer on the left channel. It is also clear that the machining process was more efficient—the channel is larger. The edges are straighter. Overall the quality of the micromachining is much higher.

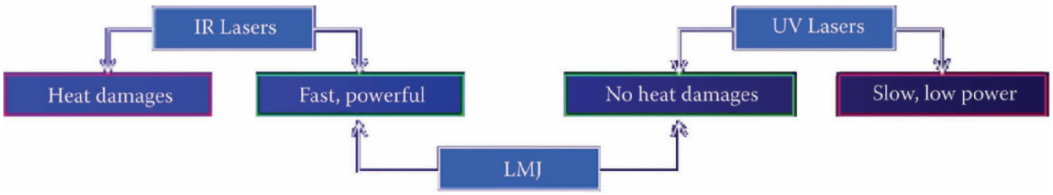


FIGURE 5.22 The laser dilemma. According to Synova, S.A., their water jet-guided laser, the Laser-Microjet (LMJ) (see also Figure 5.23), promises a solution.

Nevertheless, in some applications, as for example in the electronic or medical field, the laser is still too imprecise, too slow, or unacceptable because of material damages. High-power IR lasers, widely used in industrial material processing, are too "hot," creating heat damage and debris; meanwhile, shorter wavelengths found in UV lasers create less heat and debris but lack power for efficient material processing. This dilemma is summarized in Figure 5.22. According to some, water jet-guided laser machining, discussed below, might provide the solution for this dilemma.

Shot-to-Shot Repeatability

The process of laser ablation requires threshold fluency, and, as we have seen above, fluency is the laser pulse energy per unit area, and it can be changed either by changing the laser pulse energy or by changing the ablation area. At fluencies below the threshold, only miniscule etching (<0.05 μm/laser pulse) is observed, if any at all. However, when laser fluency is above the threshold, the etch depth per pulse increases rapidly with increasing laser fluency. The threshold fluency depends mainly on the target material structure and target absorption coefficients at the laser wavelengths. Absorption of laser radiation leads to a temperature increase in the target material. When laser pulse fluency is below the ablation threshold, the temperature increase is directly proportional to the laser pulse fluency, whereas when laser pulse fluency rises above the threshold, the temperature increases considerably more slowly. The absence of free electrons in a given substrate prevents the avalanche process from significantly starting and ultimately hinders the material ablation. In metals, there are always plenty of free electrons, and the avalanche process starts immediately, leading to reproducible machining (as we have seen, there may be other problems associated with heat diffusion). In semiconductors or insulators, there are naturally very few free electrons. The avalanche process may start right away or may not, depending on the presence or absence of a free electron in the beam path. This variability, which is inherent to the physical process, leads to unstable machining rates. The laser may be perfectly stable, and the beam spot size and amount of energy in the pulse may be precisely the same from shot to shot, yet the material ablation can vary significantly. Using lasers working in the UV range, or using ultrafast lasers, the electromagnetic field is so high that "bound" electrons are knocked free so that a large quantity of free electrons can start the avalanche ionization process immediately, reliably, and reproducibly. The latter ensures an excellent shot-to-shot reliability independent of the material.

Water Jet-Guided Laser Machining

In water jet-guided laser machining, a thin jet of high-pressure water (the diameter of the jet is between 40 and 100 μm, and the water pressure is between 20 and 500 bars) is forced through a nozzle (made of diamond or sapphire). The laser beam is focused through a water chamber (the water is deionized and filtered) into a nozzle as shown in Figure 5.23.

Leaving the nozzle, the laser beam is guided inside the water jet by total reflection at the water-air interface and is directed onto the work piece where it is absorbed by the material. The laser beam melts or vaporizes the material, and the water jet expels the molten debris. At the end of each laser pulse, the plasma disappears, and only the water jet remains in contact with the surface. Before the next pulse is emitted, the water jet removes any heat induced in the material during the laser ablation. Thus, the pulsed mode of the laser ensures continuous alternation of

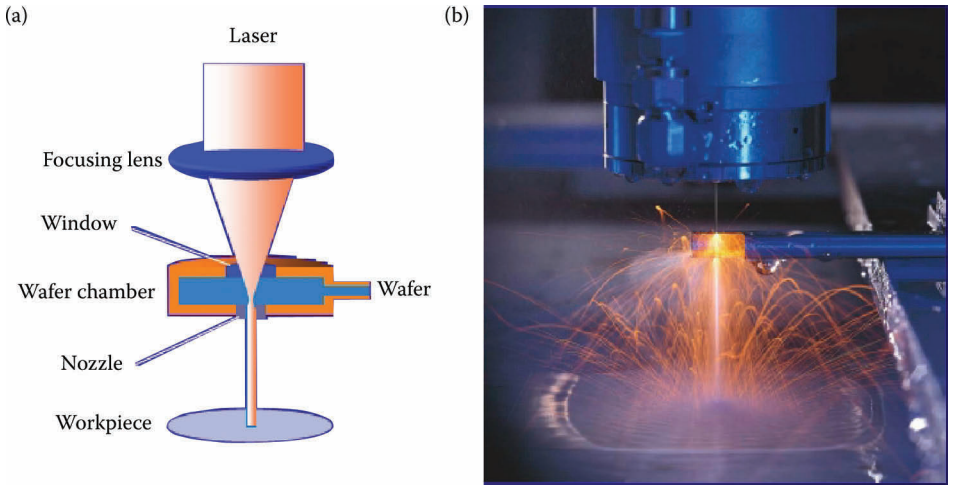


FIGURE 5.23 (a) Schematic water jet-guided laser machining setup. See text. (b) Water jet-guided setup from Synova, S.A., the Laser-Microjet®.

heating (during laser pulse) and cooling (between laser pulses). The water pressure value (20–500 bars) depends on the application, but the resulting force acting on the work piece is always very small, smaller than the force as a result of the cutting gas of conventional lasers. Thus, the water jet is obviously not abrasive in this case (as it is in mechanical water jet machining described in Chapter 6); the material is solely ablated by the laser (thermally). The high speed of the water, up to 300 m/s at 500 bars, ensures fast removal of debris. Water jet-guided laser etching prevents the problem of heat damage (microcracks, structural changes) inherent in any laser ablation process. The result is a heat damage-free material ablation.

The laser dilemma of high-power IR lasers that are too "hot," creating heat damage and debris, and the shorter wavelengths UV lasers that create less heat and debris, but lack power for efficient material processing, was summarized in Figure 5.22. Water jet-guided laser machining discussed here might provide a solution for this dilemma because it transforms a "hot" IR laser into a "cool" laser, avoiding heat damage. Water jet-guided machining combines high ablation efficiency, speed, and precision while avoiding material damage as demonstrated in Figure 5.24, where we compare laser ablation in Si using a conventional laser (Figure 5.24a) with jet-guided laser machining using the Laser-Microjet from Synova (Figure 5.24b). The Laser-Microjet was conceived in 1993 by scientists at the Institute of Applied Optics of the Swiss Federal Institute of Technology in Lausanne, Switzerland. Ten years later, this technology has proven its efficiency and reliability in many applications, especially of the semiconductor and microelectronics industries. The water jet-guided laser has indisputable advantages over the classical methods. The conventional laser kerv in Si exhibits typical heat damage with burrs and deposits, whereas the Laser-Microjet result exhibits neither burrs nor deposits.

Laser Etching with Chemical Assist

We just saw how water jet-guided laser machining helps keep a work piece clean; an air jet also may help to physically project the melt phase away from the work zone. This again results in a cleaner-looking cut but may contaminate the sample downstream. A third way to keep the work piece clean is the use of laser-assisted chemical etching (LACE).

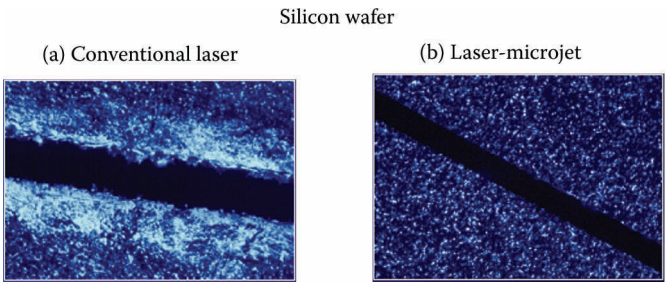


FIGURE 5.24 Silicon wafer laser machining. (a) With a conventional laser and (b) with a Laser-Microjet. The conventional laser kerv in silicon shows burrs and deposits, whereas the Laser-Microjet result has neither burrs nor deposits.

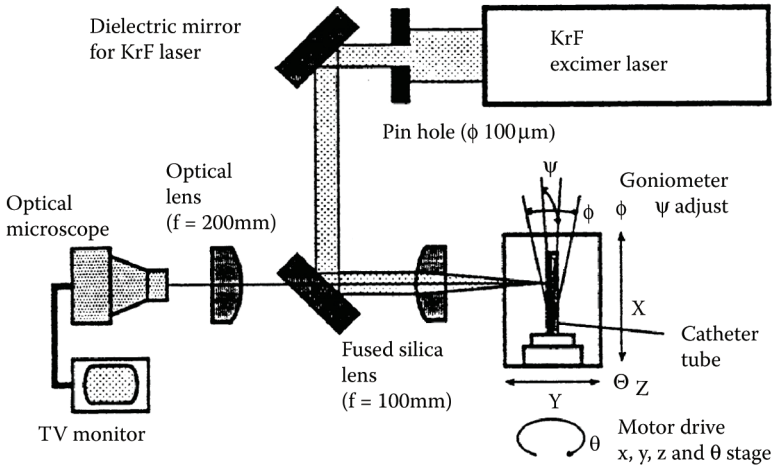


FIGURE 5.25 Laser microetching experimental setup: laser-assisted chemical etching (LACE). Etching/ablation of a catheter tube.

The laser cutting action can be assisted by dry or wet chemical etchants for building of higher-quality complex 3D micromachines. The process is also called *laser-induced chemical etching*, in which, for example, chlorine is used as the etchant and is photodissociated to react with the substrate.[14,15] If laser etching is performed in chlorine gas instead of open air, the silicon vapor and droplets leaving the work area combine with the chlorine to form $SiCl_4$ gas and are carried away from the work area. Silicon, when heated to its melting point, reacts nearly at the gas transport-limited rate with chlorine, enabling removal rates exceeding 10^5 μm³/s.[16] The resulting hole is cleaner and more sharply defined, and debris does not splatter on the surface of the substrate; thus, an excellent spatial resolution results (1 μm *x*,*y*-resolution at a removal rate of 2.10^4 μm³/s).[16,17] The latter removal rate is several orders of magnitude faster than with electrical discharge machining (EDM) methods.

Flexible laser machining setups as shown in Figure 5.25 can be used for the machining of 3D parts by using highly localized laser microchemical reactions and an accurate computer-controlled *x*, *y*, *z* and θ translation table. This same type of setup can accommodate additive processes (see Chapter 8), but it is more convenient to build separate, dedicated equipment, as illustrated in Figures 5.25.[18]

Remember LACE is not a parallel process and not fast enough for most manufacturing applications. The technique may have utility in specialty micromachining or for making molds.

At the Massachusetts Institute of Technology Lincoln Laboratories, scanning of the laser beam at a rate of 7500 μm/s is used to remove Si layers in a Cl_2 etchant. After a plane of 1 μm is removed, the focusing objective is lowered 1 μm, and a new pattern is etched. With this technique, it is very easy to produce multilevel-type structures. Some example devices made this way are shown in Figure 5.26. The Lincoln group also demonstrated that structures as shown can be replicated by using simple hot embossing molding techniques.[16]

In addition to Cl_2, reagents such as Br_2, HCl, XeF_2, KOH, and so on can also be used as the etchant in the laser photochemical etching of silicon. Table 5.7 summarizes several kinds of substrates etched by laser processes with different etchants and laser wavelengths.[11,20]

Laser Machining Applications

Overview

Advantages of laser cutting over mechanical cutting vary according to the situation, but important factors are lack of physical contact (because there is no cutting edge that can become contaminated by the material or contaminate the material) and, to some extent, precision (because there is no wear on the laser). There is also a reduced chance of warping the material that is being cut because laser systems have a small heat-affected zone (HAZ). Some materials are also very difficult or impossible to cut by more traditional means. The laser has other distinct

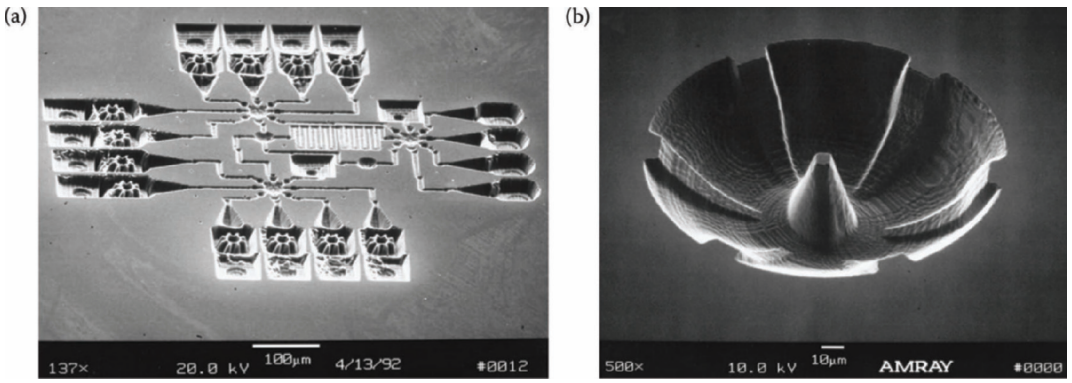


FIGURE 5.26 Some examples of laser-machined shapes. [Bloomstein, T. M., and D. J. Ehrlich. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91).* San Francisco.[19]] The devices are built using chlorine-assisted laser etching (100 Torr Cl_2). The beam spot used is 1 µm, and the objective is lowered in 1-µm increments after each plane is scanned at 7500 mm/s. Etch rate 100,000 mm^3/s. (a) SEM micrograph of a microfluidic device layer cut into silicon. The white bar is a 100-µm marker. (b) SEM micrograph of a cone inside a dish cut into silicon. The dish has an upper radius of 80 µm and a lower radius of 35 µm with a radius curvature of 82 µm and spacer layers extending 8 µm inward. The clipped cone has an upper radius of 4 µm and lower radius of 16 µm. The laser was swept in a circular scan in each plane. The white bar is a 10-µm marker. (Courtesy of Dr. D. Ehrlich, Revise, Inc.)

advantages: aspect ratios for hole diameter to depth on the order of 1:50 are possible, and holes can be drilled in hard-to-reach areas and at difficult angles (e.g., 10° to the surface). A laser-based scribe can make narrower trenches than a mechanical diamond stylus. Disadvantages of laser cutting may include the high energy required. Also, when hundreds of holes all at the same angle are to be drilled, the job is better handled by an electron beam machine (EBM). However, if holes have to be drilled at many

TABLE 5.7 Examples of Etchants and Substrates Etched by the Laser Photochemical Process

Substrate	Etchant(s)	Etch Rate	Laser/Intensity
Semiconductors			
Si	Cl_2,HCl	7 µm/s	Ar^+/>5 MW/cm^2
Si	KOH	15 µm/s	Ar^+/~10^7 W/cm^2
Si	NaOH	2 µm/min	CO_2/NA
Ge	Br_2	36 µm/s	Ar^+/0.1 kW/cm^2
GaAs	HNO_3	2 µm/s	Ar^+/60 MW/cm^2
GaAs	CH_3Br (750 Torr)	60 Å/s	Ar^+ (257 nm)/1 KW/cm^2)
GaP	KOH	600 Å/s	Ar^+ (351 nm)/3.5 KW/cm^2
InP	CH_3Br (750 Torr)	9.4 Å/s	Ar^+ (257 nm)/100 W/cm^2
Insulators			
SiO_2	SiH_4	40 Å/s	KrF/0.3 J/cm^2
SiO_2	NF_3:H_2	0.12 nm/s	ArF/7.7 mJ/cm^2
Polyimide	KOH	0.3 µm/s	Ar^+/0.03 MW/cm^2
Polyimide	Air	0.15 µm/pulse	KrF/0.3 J/cm^2
PMMA	Air	0.3 µm/pulse	ArF/0.25 J/cm^2
Diamond	Cl_2, O_2,NO_2	1,400 Å/pulse	ArF/30 J/cm^2
Metals			
Ti	NF_3	0.29 Å/pulse	0.115 J/cm^2
W	Cl_2 (0.1 Torr)	240 Å/min	0.12 J/cm^2
Al	Cl_2 (0.1 Torr)	1.4 µm/s	1.0 J/cm^2
Ag	Cl_2 (0.1 Torr)	500 Å/min	N_2 (337 nm)/0.12 J/cm^2

Source: Helvajian, H. 1995. *Microengineering technology for space systems.* Ed. H. Helvajian. El Segundo, CA: Aerospace Corporation; and Eden, J. G. 1986. Photochemical processing of semiconductors: new applications for visible and ultraviolet lasers. *IEEE Circuits Devices Mag* 2:18–24.[11,20]

angles, the laser may be the most cost-effective method. Lasers are most cost-effective for drilling holes between 0.01 and 1.5 mm. For larger holes, the cost of the large laser oscillator needed to produce the energy for vaporization becomes prohibitive, and because of the large amounts of material being vaporized, the hole produced suffers in terms of quality and controllable tolerance. With the very small holes, the high-power focusing lenses required to withstand the energies necessary for drilling holes smaller than 0.01 mm are limiting factors. The requirement for even smaller features necessitates the use of EBM (a thermal technique, see below), focused ion beam (FIB, a mechanical technique; see Chapter 6), or ultrafast laser pulsing (see above).

Both solid-state and gas lasers are commercially available and are touted for traditional machining operations in electronics such as postassembly processing, where site-specific action is necessary, for example, for drilling via holes, repair of embedded interconnects, and thick/thin film resistor and capacitor trimming. The most popular lasers for cutting materials are CO_2 and Nd:YAG, although semiconductor lasers are gaining prominence because of their higher efficiency. Because laser micromachining or direct ablation needs high laser intensity or fluence, usually only excimer lasers are used, especially for hard materials.

In recent years, excimer lasers, providing ultrafast durations of < 10 ps (10^{-11} s) based on diode-pumped solid-state gain media, have demonstrated remarkably clean micromachining of many materials. Despite initially gaining a poor reputation for reliability, during the past 25 years, excimer lasers have developed into mature industrial tools. For laser ablation of polymers, argon ion lasers, a kind of continuous laser, can also be used, and they can be modulated into different pulse widths, frequencies, and wavelengths. However, the excimer laser has several distinct advantages, even for the dry etching of polymers, over the CO_2 and Nd:YAG lasers. For example, the wavelengths are more compatible with the chemical bond energies in organic compounds and tend to produce less thermal damage. Excimer lasers, as we ascertained in Chapter 1, are also used in the newest lithography processes. A visual illustration is obtained comparing the hole-drilling capabilities of mechanical twist drill bits and excimer lasers in Figure 5.27. Here we compare 100-μm-diameter holes drilled in 75-μm-thick high-density polyethylene with a drill bit and a KrF excimer laser. Although mechanically the hole is close to the minimum size that can be drilled, the improved quality of this and much smaller holes drilled by excimer lasers is obvious.

Laser machining with chemical assist takes care of some of the problems associated with long laser pulse operation, but, as we have already seen, ultrashort laser pulsing takes care of most of them. Laser cutting machines have been used extensively in industry for cutting complex profiles in sheet and plate materials. Hard materials such as diamond, tungsten, and titanium can be cut without problems of tool wear because there is no contact between tool and work piece. For many applications, laser micromachining offers the advantage of more rapid prototyping as compared to, e.g., FIB machining or lithography techniques.

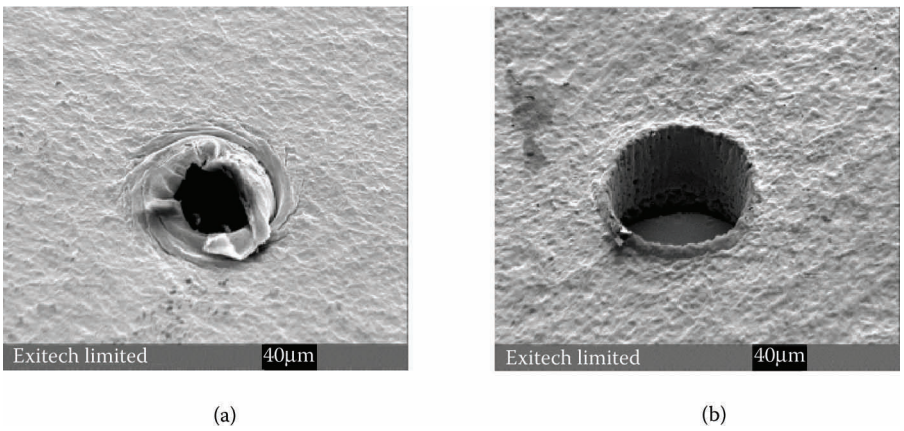


FIGURE 5.27 Holes drilled with twist drill bit (a) and KrF laser (b). (From Exitech.)

For a particular micromachining application the choice of laser type is now judged as much by criteria such as process speed, part throughput, reliability, service intervals, capital and operating costs of the overall machine tool rather than solely by the quality of the processed part.

Long Laser Pulse Machining Applications

A simple example of long laser pulse machining is a hole, laser drilled through glass or Si, that can be used as a high-aspect-ratio, through-the-wafer via.[21,22] This is illustrated for the electrical contact via of a solid-state capacitive pressure sensor in Figure 5.28.[21] This compact pressure sensor incorporates electrically conducting vias drilled through the glass cap by a CO_2 laser. In contrast, the through-the-wafer via connection fabricated by anisotropic etching has a large opening angle (54.7°) associated with the slow etching crystallographic planes of the Si substrate. Typically, the closest spacing for vias that are wet etched in a 300-µm-thick wafer is approximately 425 µm, thus limiting this method to a low density of via holes (different means of making vias are discussed in Volume III, Chapter 4).

When holes as shown in Figure 5.28 are laser drilled in a substrate, the expelled material solidifies into dust; therefore, a moving protective tape is used in front of the laser lens to keep it clean. Debris also collects around the perimeter of the openings in, e.g., a silicon substrate, when drilling, scribing, or trenching is done in air or an inert atmosphere.

To avoid the identified problems associated with long laser pulses, two methods can be used. One is to use laser-assisted chemical etching (LACE), as described above, which takes care of some of the mentioned problems; the other approach, taking care of most of the mentioned problems, is to use femtosecond laser pulses. The latter technique, as we have explained above, results in superior micromachining.

Short Laser Pulse or Excimer Laser Applications

Introduction Energy-deposition time with ultrafast excimer lasers is short compared with the electron-phonon time (a few picoseconds); thus, thermal conduction and hydrodynamic motion are generally negligible, and ejected material carries away most of the deposited energy. As a result, there is minimal collateral damage to remaining material. In metals, laser energy is deposited over the skin depth rather than spread over a thermal conduction length, yielding very high energy densities, temperatures (10 eV), and pressures (Mbar) during the pulse that are not that different from those in the sun (Figure 5.29). In Figure 5.29 we compare existence domains for various types of plasmas as a function of temperature and density. In dielectrics, electrons that are produced by nonlinear light absorption initiate the breakdown, and full breakdown is achieved during each pulse.

Excimer laser materials processing, by providing solutions to critical problems in manufacturing of integrated circuits, hard disks, displays, interconnects, desktop printers, and telecommunication

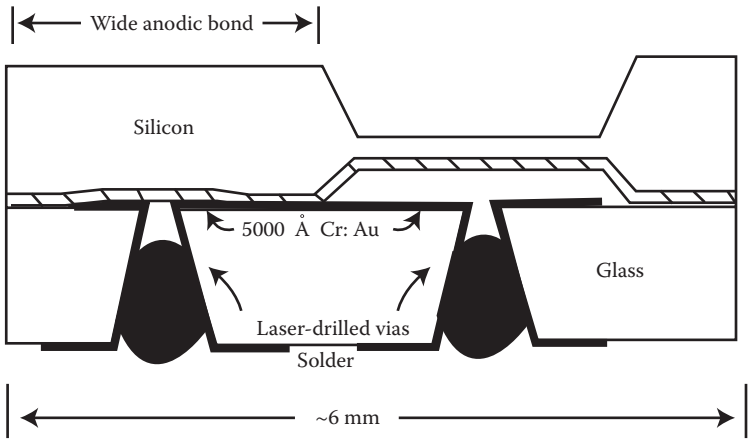


FIGURE 5.28 Electrical contact in a silicon capacitive pressure sensor using laser-drilled vias. (From Bowman, L., J. M. Schmitt, and J. D. Meindl. 1985. *Micromachining and micropackaging of transducers.* Eds. C. D. Fung, P. W. Cheung, W. H. Ko, and D. G. Fleming. Amsterdam: Elsevier. With permission.[21])

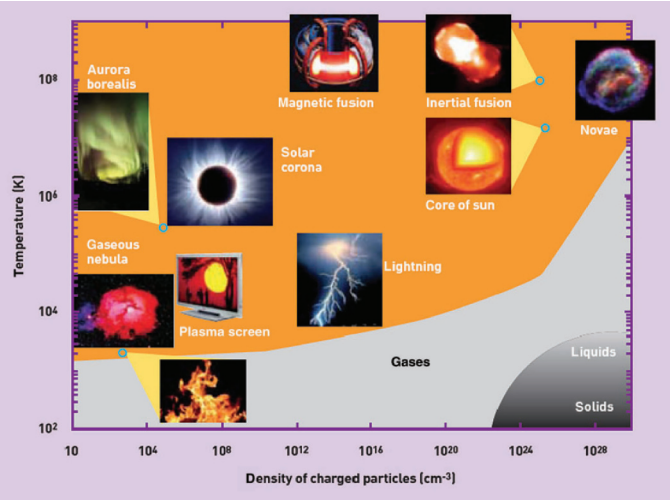


FIGURE 5.29 Existence domains for various types of plasmas as a function of temperature and density. (Extreme measurements in laser generated plasmas, CLEFS CEA-No. 4, Autumn 2006.[25])

devices, has become a key enabling technology for the current revolution in information technology.

Ink-Jet Printers, Microvias, Catheters, and DVDs

Excimer lasers on production lines in Asia, Europe, and the United States now drill nozzle-hole arrays for most ink-jet printers. Earlier 300-dpi printers consisted of a 100-nozzle row of 50-µm-diameter holes made by electroforming thin nickel foil. Trying to fabricate more holes with smaller diameters reduced the already low 70–85% production yield even further. Increased printer quality is achieved by simultaneously reducing the nozzle diameter and decreasing the hole pitch. Modern printers like Hewlett Packard's (Palo Alto, CA) DeskJet 800C and 1600C have nozzles with 28-µm diameters, which provide resolution of 600 dpi. Laser drilling of nozzle arrays allows average yields greater than 99%. Excimer laser tools with appropriate CNC programming can readily engineer custom-designed reverse-tapered, $2\frac{1}{2D}$ and 3D structures. Some applications need a rifled, tapered hole to spin the droplet and optimize its accuracy of trajectory. Most of the ink-jet printer heads being currently sold (e.g., by Hewlett Packard and Canon) are excimer laser drilled on production lines in the United States and Asia. Figure 5.30 shows an excimer laser-drilled nozzle array in a modern print head.

Another example application of excimer lasers is drilling of microvia much smaller than the ones depicted in Figure 5.30. Packages on which chips are mounted for connection to other devices have to keep pace with the rapid advances made in ICs, so that speed, power, and area (real estate) do not become compromised. In multilayer sandwiches, blind via holes provide high-speed connections between surface-mounted components on the board and underlying power and signal planes while minimizing valuable real estate occupation. The cost of drilling these vias on such high-density packages can represent 30% of the overall cost of the board. The process was first applied to volume production when the Nixdorf computer plant of Siemens introduced polyimide ablative drilling of 80-µm-diameter vias in multichip modules. Other mainframe computer manufacturers such as IBM rapidly followed suit and installed their own production lines for this application. Trillions of vias have now been drilled with excimer lasers at

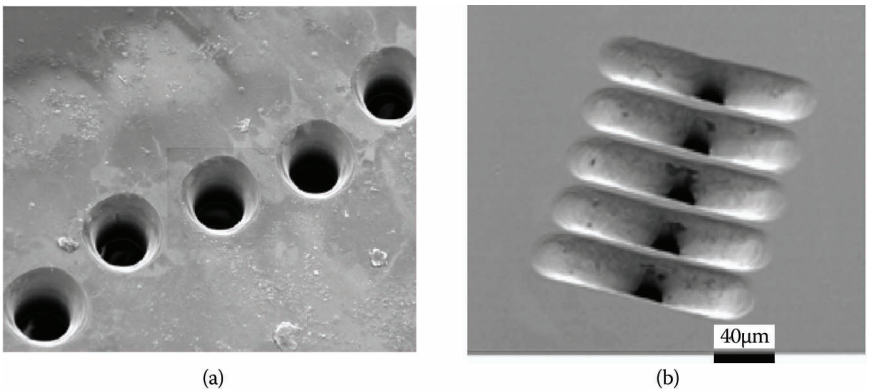


FIGURE 5.30 KrF laser-drilled ink-jet printer nozzles in polyimide. (a) 30-µm diameter nozzle array. (b) Nozzles with nonlinear tapers to aid the laminar flow of the droplet through the orifice. (From Excitech.)

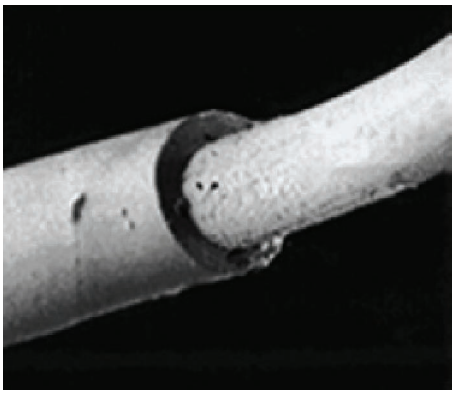


FIGURE 5.31 Stripping of wires with excimer lasers.

yields greater than 99.99% and mean time between failure at greater than 1000 hours.

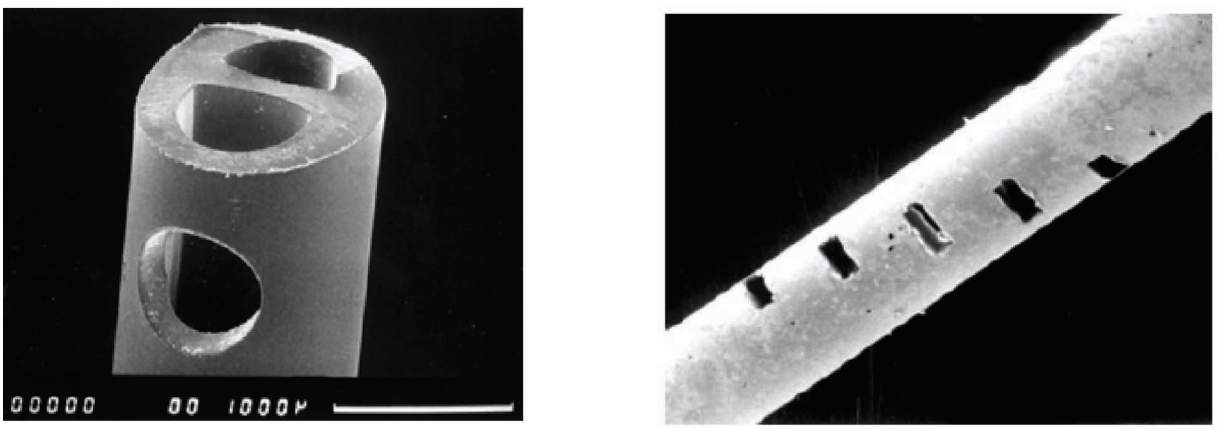
Preferential excimer laser etching of plastics against metals is illustrated by the stripping of insulation from insulation sleeving of wires with widespread use in preparing connection wires to computer hard-disk reader heads (see Figure 5.31).

In the biomedical industry, the drive for increased miniaturization with improved device functionality is crucial to rapid progress, for example, in novel catheter-based sensors and actuators. Precision microdrilling with excimer lasers is routine when making delicate catheter-based probes. Here we explore, as a futuristic example, a fiberoptic-based sensor for analyzing arterial blood gases (ABGs) and pH. The use of fiberoptic sensors for ABG analysis provides clinical diagnostics at the patient's bedside without the need for taking blood samples or doing an analysis in the blood gas laboratory (except for a much less frequent calibration of the in vivo sensors). Envision one lumen of a bilumen catheter for blood drawing and one for the sensing optical fiber as shown in Figure 5.32.

The hole at the side of a polyvinyl chloride bilumen sleeving tube through which blood is drawn (to collect blood for the blood gas lab for calibration of the in vivo sensor) is machined using a krypton fluoride (KrF) laser. The clean-cutting capability of the laser provides the catheter with the necessary rigidity that prevents kinking and blockage of the tube when inserted into the artery. The other important components of this in vivo catheter are the oxygen, CO_2 sensors, and pH sensors on the side of a fiberoptic cable placed in the other lumen. These consist of ~50 × 15-μm rectangular holes machined in a 100-μm-diameter acrylic [poly(methylmethacrylate) (PMMA)] optical fiber with an argon fluoride (ArF) laser. The holes are filled with reagents whose optical transmission depends on the O_2, CO_2, and pH levels of the surrounding blood.

For an alternative ABG and pH sensor made with Si bulk micromachining, see Figure 4.92 depicting an electrochemical ABG and pH sensor.

Another typical application of laser micromachining is to increase the density of optical storage media such as DVDs. Currently, the track spacing on a DVD is 400 nm, yielding approximately 6 GB of data storage space. The initial target spacing for the next generation of laser micromachining systems is 100 nm, resulting in 25-GB storage capacity on a DVD.



(a) (b)

FIGURE 5.32 Excimer laser manufacture of blood gas sensor. (a) KrF laser-drilled hole in side of polyvinyl chloride bilumen catheter. (b) ~50 × 15-μm rectangular holes machined in a 100-μm diameter acrylic [poly(methylmethacrylate) (PMMA)] optical fiber with an ArF laser. (From Excitech.)

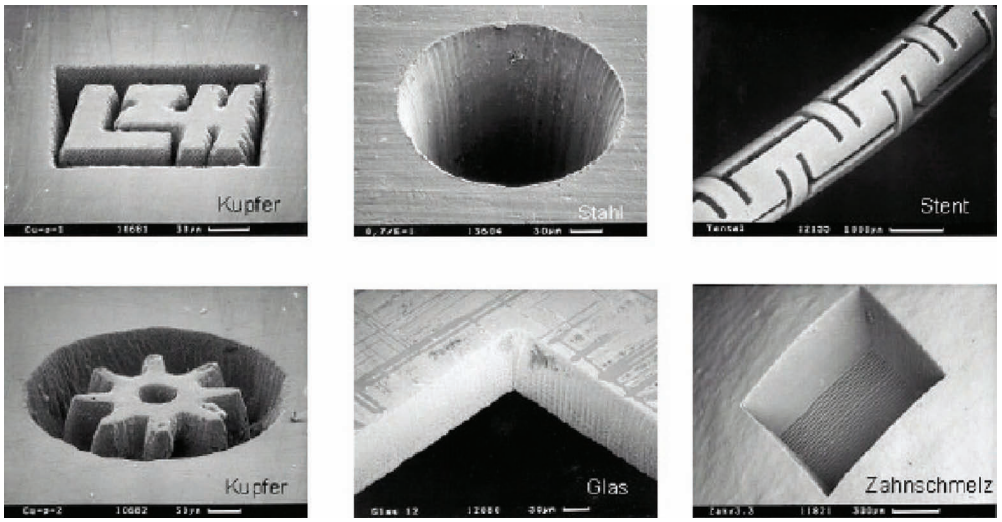


FIGURE 5.33 From the Laser Centrum at Hannover, Germany (LZH): laser machining in a wide variety of materials.

The excimer laser also has brought more accuracy to corneal surgery and vision correction than ever before. One pulse of the excimer laser removes 0.25 µm of tissue. To put this into perspective, a typical human hair is 70 µm in thickness. Surgeons have never had a device as accurate as the excimer laser in eye surgery.

At the Laser Centrum in Hannover (LZH), Germany, researchers use femtosecond lasers for hole drilling with pulse durations shorter than 100 fs and pulse peak powers in the gigawatt range.[10] Grooves of only 200-nm thickness can be cut. Some demonstration pieces are shown in Figures 5.33 and 5.34.

Machining inside Transparent Materials Ultrafast laser pulsing enables machining inside transparent bulk materials. A material such as a glass is transparent and does not absorb visible light—provided the intensity stays below the threshold for multiphoton absorption. One can exceed that threshold by focusing ultrafast laser pulses to a spot inside the material. When the intensity exceeds the threshold for plasma formation, very localized absorption occurs at the focal spot, and, as before, the plasma expands, but this time it is confined by the surrounding material. The effect of the expansion is to create a void within a very dense shell of material—a cavity within the glass itself. The process is not limited to glass. Pits can be created in any material by focusing an ultrafast laser pulse inside the material, whether it is amorphous or crystalline.

Based on the above properties of femtosecond lasers, one can manufacture single-mode waveguides in glasses. Glass is transparent to incident, ultrafast laser pulses at 775 nm, enabling the local melting of the glass via confined multiphoton absorption and avalanche ionization. The glass then resolidifies, changing its physical properties. The result is an index gradient that acts like a waveguide. Only ultrafast lasers are capable of producing this effect in transparent materials; with longer pulse lasers (nanoseconds), the glass incurs damage before the intensity reaches the threshold at which guides are formed. Femtosecond laser pulses can be used to induce localized refractive index increase in a wide variety of glasses, and this way optical waveguides have been produced in silicate, borosilicate, chalcogenide, and fluoride glasses.

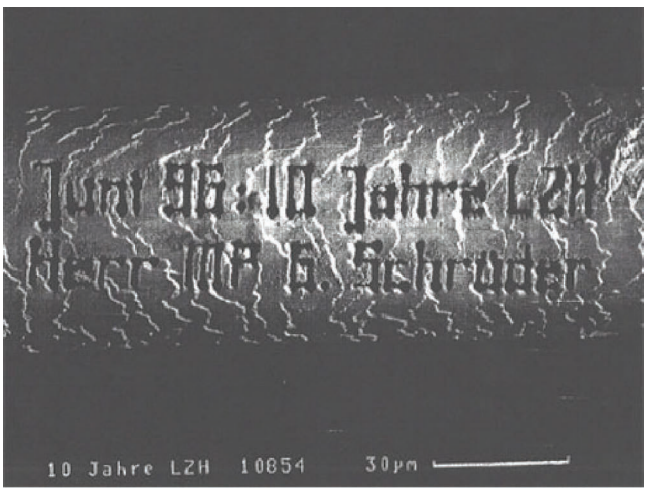


FIGURE 5.34 From the Laser Centrum at Hannover, Germany: Micromarking (G. Schröder's hair). ArF laser-micromachined 40-µm square holes through a human hair.

Table 5.8 Laser Applications in Various Industrial Fields

Heavy Manufacturing	Light Manufacturing	Electronics	Medical	General
Profile cutting in sheet and plate metals	Profile cutting in plastics and wood	Via formation insulating material: Tab and MCM	Flow orifices <100 μm diameter	CVD diamond cutting
Seam and spot welding	Engraving	High-accuracy wire stripping	Drilling and cutting delicate or thermally sensitive materials	Ceramic and glass micromachining
Cladding and drilling	Drilling	Skiving of flexible circuits	Micromachining applications where edge quality and cleanliness are critical	Thin film patterning; microlithography
		IC repair		

Summary: Laser Machining Applications and Comparisons

Subtractive and additive laser machining (see Chapter 9) present several application opportunities in micromachining because of their versatility, site-specific operation, and rapid prototyping capability. Ultrafast laser machining, when lasers become less expensive, will make laser micromachining yet more attractive. There are some obvious drawbacks with these direct write techniques: they are serial processes, and the deposition or removal rates limit the speed of the micromanufacture. However, for rapid prototyping, mold fabrication, and site-specific manufacturing, laser machining has a very bright future, and, as we have remarked before, many microsystems might carry a bigger price tag than an IC, making a more expensive manufacturing technology acceptable.

In Table 5.8 we list typical laser applications in different fields, and in Table 5.9 we compare laser micromachining with some other typical competing machining technologies.

A good further introduction to laser machining can be found in *Laser Machining: Theory and Practice* by George Chryssolouris[23] and in an excellent chapter by Henry Helvajian, "Laser Material Processing: A Multifunctional in situ Processing Tool for Microinstrument Development."[11] A very informative laser handbook on the Internet can be found at http://www.clark-mxr.com/industrial/handbook/introduction.htm. An excellent tutorial on the Internet is by Dieter Schuöcker, Chair of Nonconventional Processing, Forming, and Laser Technology at the TU Vienna, Austria, in cooperation with Arnold Braunsteiner, at http://www.argelas.org/files/tutorium/tutorium.htm. Laser tutorials may also befound at http://www.ilt.fraunhofer.de/eng/100048.html by Fraunhofer Institut Lasertechnik, http://www.coherentinc.com/Literature/ by Coherent, and http://www.colemaneye.com/fundamentals/laser.htm by Aldrich Computers & Consultants.

Also Rami Arieli's "The Laser Adventure" at http://stwww.weizmann.ac.il/lasers/laserweb/index.htm is a good laser introduction. A service

Table 5.9 Comparing Laser Micromachining with EDM, Mechanical Machining, and Chemical Etching

	Practical Resolution Limit	Attainable Aspect Ratio	Taper	Undesirable Side Effects	Status of Technology Development
Excimer laser	5 μm	>100:1	Yes	Recast layer	Low
CO_2 laser	200 μm	100:1	Yes	Recast layer, burring, thermal	High
Nd:YAG	50 μm	100:1	Yes	Recast layer, burring, thermal	High
EDM	100 μm	20:1	No	Surface finish	Moderate
Chemical etch	250 μm	1:1.5	Yes	Undercutting	Moderate
Mechanical	Ø 100 μm	10:1	No	Burring	Moderate

provider in the laser micromachining area with an excellent educative web site is Resonetics (http://www.resonetics.com/MDmfg.htm). Another web site to bookmark is Sam's Laser FAQ (http://www.repairfaq.org/sam/lasersam.htm). A laser link page can be found at http://www.mysteries-megasite.com/main/bigsearch/laser.html.

Electron Beam Machining

Electron beam removal of materials is another fast-growing thermal technique. Instead of electrical sparks, this method uses a stream of focused, high-velocity electrons from an electron gun to melt and vaporize the work piece material.[24] In EBM, electrons are accelerated to a velocity of 200,000 km/s or nearly three-fourths that of light. As shown in Figure 5.35, the process is performed in a vacuum chamber to reduce the scattering of electrons by gas molecules in the atmosphere. The stream of electrons is directed against a precisely limited area of the work piece; on impact, the kinetic energy of the electrons is converted into thermal energy that melts and vaporizes the material to be removed, forming holes or cuts. Originally, in the early 1960s, the method was used for welding and annealing, but today it is also used for many IC and micromachining tasks. EBM equipment now is commonly used by the electronics industry to aid in the etching of masks and the fabrication of circuits in microprocessors; today, scanning electron beam lithography remains one of the best-demonstrated methods for microfabrication in the sub-100-nm regime (see Part I). We will not repeat the electron beam lithography aspects of electron beam machines here but rather emphasize other machining applications. An electromagnetic coil or lens focuses the diverging electron beam to a power density of about 6500 GW/mm^2, high enough to vaporize the work piece material. A typical cross-sectional diameter of the beam is 10–200 μm at the point of impact on the work piece. A bias electrode switches the electron beam energy from the cathode (electron gun) on and off for pulsed drilling operations. The bias electrode also controls beam intensity. With additional coils, the beam can be deflected in any direction. With mechanical means, holes of a diameter of 0.8 mm are difficult to drill, and those less than 0.1 mm are nearly impossible. In contrast, EBM works well for deep holes of diameters less than 0.1 mm and is generally effective with a diameter-to-depth ratio of 1:10 in hard-to-machine materials. A hole in a sheet 1.25-mm thick up to 125 μm in diameter can be cut almost instantly with a taper of 2–4°. With multiple pulses, diameter-to-depth ratios can reach 1:100. For example, hole diameters of 0.1 mm can be drilled in materials up to 10-mm thick.

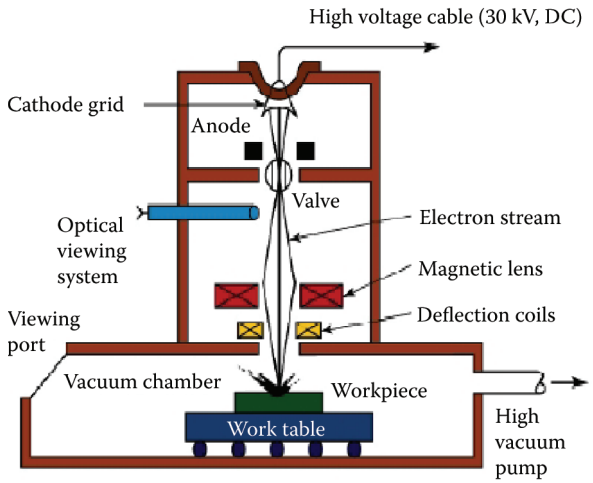


FIGURE 5.35 Schematic illustration of the electron beam machining process. Unlike LBM, this process requires a vacuum; therefore, work piece size is limited to the size of the vacuum chamber.

Once the vacuum is established (10^{-4} mm Hg), EBM can be extremely fast. The material removal rate is about 0.1 mg/s; hence drilling rates of 2000 holes/s are routine with EBM. Typical tolerances are about 10% of the hole's diameter (e.g., 0.005 mm on a 0.05-mm hole). The most suitable application is for work pieces requiring large numbers of simple small holes to be drilled or for drilling holes in materials that are hard and difficult to machine with other processes. The equipment is very expensive, and the need to operate in vacuum adds considerably to the machining time. The cost of an EBM facility ranges from $75,000 to $1.5 million.

Plasma Beam Machining

Plasma arc cutting (also plasma arc machining) is mainly used for cutting thick sections of electrically conductive materials (Figure 5.36). A high-temperature plasma stream (up to 60,000°F)

FIGURE 5.36 Plasma arc cutting.

interacts with the work piece, causing rapid melting. A typical plasma torch is constructed in such a way that the plasma is confined in a narrow column about 1 mm in diameter. The electrically conductive work piece is positively charged, and the electrode is negatively charged. Relatively large cutting speeds can be obtained: for example, 380 mm/min for a stainless steel plate 75-mm thick at an arc current of 800 A. Tolerances of ±0.8 mm can be achieved in materials of thicknesses less than 25 mm, and tolerances of ±3 mm are obtained for greater thicknesses. The HAZ for plasma arc cutting varies between 0.7 and 5 mm in thickness, and the method is used primarily for ferrous and nonferrous metals. The inert gas to form plasma also serves as a shielding gas. The torch and sometimes the work piece require cooling, and the process produces high noise levels. The technique is of marginal use in micromachining, except perhaps to cut the plasma arc-coated sensor substrates discussed in Chapter 9.

For a detailed history of plasma cutting, visit http://www.hypertherm.com/technology/plasma_history.htm. A detailed analysis of the quality of the cut of a plasma arc is presented at http://archive.metalformingmagazine.com/1998/05/plasma/plasma.htm.

Acknowledgments

Special thanks to Dr. Rodrigo Martinez, UCI Graduate now at ETH in Laussane.

Questions

Questions by Mr. Chuan Zhang, UC Irvine

5.1: In the electrical discharge machining (EDM) process, the spark-gap width is an important parameter in determining the accuracy of the process. Discuss the pros and cons of a very small spark-gap width.

5.2: What are the merits and drawbacks of electrical discharge machining? List some applications of EDM.

5.3: Sketch a way to make a 300-μm-diameter Ni rod with a polygonal cross-section using electro discharge machining wire cutting.

5.4: Draw two shapes that are better produced by lithography, and two shapes that are better made by wire cutting discharge machining (WC-EDM).

5.5: How can we perform μ-EDM in batch mode by relying on LIGA?

5.6: Explain why the metal removal rate in EDM is much slower than that of ECM.

5.7: Explain the energy conversion process in laser beam machining (LBM). How can we thermally remove materials by consuming electrical power?

5.8: Why is the mechanism of laser ablation different for different laser light wavelengths?

5.9: Why is a laser more efficient in machining than visible light? List the common applications of laser machining.

5.10: How can the quality of the ablated region be improved in the laser ablation process? Which factors affect the thickness of the heat-affected zone (HAZ)?

5.11: What are the important factors influencing the efficiency and spatial resolution of laser ablation? Explain why.

5.12: What is the difference between optical penetration depth and damage penetration depth in laser machining?

5.13: How does the laser pulse length affect laser beam machining? Compare long laser pulse etching and short laser pulse etching.

5.14: How can water jet guided laser machining keep a workpiece clean? How does it work?

5.15: What are the merits of laser etching with chemical assist (LACE)? What are the technical barriers for applying this technique in most manufacturing fields?

5.16: Laser machining offers advantages compared to methods such as EDM, ECM, chemical etching, or stamping. A laser can easily cut complex parts with high tolerances. Excessively hard, pressure- and force-sensitive, or very thin materials can be cut with a small focused beam. In many cases, laser cutting is preferred over stamping due to its high flexibility and noncontact characteristics. Lasers also enable low-temperature deposition processes, rapid prototyping, mold fabrication, and site-specific manufacturing. It also has obvious drawbacks. What are they?

References

1. Kalpajian, S. 1984. *Manufacturing processes for engineering materials.* Reading, MA: Addison-Wesley.
2. Snoeys, R., F. Staelens, and D. F. Dauw. 1986. Adaptive control optimization as basis for intelligent EDM die-sinking machines. *Proceedings of the Winter Annual Meeting of ASME.* Anaheim, CA, 22:63–78.
3. Guitrau, E. B. 1997. *The EDM handbook.* Cincinnati, OH: Hanser Gardner Publications.
4. Rain, C. 1981. Nontraditional methods advance machining industry. *High Technology* November/ December:55–61.
5. Saito, N. 1984. Recent electrical discharge machining (EDM) techniques in Japan. *Bull Jpn Soc Precis Eng* 18:110–16.
6. Sommer, C., and S. Sommer. 2000. *Wire EDM handbook.* Houston: *Advance Publishing.*
7. Kuo, C.-L., T. Masuzawa, and M. Fujino. 1991. *Proceedings: IEEE micro electro mechanical systems (MEMS '91).* Nara, Japan.
8. Kuo, C.-L., T. Masuzawa, and M. Fujino. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92).* Travemunde, Germany.
9. Takahata, K., and Y. B. Gianchandani. 2002. Batch mode micro-EDM for high-density and high-throughput micromachining. *IEEE/ASME J Microelectromechanical Systems* 11:102–110.
10. Ehmann, K. F., D. Bourell, M. L. Culpepper, T. J. Hodgson, T. R. Kurfess, M. Madou, and K. Rajurkar. 2007. *Micromanufacturing: international research and development.* New York: Springer.
11. Helvajian, H. 1995. *Microengineering technology for space systems.* Ed. H. Helvajian. El Segundo, CA: Aerospace Corporation.
12. Srinivasan, R., and V. Mayne-Banton. 1982. Self-developing photoetching of poly(ethylene terephthalate) films by far-ultraviolet excimer laser radiation. *Appl Phys Lett* 41:576–78.
13. Liu, Y. S. 1989. *Laser microfabrication: thin film processes and lithography.* Eds. J. Y. Tsao, and D. J. Ehrlich. New York: Academic Press.
14. Ehrlich, D. J., D. J. Silversmith, R. W. Mountain, and J. Tsao. 1982. Fabrication of through-wafer via conductors in Si by laser photochemical processing. *IEEE Trans Compon Hybrids Manuf Technol* CHMT-5:520–21.
15. von Gutfeld, R. J., and R. T. Hodgson. 1982. Laser enhanced etching in KOH. *Appl Phys Lett* 40:352–54.
16. Bloomstein, T. M., and D. J. Ehrlich. 1992. Laser-chemical three-dimensional writing for microelectromechanics and application to standard-cell microfluidics. *J Vac Sci Technol B* 10:2671–74.
17. Shlichta, P. J. 1988. Laser micromachining in a reactive atmosphere. *NASA Tech Briefs* 12:84.
18. Maeda, S., K. Minami, and M. Esashi. 1994. *IEEE international workshop on micro electro mechanical systems (MEMS '94).* Oiso, Japan: IEEE.
19. Bloomstein, T. M., and D. J. Ehrlich. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91).* San Francisco.
20. Eden, J. G. 1986. Photochemical processing of semiconductors: new applications for visible and ultraviolet lasers. *IEEE Circuits Devices Mag* 2:18–24.
21. Bowman, L., J. M. Schmitt, and J. D. Meindl. 1985. *Micromachining and micropackaging of transducers.* Eds. C. D. Fung, P. W. Cheung, W. H. Ko, and D. G. Fleming. Amsterdam: Elsevier.
22. Anthony, T. R., and H. R. Cline. 1976. Deep-diode arrays. *J Appl Phys* 47:2550.
23. Chryssolouris, G. 1991. *Laser machining.* New York: Springer Verlag.
24. Taniguchi, N. 1984. Research on and development of energy beam processing of materials in Japan. *Bull Jpn Soc Precis Eng* 18:117–25.
25. Extreme measurements in laser generated plasmas, CLEFS CEA-No. 4, Autumn 2006.

6

Mechanical Energy-Based Removing

Manufacturing dominates world trade. It is the main wealth creating activity of all industrialized nations and many developing nations. A manufacturing industry based on advanced technologies with the capability of competing in world markets can ensure a higher standard of living for an industrial nation.

P. McKeown
(*1996*)

Outline

Introduction

History of Mechanical Machining

Absolute and Relative Tolerances in Manufacturing

Ultraprecision Mechanical Machining

Desktop Factory

Precision Mechanical Machining versus IC-Based Micromachining

Questions

References

Commercial desktop factories (DTFs) at Sankyo Seiki.

Fanuc's the ROBOnano U_i, an ultraprecision micromachining station (cost $1 million) and a Noh mask made with this machine (http://www.fanuc.co.jp/en/product/robonano/index.htm).

Introduction

In mechanical subtractive machining, physical removal of unwanted material is achieved by mechanical energy applied at the work piece. In the ultraprecision range, the two major categories of mechanical material-removing technologies we detail in this chapter are single-point machining and abrasive machining (basically multipoint machining). Abrasive machining is a process that removes small amounts of material from a work piece in the form of tiny chips through the contact of small, hard, sharp, nonmetallic particles. In abrasive machining we review abrasive wheel machining, in particular electrolytic in-process dressing (ELID), ultrasonic machining (USM), water jet, abrasive water jet (AWJ), and abrasive jet machining (AJM). We also consider briefly two energetic beam mechanical machining methods, i.e., focused ion beam (FIB) and fast atom beam (FAB) machining. We also cover the miniaturization of mechanical machining equipment in a section on "desk-top factories." The fact that it often takes a two-ton machine tool to fabricate microparts, where cutting forces are in the milli- to micro-Newton range, is a clear indication that a complete machine tool redesign is required for the fabrication of micromachines.

Whereas primary manufacturing processes involve casting and molding, secondary manufacturing processes constitute the main mechanical removing techniques involving turning, drilling, and milling. Abrasive processes to superfinish a work piece are called tertiary manufacturing processes.

History of Mechanical Machining

Tool making for mechanical machining began perhaps when humans learned to walk erectly and had their hands free to grasp objects of wood or stone and to shape them. Although tool-making hominids probably emerged millions of years earlier, we can only trace tool-making history back 2 million years to the oldest surviving and recognizable flint hand axes.[1] Stronger, longer lasting tools resulted when metals such as copper and iron replaced wood and stone. Humans have been cutting metals for thousands of years to make the myriad tools, machines, and other devices that civilization demands.

The birth of more precise mechanical machining dates back to the invention of the lathe, attributed to Anacharis the Scythian and first depicted in an Egyptian low relief about 300 B.C. (Figure 6.1).[1] Hipparchus, in the second century B.C., and Ptolemy, in 150 A.D., used graduated instruments for astronomy and sailing. Ptolemy's astrolabe became the basic tool for observers of the skies for the next thousand years; the Arabs perfected it in the ninth century, and it remained in use for another 700 years (Figure 6.2).[2] Precision machining with a hand-operated lathe got a boost in the mid-1600s from demands for more precision in building better instruments for time measurement.[3] Ruling engines for the control of diffraction gratings have driven state-of-the-art precision machine tool design since the late 1700s. More than a century ago, Rowland at the John Hopkins University built ruling engines with part-per-million accuracy.[4,5] Today, the mechanical technology on which photolithographic alignment systems are based is still grounded in the same ruling engine design. An American astronomer, David Rittenhouse, made the first diffraction grating in 1785, when he fabricated a half-inch-wide grating with 53 apertures.[6] It is difficult to identify another single device that has provided more important experimental information to every field of science than the diffraction grating.[6] Rowland not only constructed sophisticated ruling engines, but he also invented

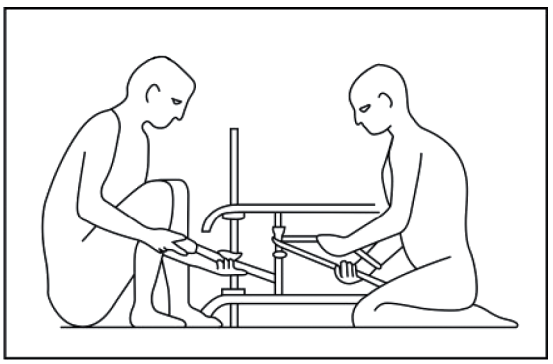


FIGURE 6.1 First lathe as depicted in an Egyptian bas relief, about 300 B.C. Shown here in a line drawing. The man at left is holding the cutting tool. The man at right is making the work piece rotate back and forth by pulling on a cord or thong.

FIGURE 6.2 Ptolemy's astrolabe; front view (a) and back view (b).

the concave grating, a device of spectacular value to modern spectroscopists.

Tools driven by mechanical power brought on the machine age in the late 1700s.[7] Diamond as a cutting tool was apparently first used in 1779 for making threads in hardened steel. The significance of diamond tools in precision machining was realized in the 1850s. By 1920, the Lord's Prayer had been engraved with a diamond tool into an area of 100 μm × 40 μm. In the 1930s, watchmakers used diamond turning to make watch dial components with excellent surface finishes. In the 1940s, Moore Special Tool Company designed a jig grinder that allowed toolmakers to work with accuracies that were not achievable by any of the existing methods (http://www.mooretool.com).

In modern times, precision engineering was initially pushed by nuclear programs and eventually by semiconductor manufacturing. By the 1960s demands from electronics, defense, and energy applications led to pioneering work on diamond turning from Lawrence Livermore National Laboratory and Oak Ridge National Lab. In the 1970s, the need for large optics in space telescopes and "defense" systems triggered the application of precision machining to complex optical components.

Norio Taniguchi started working on processing of materials to nanoscale precision as early as 1940 and worked on achieving precision finishes of quartz crystals, silicon, and other hard and brittle materials by using ultrasonic machining (USM). Taniguchi coined the term *nanotechnology* and in 1974 used the term to define ultraprecision machining. According to Taniguchi, ultraprecision machining is "the process by which the highest possible dimensional accuracy is achieved at a given point in time." In an ultraprecision manufacturing system, all the material processing methods used in a conventional manufacturing system, like removal, addition, joining, surface treatment, and assembly, are scaled down by orders of magnitude. The machine tool capabilities, metrology, and instrumentation used in these processes range in the orders of micrometers to nanometers.

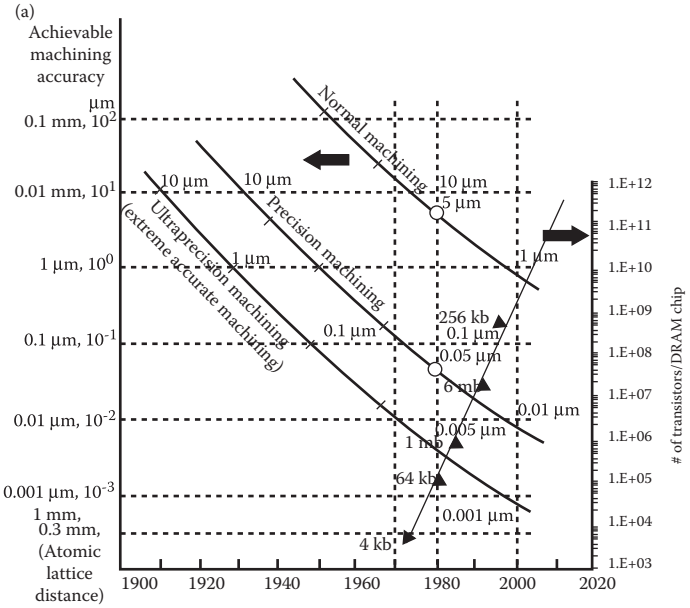
Figure 6.3 includes the machine accuracy capabilities Norio Taniguchi predicted along with the processes or tools used to achieve it.

In the past 25 years, great progress has been made in building machines that can be operated and controlled automatically.

By 1977, highly precise instruments such as servomotors, feedback devices, and computers were implemented, paving the way for computer numerical control (CNC) machining, which is now standard in many types of machine tools. At the start, the smallest movement these machines could reproducibly make was 0.5 μm.

A crucial development came in the 1980s, when advanced machine tools became equipped with precision metrology and control tools. These machines used laser interferometer and capacitance probe feedback controls, temperature control, and hydrostatic

(a)



(b)

Machine Tools (Processing equipment)	Measuring Instruments (Inspection equipment)
Lathes, milling machines	Vernier calipers
Precision lathes, grinding machines, lapping machines, honing machines	Mechanical comparators, micrometers, dial indicators
(Curve generators) Precision jig borers, jig grinding machines, super finishing machines	(Electric or pneumatic micrometers) Optical comparators
Precision grinding machines, precision lapping machines (optical lens grinding machines) Precision diamond lathes	Optical or magnetic scales, electrical comparators Electronics comparators (noncontact)
(DUV mask aligner) Optical lens finish grinding machines, precision diamond grinding machines, ultraprecision grinding machines	Laser measuring instruments Optical fibers, Talysurfs, Talyronds
(Diffraction grating ruling engine) (Electron beam lithographer) Super high-precision grinding machines, super high-precision lapping and polishing machines	High-precision laser measuring instruments (Doppler, multireflection) Talysteps
Atom, molecule, or ion beam machining, atom or molecule deposition (Substance synthesizing)	Scanning electron microscopes, transmission electron microscopes, electron diffraction equipment, ion analyzers, x-ray microanalyzers, Auger analyzers, ESCA
Atom manipulation with proximal probes	STM and AFM

FIGURE 6.3 Progress of accuracy in machining. The development of achievable machining accuracy during the past 100 years is shown in this figure under the generalized classification of normal machining, precision machining, and ultraprecision machining. (a) Left ordinate, achievable manufacturing accuracy over time based on Taniguchi. (Taniguchi, N. 1983. Current status in, and future trends of, ultraprecision machining and ultrafine materials processing. *Annals CIRP* 32:S573–82.[8]) Right ordinate, increase in transistor density over time according to Moore's Law. (Gibbs, W. W. 1997. *Scientific American* September 22.[9]) (b) Machines, processing equipment, and metrology equipment by which the indicated resolution can be achieved and checked. By "ultraprecision machining," Taniguchi means those processes/machines by which the highest possible dimensional accuracy is or has been achieved at a given point in time.

bearings and featured accuracies better than 0.1 μm. Precision manufacturing methods were extended for industrial use for cutting aluminum, which was used for making components for scanners, photocopying machines, and computer memory disks. Also in the 1980s, cutting with very small diamond tools (e.g., 22-μm diameter) was developed in Japan.

In the 1990s, precision manufacturing became accepted as another advanced tool in manufacturing systems with some machining stations even allowing freeform machining. Large-scale adaptation of CNC machining also occurred during this period as a result of the surge of the electronics and computer industries.

By 1993, 0.05 μm became possible, and today there is equipment available featuring 0.01 μm and even nanometer step resolution (http://www.fanuc.co.jp/eindex.htm).[10]

This evolution closely follows the predictions sketched in Figure 6.3, where the Taniguchi curves show a machining accuracy for ultraprecision machining of subnanometer resolution for 2008. The resolution of the steps a machine can make, of course, is a determining factor for the manufacturing accuracy of the work piece.

Historically important milestones in new machining concepts include the introduction of ultrasonic techniques, focused ion beam, and desktop factories (DTFs).

The roots of ultrasonic technology can be traced back to research on the piezoelectric effect conducted by Pierre Curie around 1880. He found that asymmetrical crystals such as quartz and Rochelle salt (potassium sodium titrate) generate an electric charge when mechanical pressure is applied. Conversely, mechanical vibrations are obtained by applying electrical oscillations to the same crystals. This led to the discovery of SONAR (sound navigating and ranging). In USM, the tool, made of softer material than that of the work piece, is oscillated at a frequency of about 20 kHz with an amplitude of about 25.4 μm.

A liquid metal ion source used in focused ion beam machining (FIBM) was demonstrated between 1978 and 1980. The latter technique, although very slow and expensive, represents the latest and most accurate means of mechanical precision machining.

FIGURE 6.4 An example of a desktop factory at AIST, Japan.

In this case, a powerful, very thin beam of ions replaces the familiar mechanical tool. A variation of FIBM is fast-atom beam (FAB) machining, pioneered by Hatamura et al.[11] in the early 1990s. In FAB, ionized atoms, accelerated to about 100 km/s, are neutralized and used as the mechanical cutting tool.

Desktop factories (DTFs) constitute a rather interesting new manufacturing philosophy involving flexible and modular table top-sized automated factories that feature minimal human participation in the manufacturing process. An example of such a factory is shown in Figure 6.4. Since the early 1990s progress has been made toward making such DTFs a reality. A DTF as shown here has the potential of becoming the factory of the future: a totally self-contained, robotic, desktop-size machine tool that only requires materials, power, and water as outside inputs, and out come the finished machined products. The first DTFs in research and development incorporated lathes, cleaning, gluing, punching, and drilling stations. The work piece is transported between these different machining functions by a "cart" moving from station to station. A first commercial manufacturing unit by Sankyo Seiki is shown as the inset in the header of this chapter. Sankyo Seiki Mfg. Co. Ltd. believes that DTFs might help manufacturing survive in Japan.[12]

Absolute and Relative Tolerances in Manufacturing

The application ranges of normal, precision, and ultraprecision mechanical machining, according to Norio Taniguchi, are shown in Figure 6.3. The Taniguchi curves, which dramatically illustrate improvement in mechanical manufacturing accuracies achieved in the second half of the twentieth century (left coordinate) are compared with Moore's Law (right coordinate), which demonstrates the logarithmic progress in density of transistors on a chip during the past few decades. Mechanical machining and integrated circuit (IC) technology-based machining methods, as well as appropriate metrology tools, are listed in the table on the right side of this figure (see also Evans[13]). As illustrated, ultraprecision diamond machining reached a manufacturing accuracy of better than 1 nm in the year 2000. In Chapter 2 we mentioned that the mirrors for an extreme ultraviolet camera are machined flat with an accuracy of 0.25 nm rms or better. At the same time, lithography methods only reached critical dimensions of 0.18 μm. However, this comparison is a bit misleading because the accuracy for mechanical machining, which is truly 3D, pertains to the product surface finish, whereas for lithography-based methods, it is based on the accuracy of carving out very small features (e.g., 0.18 μm) in the *x,y*-plane. Photolithography-based IC methods are limited to 2–2.5D shapes (projected shapes), and although the dimensions along the *z*-axis can also be controlled to the nanometer and subnanometer range, in the *x,y*-plane, the accuracy, which is mostly determined by the alignment accuracy of a large number of photomasks, is considerably less. In terms of attaining the smallest absolute sizes, mechanical machining is considerably inferior to lithography-based methods because the writing tool in mechanical machining is crude compared with writing with light or electrons—the smallest diamond styluses are still about 22 μm in diameter! In mechanical machining (for example, polishing of a mirror), the size of the grinding tool does not influence the resulting accuracy in the *z*-direction.

The discrepancy between the progress in tolerances predicted by the Taniguchi and Moore plots and the tolerances that are being achieved in practice have been increasing, i.e., progress is slowing down (see also Part I on lithography). Progress in mechanical machining is in large part based on continued improvement in positioning stages and

computer control. Because positioning stages for lithography and ultraprecision machining are the same, the fact that the progress is slowing for both is not that surprising. Reasons for the slowdown of IC technology are a bit more complicated, however. In lithography, subsequent layers need to be aligned, an operation not needed when diamond milling, e.g., a high-precision lens. The alignment operation in lithography may be checked by laser interferometer-controlled stages to an absolute positioning accuracy of <10 nm, thereby allowing the registration of successive layers to a similar precision. The reasons for the slowdown in electronic device miniaturization are obviously not because of lack of appropriate metrology equipment, but rather hinge on securing more mechanical stability in the positioning stages and even more so on manufacturing cost and the fact that transistors are becoming too leaky at less than 30 nm, where quantum effects start playing a role.

To further extend manufacturing accuracy in electronics, one has to look beyond current manufacturing technologies and consider both top-down and bottom-up manufacturing methods or perhaps hybrid approaches. Bottom-up methods, such as molecular engineering, and top-down methods, such as e-beam nanofabrication of quantum devices, are bound to lead the way beyond the current limitations dictated by the Moore and Taniguchi curves. Micromachining of special manufacturing tools will play a pivotal role for both those new bottom-up and top-down techniques. For example, the scanning tunneling microscope, based on micromachining ultrasharp tips on Si cantilevers, creates machining opportunities with accuracies on the atomic level, well beyond the predictions of the Taniguchi and Moore curves.

For mechanical structures, operation in the micro- and nanodomain imposes barriers other than those described for electronic devices. New design philosophies that are more error-insensitive with simpler mechanisms, smaller numbers of parts, and, wherever possible, flexible members rather than rigid ones must be adopted. Also here bottom-up machining methods will play an important role (see Volume III, Chapter 3).

In Figure 6.5, we show, in a most generalized fashion, human manufactured objects (from an IC fabrication to a house) in terms of absolute size and relative tolerances (feature tolerance/object size). Mechanical machining is typically used to machine the bigger objects, e.g., a lens of an absolute size of 10 cm or a parabolic astronomic mirror with a diameter of several meters.[3] Mechanical engineers define precision machining as machining in which the relative accuracy (tolerance/object size) is 10^{-4} or less

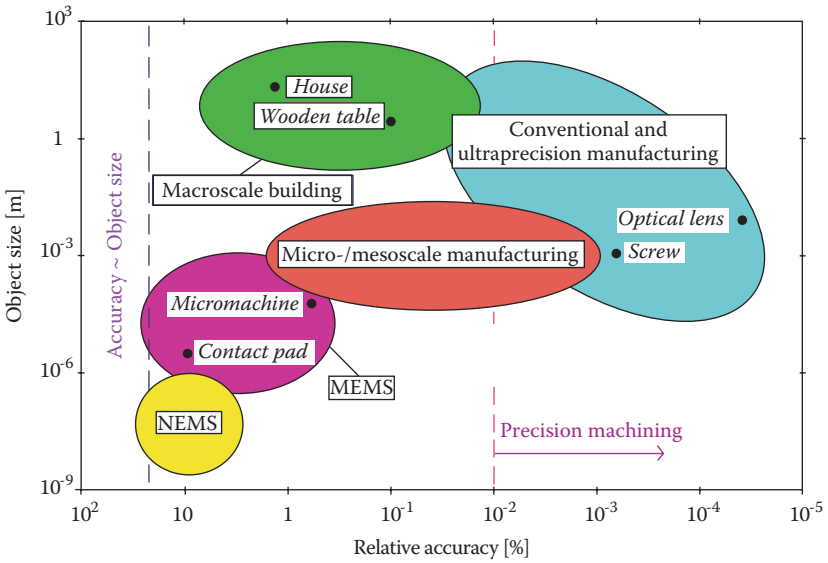


FIGURE 6.5 Various examples of human-engineered objects as a function of their relative (tolerance/object size) and absolute tolerance (size). The application field for precision machining at relative tolerances of 10^{-4} and less is also indicated. Object size versus relative accuracy for MEMS, NEMS, micro-/mesoscale manufacturing, and conventional/ultraprecision engineering.

of a feature/part size (see dividing line in Figure 6.5 at a relative accuracy of 10^{-2}%).[4,5] For comparison, a relative accuracy of 10^{-3} in the construction of a house is considered excellent. It is important to realize that, although IC techniques and silicon micro- and nanomachining (MEMS and NEMS) can achieve excellent absolute tolerances, relative tolerances here are rather poor compared with those achieved by most mechanical machining techniques. The decrease in manufacturing accuracy with decreasing size is rarely mentioned in discussions of Si micromachines; this probably is because Si micromachining originated from electrical engineering practice rather than mechanical engineering. The contrast is striking; in mechanical machining, relative tolerances of 10^{-6} (ppm) are becoming standard, whereas in the IC industry, a 10^{-2} relative tolerance is considered good. The definition of precision machining, with relative tolerances of 10^{-4}, excludes micromachining! To describe micromachining with the same terminology, it must be expanded to cover all machining methods where the relative accuracies are at least 10^{-4} or where absolute size in one or more dimensions is in the micrometer range.

In Figure 6.5 we also show the domains covered by the various machining options such as MEMS, NEMS, micro-/mesoscale manufacturing, and conventional/ultraprecision engineering. We see that IC, MEMS, and NEMS processes cover smaller absolute sizes, whereas their relative accuracies remain poor compared with those of conventional and ultraprecision manufacturing. IC-based machining in the relatively poor 10^{-1} to 10^{-2} relative accuracy range is also restricted in the types of engineering materials that can be used (Si, SiO_2, Si_3N_4, Al, …). There is an important micro-/mesoscale machining size range for applications with sizes between 10 µm and 1 mm, bridging the gap between silicon-based IC manufacturing and MEMS/NEMS and conventional miniature machining (Figure 6.6). This is a difficult size range to conquer: too large for IC manufacturing techniques and too small for traditional machining. It forms the needed hand-shake technology between the micro- and the macroworld. The bandwidth of sizes that can be attained with ultraprecision machining makes it an attractive candidate for this application range. Ultrahigh-precision techniques cover a wide range of materials, and they are—in contrast to IC methods—truly 3D and may even produce freeform surfaces. A freeform surface is a kind of complex, irregular, asymmetric aspherical surface. Generally, it cannot be described with an equation but usually only by a series of digital points. Unfortunately, whereas microscale technologies are well established in the microelectronics industry, the same cannot be said for mesoscale products involving complex 3D geometry and high relative accuracies in a range of nonsilicon materials. Thus, there is need for a strong focus on mesoscale manufacturing methods because our NEMS and MEMS devices would be rendered worthless if we cannot contact or package them.

"Precision machining," when used by mechanical engineers, has typically been reserved for removal processes only, whereas "micromachining," as used to describe IC, MEMS, and NEMS fabrication technology, covers both removal and forming processes. Precision machining and IC, MEMS, and NEMS are

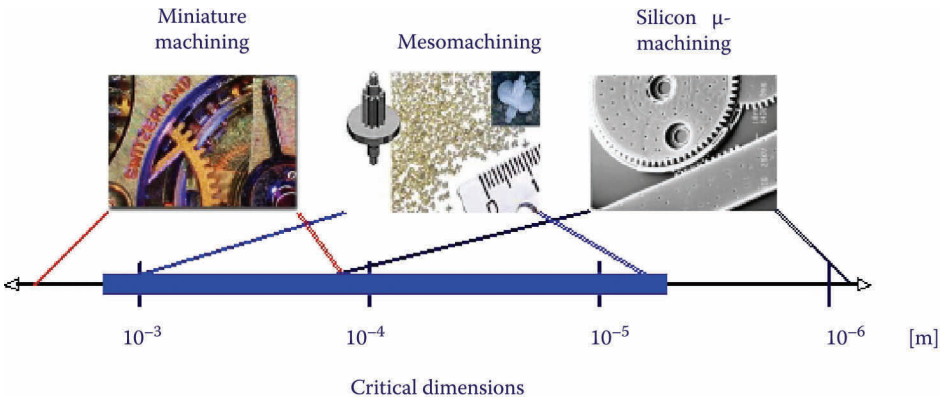


FIGURE 6.6 Mesoscale machining.

complementary; all strive to improve absolute and relative tolerances, with IC-, MEMS-, and NEMS-based technology better at obtaining smaller features (absolute tolerance) and traditional methods better at obtaining tighter relative tolerances. In this book, precision and ultraprecision machining encompasses removal and forming processes, with MEMS and NEMS as one of its newest disciplines.

The adjectives *nontraditional* and *nonconventional* have been in use by mechanical engineers for at least the past 30 years to describe nonmechanical machining methods such as thermal, electrochemical, and chemical machining of high strength and corrosion- and wear-resistant materials. However, the name *nontraditional* has become obsolete because "nonconventional" machining methods have found a wide range of commercial applications even with more readily machined materials.[14] In this book we refer to all top-down, non-IC manufacturing techniques as *traditional* machining techniques. In our nomenclature, IC-, MEMS-, and NEMS-based micromachining techniques are the nontraditional machining methods of our day. In making a manufacturing decision, it is not so much a question of which manufacturing technique is better or newer (very often, since the 1960s, silicon was the answer, sometimes independent of the question), but more which one is appropriate to use for the problem at hand.

Ultraprecision machining is significantly more expensive than normal machining and precision machining, especially when many abrasive machining steps are involved. Abrasive processes are finishing processes that are used to remove surplus material from the work piece surface. It is usually used on almost any surface that has been previously rough machined and is among the most expensive processes because it is generally slow in removing material. However, as we will learn, they can machine extremely hard surfaces and can be used on materials that are beyond the limits of virtually all other mechanical machining processes. In Table 6.1 available mechanical, electronic, and optical products are listed and categorized according to their tolerance band.

In general, to obtain the best possible accuracies with a given technique, materials such as wood or brick cannot be used; form-stable and workable materials such as aluminum, stainless steel, ceramics, or glasses must be used. In other words, material properties and machining accuracy are intertwined.

Ultraprecision Mechanical Machining

Introduction

In mechanical removal processes, stresses induced by a tool overcome the strength of the material. The process can produce complex 3D shapes, with very good dimensional control and good surface finishes. The method tends to be wasteful of material and expensive in terms of labor and capital. How well a part made from a given material holds its shape with time and stress is referred to as the *dimensional stability* of the part and the material. To maximize dimensional stability, the machine design engineer tries to minimize the ratios of applied and residual stress to yield strength of the material. A good rule of thumb is to keep the static stress less than 10–20% of yield. Increased heat at the work piece causes uneven dimensional changes in the part being machined, making it difficult to control its dimensional accuracy and tolerances. Thermal errors are often the dominant type of error in a precision machine, and thermal characteristics such as thermal expansion coefficient and thermal conductivity deserve special attention.[5]

Mechanical removal processes can be broken down into four commonly recognized categories: turning, milling, drilling, and grinding. *Turning* operations involve rotating a part and bringing it into contact with a fixed cutting tool. Mostly, turning makes parts that have a round cross-section such as screws, shafts, and pistons. *Mills* are used on a variety of materials, ranging from wood to metals, to shape a part. A mill's working feature is its revolving armature that can have a number of different tool-bit shapes. These shapes are used to cut out the different features of a part. A "block" of bulk material (e.g., steel) or a rough casting is secured to the mill, and either a manual operator or program (CNC) operates the machine, maintaining the cutting processes until the final milled product is produced. Machine *drilling* is normally done with a drill

Table 6.1 Tolerances of Components of Some Commonly Available Products

	Tolerance Band (±)	Mechanical	Electronic	Optical
Normal machining	200 μm	Normal domestic appliances and automotive fittings, etc.	General purpose electric parts, e.g., switches	Camera, telescope, and binocular bodies
	50 μm	General purpose mechanical parts for typewriters, engines, etc.	Transistors, diodes, magnetic heads for tape recorders	Camera shutters, lens holders for cameras and microscopes
Precision machining	5 μm	Mechanical watch parts, machine tool bearings, gears, ball screws, rotary compressor parts	Electrical relays, resistors, condensers, silicon wafers TV color masks	Lenses, prisms, optical fiber and connectors (multimode)
	0.5 μm	Ball and roller bearings, precision drawing wire, hydraulic servo-valves, aerostatic bearings, ink jet nozzles, aerodynamic gyro bearings	Magnetic scales, CCD (charged-coupled device), quartz oscillators, magnetic memory bubbles, magnetron, IC line width, thin film pressure transducers, thermal printer heads, thin film head discs	Precision lenses, optical scales, IC exposure masks (photo, x-ray), laser polygon mirrors, x-ray mirrors, monomode optical fiber and connectors
Ultraprecision machining	0.05 μm	Gauge blocks, diamond indenter tip radius, microtome cutter edge radius, ultraprecision *x-y* tables	IC memories, electronic video discs, LSI	Optical flats, precision Fresnel lenses, optical diffraction gratings, optical video discs
	0.005 μm		VLSI, superlattice thin films	Ultraprecision diffraction gratings

Source: McKeown, P.A. 1986. *Institute of Mechanical Engineers Proceedings* 200:76.[15]

press or a boring machine. Tools that are used to produce holes are called bits. They are held in place in the drill press by clamping them in a chuck or by inserting them into a tapered sleeve. *Grinding* is a finishing process that is used to remove surplus material from the work piece surface. It is usually used on almost any surface that has been previously rough machined and is among the most expensive processes because it is generally slow in removing material. These four processes are characterized by the fact that each process generates chips, and each of these process categories has unique characteristics such as available materials, acceptable tolerance ranges, surface finish, part, and tooling costs.*

A lathe is a machine for shaping a piece of material, such as wood or metal, by rotating it rapidly along its axis while pressing a fixed cutting or abrading tool against it. Lathes are among the more versatile of the machine shop tools. They have the capability of generating true round surfaces, flat surfaces, and also screw threads. The various operations of the lathe include turning, facing, boring, drilling, parting, and threading.

Given that we are mostly interested in micro- and nanomachines, we focus our attention here on ultraprecision single-point machining and abrasive machining for finishing a product. The former is a type of ultraprecision machining where submicrometer form accuracy and nanometer-level surface finish are achieved by using high-precision, high-stiffness machine tools and single-point cutting tools. The latter is a process that removes small amounts of material from a work piece in the form of tiny chips through the contact of small, hard, sharp,

* A useful glossary for terms used in metal removal and manufacturing processes in general can be found at http://instruct1.cit.cornell.edu/courses/orie310/mfgproc/mfgprocsummary.html, and an introduction to machine tools can be found at http://me.mit.edu/2.70/machine/outline.html and at http://engide.bizland.com/mfg-5-machine.htm.

nonmetallic particles. In the abrasive machining section we review abrasive wheel machining, in particular electrolytic in-process dressing (ELID), ultrasonic machining (USM), water jet, abrasive water jet (AWJ), and abrasive jet machining (AJM).

Single-Point Mechanical Machining

In mechanical single-point machining, various factors such as deformation of the work piece and tool, vibration, thermal deformation, inaccuracies of machine tools, and so on affect the machining accuracy. Tools for single-point machining include the following materials: carbon and medium alloy steels, high-speed steels, cast-cobalt alloys, carbides, coated tools, alumina-based ceramics, cubic boron nitride, silicon-nitride-base ceramics, diamond, and whisker-reinforced materials. Single-point machining falls into these four types of machining: 1) straight turning, 2) cutting off, 3) slab milling, and 4) end milling as illustrated in Figure 6.7.

The types of chip formation in these mechanical single-point machining processes are continuous, discontinuous, built-up edge, and oblique cutting. Most machining operations involve oblique cutting and lead to the familiar rolled-up chips. Built-up edges affect the performance of a cutting tool because of their hardness and the perceived dulling of the tool. Chip breakers are used to prevent discontinuous chips from forming (see Figure 6.8).

As in all metalworking processes where plastic deformation is involved, the energy dissipated in cutting is converted into heat that in turn increases

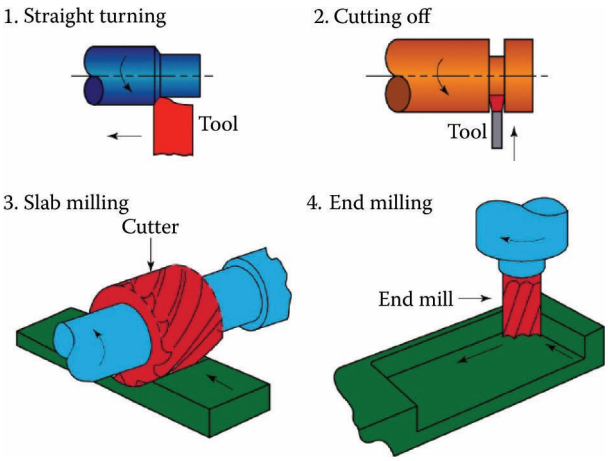


FIGURE 6.7 Single-point mechanical machining types.

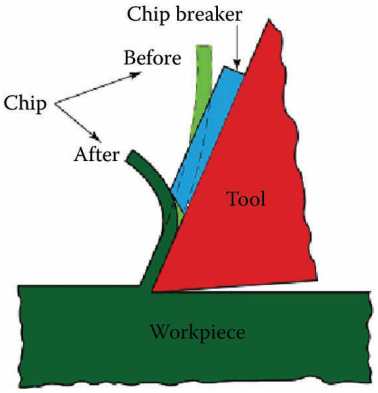


FIGURE 6.8 Chip formation during diamond milling.

the temperature in the cutting zone. Excessive heat decreases the strength, hardness, stiffness, and wear resistance of the cutting tool; tools also may soften and undergo plastic deformation, and thus the tool shape is altered.

Ultrahigh-precision machines with sharp single-crystal diamond tools have made submicrometer precision machining possible. Single-point diamond turning is less labor intensive than grinding and polishing (see "Abrasive Machining" below). A well-defined single-point tool is more predictable than a multipoint grinding wheel. Diamond-turned optical components have surfaces with a better metallurgical structure compared with the ones made from polishing and lapping. The time required to make optical quality mirrors and spheres using polishing and lapping methods is in the range of weeks, whereas the same or better quality parts can be completely fabricated from blanks using diamond turning in a few days.

A desktop high-precision four-axes CNC machine (four degrees of freedom)* for making microparts using milling and μ-EDM in a desktop machine was introduced in Chapter 5 (Figure 5.8). The single-crystal diamond cutter of this hybrid desktop machine was featured in Figure 5.11. In Figure 6.9a, we show a scanning electron microscope (SEM) photograph of a single-crystal diamond tool bit. The

* The degrees of freedom (DOF) of a machine refers to the number of independent axes along which one can obtain relative motion between the tool and the work piece. One can achieve this either by moving the tool or moving the work piece. The DOF does not include the spindle rotation or tool translation that is responsible for the cutting action. More than three DOF generally allows one to rotate the table or the work piece to achieve the correct angular relationship between the work piece and the cutter. This can help to reduce the number of separate fixturing steps and achieve smoother contours.

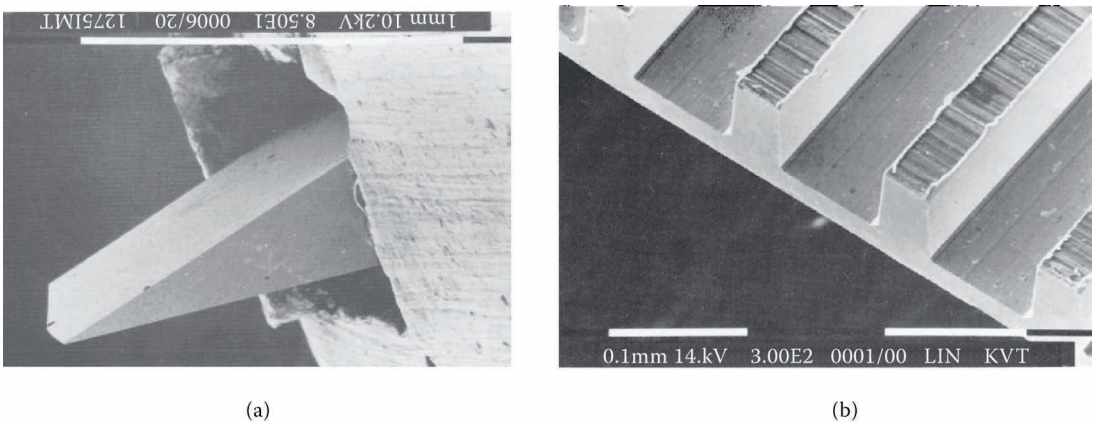


FIGURE 6.9 Diamond ultrahigh-precision tools. (a) Single-crystal diamond tool. Angle of tool is 20°. Maximum depth of cut is 500 mm. (b) Microgroove produced by diamond machining a 100-mm-thick Al foil. Width is 85 mm and depth is 70 mm.

radius of the edge on this bit is estimated at less than 0.05 μm; with such a tool, submicrometer grooves equivalent to those produced by silicon micromachining can be generated (Figure 6.9b).[16] The single hard point on the lathe in Figure 6.10 may be made from natural or synthetic diamond. Synthetic diamonds are preferred because natural diamonds have flaws that at times make them unpredictable. Ultraprecision cutting tools need to be hard and sharp, as well as have enhanced thermal properties to maintain their size and shape while cutting. High hot hardness means higher speeds and feed rates (higher production rates and lower costs). Advantages offered by diamond, the hardest material known today, include:

- Crystalline structure with a well-maintained sharp cutting edge
- High thermal conductivity, the highest of any material at room temperature
- Ability to retain high strength at high temperature
- High elastic and shear moduli, which reduce deformation during machining
- High wear resistance
- Low friction
- Produces a very accurate cut and good surface finish
- Most effective in light uninterrupted finishing cuts

However, diamond is chemically attacked by ferrous materials at high temperatures and is generally unsuitable for the machining of steels and nickel alloys. This is because of the very high wear rate of the diamond, resulting in nonviable tool costs for those applications. Diamond bits can be applied to most nonferrous metals, polymers, and crystals. Boron nitride in the forms of single-crystal cubic boron nitride (CBN) and microcrystalline cubic boron nitride (MCBN) under trade names such as Borazon or Borpax is used for precision cutting of hardened ferrous metals and high-temperature alloys.

Diamond machining is used for machining of nonferrous metals such as aluminum and copper, which are difficult materials on which to obtain a mirror surface by grinding, lapping, or polishing. This is because these metals are relatively soft, and the abrasive processes scratch the finished surface and, furthermore, are unable to produce

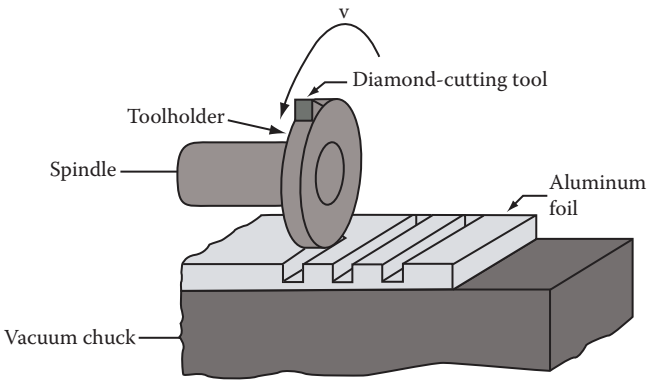


FIGURE 6.10 Diamond tool on a lathe for precision mechanical machining. Ultrahigh-precision milling of grooves in aluminum foil. The aluminum foil is 100 μm thick, and the grooves are 85 μm wide and 70 mm deep.

high levels of flatness at the edges of the machined surface.

The single-crystal diamond tool refinements are not the only reason for the high precision achieved today; submicrometer precision is also achieved through high-stiffness machine beds, air bearings (air bearings with a rotational precision of 0.01 µm and better are available), and measurement systems such as laser interferometry. Furthermore, highly precise instruments such as servomotors, feedback devices, and computers have been implemented, and many types of machine tools are now equipped for computer numerical control (CNC), further improving precision and reproducibility of the manufactured parts.

In mechanical micromachining processes, the effects of scaling on the process mechanism—including chip formation, cutting forces, vibration, process stability, and the surface integrity of the machined components—are some of the important issues that are not fully understood. The impact of work piece size on factors such as minimum chip thickness, surface quality, burr formation, built-up edge formation, and microstructure effects has been reported as summarized in Ehmann et al.[12] These reports reveal that ploughing and elastic recovery of the work piece plays a more significant role the more closely the chip thickness approaches the cutting edge radius of the cutting tool.

Various modeling techniques—including molecular dynamics (MD) simulation, finite element method (FEM), multiscale simulation modeling, and mechanistic modeling—have been attempted to characterize micromachining processes better. Simulation technologies for mahining processes such as FEM simulation of cutting, kinematical simulation of grinding, computational fluid dynamics (CFD) simulation of abrasive flow machining, and CFD simulation of electrical discharge machining (EDM) have been used for machining at the microscale level. These methods expand the process know-how and help to control development costs and define process limits. Using a diamond tip and numerical control (NC), rice grain-sized cars were machined at Nippondenso. One of these cars was 4.785 mm long, 1.73 mm wide, and 1.736 mm tall as shown in Figure 6.11.[17] In 1995, the Guinness Book of Records awarded a world record certificate to Nippondenso Co., Ltd. (now DENSO Corporation) for the development of this world's smallest motorized car. The DENSO Micro-Car is a miniature version of Toyota's first passenger car, the 1936 Model AA sedan. Its size is astounding: 1/1000th the size of the original car, or about the size of a grain of rice.

The Micro-Car has 24 parts that come in 13 different types, including body, tires, spare tire, wheels, axle, bearings, headlights, rear lights, front bumper, rear bumper, step, number plate, and emblem. The shell of the car was made by the sacrificial mold technique. A piece of aluminum was NC-machined and plated with 30-µm-thick electroless Ni; after cutting off the lower part of the body with EDM, the Al was removed by KOH etching. Finally, the Ni

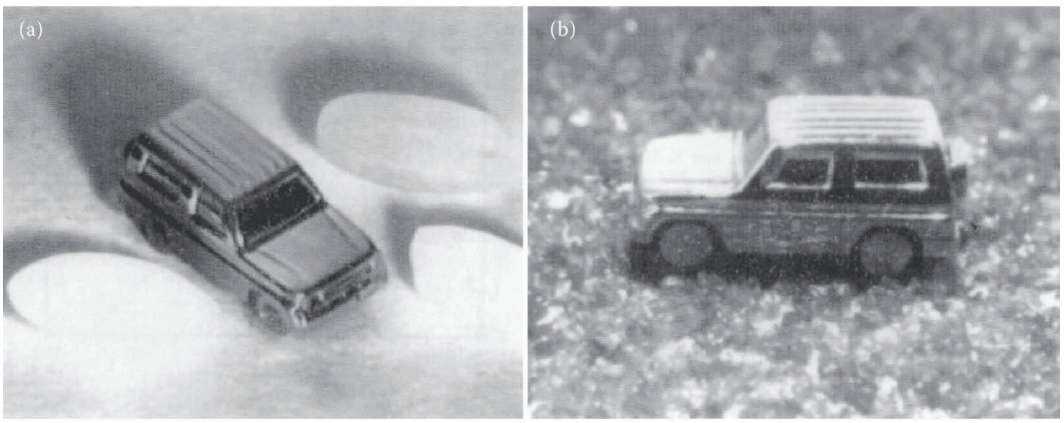


FIGURE 6.11 Micro-Car and rice grains (a) and Micro-Car on sandpaper (grain size = 200 mm) (b). The car is 7 mm long, 2.3 mm wide, and 3 mm high. (From Bier, W., G. Linder, D. Seidel, and K. Schubert. 1991. *KfK Nachrichten* 165;[16] and Teshigahara, A., et al. 1995. *J Microelectromech Syst* 4:76. With permission.[17])

car shell was Au coated to protect it from oxidation. The stainless chassis and wheels were also made by NC machining, and the core shaft and coil of the motor were made by NC machining and EDM. Wet etching of aluminum-coated glass was used for the registration plate and the emblem. A 250-µm bearing hole was drilled in the chassis for the zirconia wheel shaft. A stepping motor drives the car electromagnetically. The tube-shaped barium ferrite rotor, formed using grinding, was permanently magnetized with a special tool. The Micro-Car is powered via two 25-µm copper wires. Unfortunately, lubrication has drawbacks for microsystems, including high viscosity and adhesive effects. Therefore, the wear on the bearings is a limiting factor. Despite this, the car can run as fast as 100 mm/s, which is fast at this scale. The electromagnetic step motor is driven by an external magnetic field, and the car's permanent barium ferrite rotor runs at a maximum speed of 100 mm/s, developing a torque of about 10^{-6} Nm (at 3 V and 20 mA). To assemble the various microparts into a complete car, a cyanoacrylate adhesive and a mechanical micromanipulator, ordinarily used for handling biological cells, were used.[17]

FANUC's ROBOnano U_i is an ultraprecision multifunction micromachining station capable of making such high-precision parts as the mold for diffraction gratings or aspheric lenses (http://www.fanuc.co.jp/en/product/robonano/index.htm) (see the machine in heading of this chapter). This superprecision micromachining equipment consists of a diamond milling tool rotating at high speed on a superprecision positioning table with 1-nm step resolution and a rotation resolution of 0.00001°. It is a five-axis mill with all linear slides on air bearings. The machining operation is enclosed in an air shower with an environment controlled to within 0.01°C. The single-crystal diamond cutting tool is capable of generating 3D free curved surfaces. The machine is compact with a footprint of 1.27 m × 1.42 m and a height of 1.5 m and comes at a cost of $1 million. It is not available outside Japan at this time. Using this setup, FANUC succeeded in relief engraving a minute, mirror-surfaced Noh mask in copper with a "diamond end mill," which results in a surface smoother than a mirror. The roughness of the forehead (R_{max}) of the mask is 60 nm (see heading of this chapter). Thus, FANUC has realized a superprecision micromachining technique that allows mirror finishing in any direction except in a narrow groove. The machine was commercialized after 17 years of development by roughly 30 people. The ROBOnano is used to machine mirrors, filters, gratings, liquid crystal display panels, small lenses, and other small ultraprecise parts.

FIGURE 6.12 5-Axis-HP-Milling Center Gamma High Performance 303 by WISSNER GmbH, Göttingen, Germany.

A similar milling machine from Germany is the 5-Axis-HP-Milling Center Gamma High Performance 303 by WISSNER GmbH (Göttingen) shown in Figure 6.12. In Figure 6.13 a demonstration microstructure made on this machine, by the Fraunhofer Gesellschaft Institute for Production Systems (IPK) in Berlin, is shown.

A common limitation of the single-point diamond turning and grinding machines is their inherent ability to generate only rotationally symmetrical surfaces. Although these geometries cover the majority of requirements, there is an

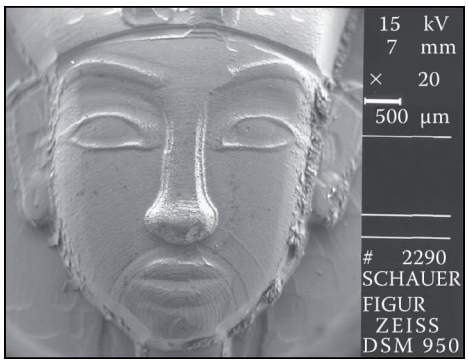


FIGURE 6.13 Demonstration on the 5-Axis-HP-Milling Center Gamma High Performance 303 by WISSNER GmbH (Göttingen, Germany) shown in Figure 6.12 by IPK (Berlin).

increasing demand for optics to incorporate more random, freeform geometries. Examples of these "freeform" geometries might be advanced laser printer optics or "conformal" windows that blend into the leading edge of an aircraft's wing. A new breed of multiaxis machines has been developed that are able to generate shapes that are no longer limited to rotational symmetry but extend to these freeform geometries. An example fabrication approach, to make, e.g., a freeform optical lens, is called computer-controlled fabrication (CCF). This may include 1) CNC grinding, 2) CNC lapping, 3) profile measuring, 4) computer-controlled corrective lapping, and 5) CNC polishing. After CNC grinding, a basic freeform surface with "peak-valley" structure is obtained. CNC lapping gets rid of the "peak-valley" and microcrack layer and decreases the surface roughness without lowering accuracy. A profile of the freeform lens is then measured with a computer-controlled coordinate measuring machine, and the machining error is obtained through a comparison between the measured data and theoretical profile data. A computer-controlled corrective lapping (CCCL) process then provides the corrective NC code by using the error data. Profile measuring and CCCL are alternated until the required accuracy is reached. Finally, the freeform lens is polished on a CNC polishing machine tool to lower the surface roughness.

Moore Nanotechnology Systems LLC, a US company, developed the Nanotech 220UPL (Figure 6.14) that is considered by many the most accurate diamond-turning machine in the world because of its groove-compensated air-bearing spindle that provides motion accuracy of less than 50 nm through its 1500-rpm speed range. Single-point diamond turning and deterministic microgrinding technologies are used for the production of plano, spherical, aspherical, conformal, and freeform optics as described above. The Nanotech 220UPL two-axis CNC single-point diamond turning lathe from Moore Nanotechnology Systems (Keene, NH) can be used to make many types of medical devices. Some applications include in vivo miniature pumps, injection-molded electroless nickel mold inserts for syringes, endoscopic lenses, or pipettes and implants.

FIGURE 6.14 Moore Nanotechnology Systems LLC, a US company, developed The Nanotech 220UPL that is considered by many the most accurate diamond-turning machine in the world because of its groove-compensated air-bearing spindle that provides motion accuracy of less than 50 nm of the whole 1500-rpm range.

Another trend in mechanical machining is toward flexible manufacturing. In response to the need for automation and the demands created by frequent design changes over a broad variety of products, the flexible manufacturing system (FMS) was developed. FMS is a combination of several technologies such as computers, CNC workstations, robots, transport bands, computer-aided design, and automatic storage. The technique was developed to produce many varieties of a certain product in smaller quantities rather than many devices of one type. CNC workstations are linked by automatic work piece transfer and handling, with flexible routing and automatic work piece loading and unloading.[18] This philosophy also underlies desktop factory design (see "Desktop Machining" below).

Ultraprecision manufacturing is the commercially preferred technique for the production of computer hard disks, mirrors for x-ray applications, photocopier drums, commercial optics such as polygon mirrors for laser-beam printers, consumer electronics such as mold inserts for the production of compact disc reader heads and camcorder viewfinders, and high-definition television (HDTV) projection lenses and VCR scanning heads.

A tutorial on mechanical drilling and milling can be found at http://www.me.mtu.edu/~microweb.

For more information on estimating manufacturing accuracies, consult http://www.isr.umd.edu/Labs/CIM/vm/ama/node19.html. Custom houses will typically report as best mechanical machining tolerances ±2 μm for cutting and ±0.4 μm for grinding (e.g., Burman tool at http://www.burman.co.uk/engineering/Machining.htm).

Abrasive Machining

Introduction

Abrasive machining is a process that removes small amounts of material from a work piece in the form of tiny chips through the contact of small, hard, sharp, nonmetallic abrasive particles. These abrasive particles may be attached to a spinning wheel (bonded), embedded in a slurry (loose), or impacted on a surface in a jet (impact). Abrasive methods in each of these categories are summarized in Figure 6.15.

Common types of abrasive machining processes using bonded abrasives include grinding, honing,* and microhoning. Lapping,† buffing, polishing, and ultrasonic machining (USM) use loose abrasive grit. Honing, microhoning, and lapping are usually termed *finishing operations* (also tertiary manufacturing methods). In USM, also called *ultrasonic impact grinding*, high-frequency vibrations delivered to a tool embedded in an abrasive slurry create accurate cavities of virtually any shape. Water jet, abrasive water jet (AWJ), and abrasive jet machining (AJM) are examples of abrasive impact machining.

The abrasive grit or cutter needs to be harder than the material to be cut (see Table 6.2). Natural abrasives include emery, corundum, quartz, garnet, and diamond. However, natural abrasives contain unknown amounts of impurities and possess

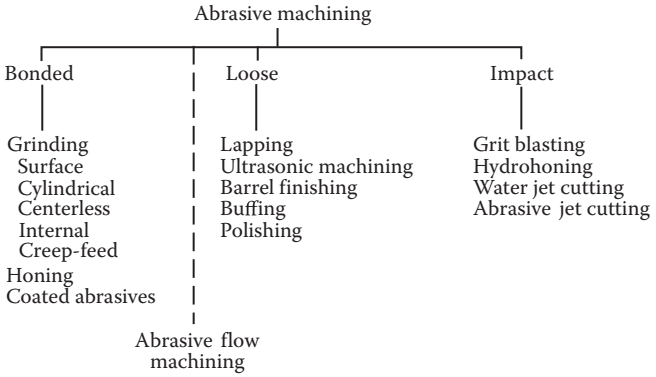


FIGURE 6.15 Categories of abrasive machining according to the matrix of the abrasive grit.

nonuniform properties, and as a result, aluminum oxide and silicon carbide abrasives are made synthetically. Conventional synthetic abrasives include aluminum oxide (Al_2O_3) and silicon carbide (SiC); superabrasives include cubic boron nitride (CBN) and diamond.

1. Aluminum oxide (Al_2O_3) known as Alundum or Aloxide. Various substances may be added to enhance hardness, toughness, and so on. Plain Al_2O_3 is white and used to grind steel, ferrous, and high-strength alloys.
2. Silicon carbide (SiC) known in trade as Carborundum and Crystalon. Harder than Al_2O_3 but not as tough. Used to grind aluminum, brass, stainless steel, cast irons, and certain brittle ceramics.
3. Boron nitride in the forms of single-crystal CBN and microcrystalline cubic boron nitride (MCBN) under trade names such as Borazon or Borpax. Used for hard materials such as hardened tool steels and aerospace alloys.
4. Diamond, a pure form of carbon, both natural and artificial. Used on hard materials such as ceramics, cemented carbides, and glass.

The temperature increase during abrasive machining can be very high (3000°F). Chips carry away heat, but a larger fraction of heat is conducted into the work piece. The effect of the temperature increase is more pronounced than in metal cutting, and the excessive temperature increase caused by grinding can temper or soften hardened metals. Abrasive machining is used when a very fine surface finish

* In honing, a simultaneous rotating and reciprocating motion between an abrasive stone and the work piece causes the removal of a small amount of material. The resulting work piece surface is one of high precision and uniform finish.

† The process of wearing down the surface of a softer material by rubbing it under pressure against the surface of a harder material that has been formed in the shape opposite to that desired on the softer material. Loose, wet abrasive generally is used between the two parts, but sometimes the abrasive is embedded in the lap in a rigid form (see diamond cutting tool).

Table 6.2 Knoop Hardness for Various Materials

Common glass	350–500	Titanium nitride	2000
Flint, quartz	800–1100	Titanium carbide	1800–3200
Zirconium oxide	1000	Silicon carbide	2100–3000
Hardened steels	700–1300	Boron carbide	2800
Tungsten carbide	1800–2400	Cubic boron nitride	4000–5000
Aluminum oxide	2000–3000	Diamond	7000–8000

and good dimensional accuracy are required and/or when the work piece materials are too hard or too brittle to work by other methods. The method is typically much slower than single-point mechanical machining (see above).

Grinding

Here, as a first example of an abrasive process and as an introduction to electrolytic in-process dressing (ELID) (see next section), we survey grinding. In grinding, the chip-removal process uses an abrasive grain or grit as the cutting tool that is bonded to a very fast spinning grinding wheel (see Figure 6.16). The grinding wheel is usually disk shaped and is precisely balanced for high rotational speeds. Grinding wheels incorporate an abrasive from the list in Table 6.2. The properties of a grinding wheel that determine how it acts are the kind and size of abrasive, how closely the grains are packed together, and amount of the bonding material.

The individual abrasive grains have an irregular geometry and are spaced randomly along the periphery of the wheel. The radial position of these cutting grains varies, and the cutting edges are extremely small. The grain size is an important parameter in determining the quality of the surface finish and the material removal rate. Small grit sizes produce better finishes, and larger grit sizes permit larger material removal rates. Also, harder materials need smaller grain sizes to cut effectively, whereas softer materials require larger grit size. The grain size of the abrasives is measured as grit size: coarse grain 8–24, medium 30–60, fine 70–180, and very fine >220. The feeds and depths of the cut in grinding are small, whereas the cutting speed is high (600 ft./min). As mentioned earlier, the temperature increase during abrasive machining can be very high (3000°F). Chips carry away heat, but a larger fraction of heat is conducted into the work piece. The effect of the temperature increase is more pronounced than in metal cutting, and the excessive temperature increase caused by grinding can temper or soften hardened metals. Burning chips are the sparks observed during grinding with no cutting fluid because the chips have heat energy to burn or melt in the atmosphere. Grinding may be classified as nonprecision or precision, according to purpose and procedure. In grinding, the chips are small but are formed by the same basic mechanism of compression and shear as in single-point machining (see above). There are three types of grain action in grinding:

1. Cutting: Grit can penetrate to the surface and performs chip removing.
2. Plowing: Grit can penetrate to the surface but cannot perform cutting. The work surface deforms.
3. Rubbing: Grits rub to the surface, and energy is consumed without cutting.

FIGURE 6.16 Common grinding wheels.

Dressing a grinding wheel is the process of resharpening the tiny cutting edges on a grinding wheel's surface. Thus, dressing a grinding wheel is shaping the grinding face of the wheel; this is often done with computer-controlled shaping features with a diamond-dressing tool kept normal to the surface at point of contact (Figure 6.17). The diamond dresses or turns the grinding wheel.

The machining energy in grinding is typically purely mechanical, but next we review a grinding

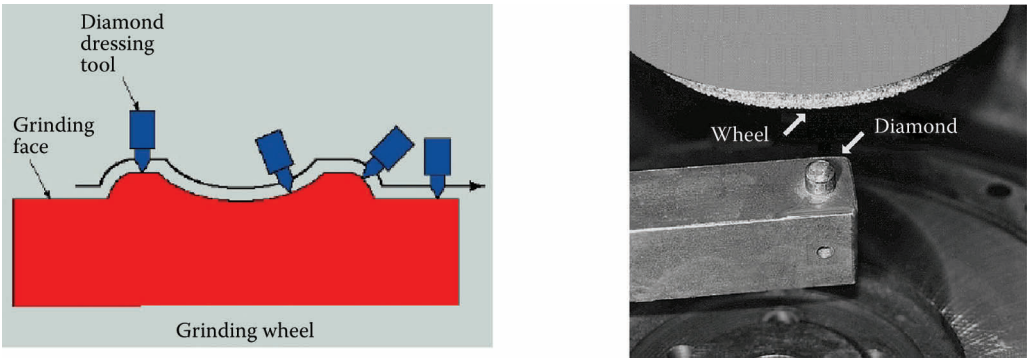


FIGURE 6.17 Shaping the grinding face of a wheel by dressing it with computer-controlled shaping features. Note that the diamond dressing tool is normal to the surface at point of contact.

process where electrochemistry plays a major role. In ELID, the material removal process (any material in this case) is purely mechanical, but an electrochemical dissolution process of the metal matrix dresses the abrasive wheel, with superabrasives embedded in a metal matrix, continuously.

Nanofinishing of Surfaces with Electrolytic In-Process Dressing

A nanosurface finish of a wide variety of brittle materials is becoming a more important demand in IC fabrication, MEMS, and NEMS. For example, a nanosurface finish is a necessity for printing nanowiring circuits using lithography techniques. The surface finish of Si wafers polished using chemomechanical polishing is found still less than ideal (see Chapter 1). To obtain a nanofinish, chip removal from the surface that is being grinded should be minimized. A *nanosurface* is only possible when the chip size of the material being removed reaches atomic level. Finishing with superabrasive grits [like diamond or cubic boron nitride (CBN)] is the method of choice to achieve these types of surfaces, but superabrasive grits need very good bonding strength to the matrix that holds them for better efficiency. Recent developments of ductile mode machining show that plastically deformed chip removal minimizes the subsurface damage of the work piece and that when chip deformation takes place in the ductile regime, a defect-free nanosurface is possible—completely eliminating the polishing process! Metal-bonded superabrasive grinding wheels in principle offer the solution here: they have the required higher bonding strength of the metal to the superabrasive grit and the required high mechanical stability. The major problems encountered using such superabrasive metal bonded wheels are that during grinding they get loaded and glazed. To avoid these difficulties, the wheel must be dressed periodically (see Figure 6.18), which

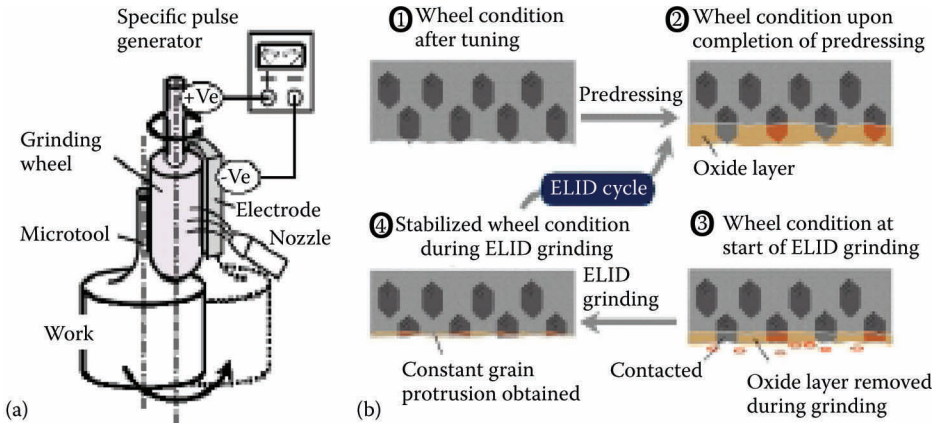


FIGURE 6.18 The ELID setup (a) and the mechanism of ELID grinding (b). (Ohmori, H., and T. Nakagawa. 1990. *Annals CIRP* 39:329.[19])

makes the grinding process very tedious. It is here that electrolytic in-process dressing (ELID) provides an elegant solution that eliminates both the wheel loading and glazing problems.

ELID is a mechanical grinding operation where the abrasive grit, embedded in a metal matrix, is kept fresh by an electrochemical dissolution process of the metal matrix. The method applies to metal and nonmetal substrates.

The concept of ELID was developed by Ohmori, chief scientist at RIKEN.[19] The system uses a metal bond grinding wheel (tool), a power supply, a grinding fluid (electrolyte), and a counter electrode as sketched in Figure 6.18a. It is important to notice that the work piece is not in the electrical circuit; thus, the grinding action here is purely mechanical, and the electrochemical assist is only there to keep the abrasive grit fresh. The metal bond wheel is made the positive electrode, whereas a negative electrode is positioned about 0.1–0.3 mm below the wheel face. The grinding fluid is supplied between the two. The grinding wheel (tool) used is the metal bond wheel in which bonding material is composed mainly of cast iron and cobalt and is sintered with the abrasive grit. The wheel is dressed using electrolysis during grinding, which helps to maintain larger grain protrusions and grit density. A DC pulse voltage is supplied between the two electrodes to electrolytically remove only the metal bond of the wheel, allowing efficient and automatic dressing of the wheel. This dressing is continued even during the grinding work to prevent reduced wheel sharpness from wear, thereby realizing highly efficient mirror-surface grinding. The mechanism of the ELID grinding method is sketched in Figure 6.18b. First the wheel bond is electrolyzed, which causes the abrasives to protrude appropriately (1). This process at the same time produces a nonconductive oxidized iron, which accumulates to form a coating on the wheel surface, automatically reducing the electrolysis current. Initial or predressing is completed at this point (2). When the grinding work is started in this state, the nonconductive oxide layer on the wheel surface comes in contact with the surface of the work piece and is removed by friction. As this takes places, the abrasives start to grind the work piece, and subsequently start to gradually wear (3). This reduces the insulation of the wheel surface, allowing the electrolytic current to flow again. As a result, the whole process starts again with the electrolysis of the wheel bond where the nonconductive oxide coating between the worn-out abrasives has become thin (4), allowing the abrasives to protrude again. This is called the ELID cycle. The ELID cycle varies according to the type of metal bond of the wheel used, electrolytic conditions, work piece, and grinding conditions. A surface finish of 2 nm may be achieved using ELID (20–50 N force for grinding).

ELID grinding is highly suitable for achieving nanosurface finishes on metals and nonmetals alike. Different types of ELID process are amenable for machining and finishing of all types of microshapes, and a nanosurface finish can be achieved using ductile mode machining on hard and brittle materials. ELID grinding can produce ductile surfaces without any subsurface damage, which eliminates the lapping and polishing processes. Thus, surface finish can be improved with higher accuracy and tolerance because the component is produced using only one process. The method is applied to process material such as silicon, glass, ceramics, extremely hard steel materials, and composite materials. Conventional applications include semiconductor substrates; electronic/optical components, such as lenses, magnetic heads, hard-disk substrates, and optical connectors; and the production of metal molds.

The major remaining difficulties with this technique are the lack of feedback devices to control the in-process dressing, optimization of machining conditions, and modeling of ductile mode grinding.

For up-to-date information on ELID, visit the web site of NEXSYS (http://www.nexsys.co.jp/index_e.htm) see Figure 6.19 for a NEXSYS desktop ELID machine. NEXSYS is an authorized RIKEN venture company established in June 1998 to provide services committed to spreading new fabrication technologies, mainly ELID mirror-surface grinding method developed at the Materials Fabrication Laboratory of RIKEN (The Institute of Physical and Chemical Research).

Ultrasonic Machining

Diamond tool machining is not only limited in the number of feature shapes and sizes that are possible it is also time consuming. Electrical discharge

FIGURE 6.19 Desktop ELID machine (500 mm × 500 mm × 560 mm) from NEXSYS (http://www.nexsys.co.jp/index_e.htm). ELID grinding is currently the best method to avoid high wheel loading and to generate mirror finish surfaces.

machining [(EDM), see Chapter 5)] is only suitable for use on conductive materials and because, just like laser beam machining [(LBM), see Chapter 5)], it is a thermal process, a heat-affected zone (HAZ) often has a negative impact on the end use of parts, especially in the case of high-reliability applications. In contrast, ultrasonic machining (USM) is a non-thermal, non-chemical and non-electrical machining process that leaves the chemical composition, material microstructure and physical properties of the work piece unchanged. This mechanical process can be used to generate a wide range of intricate features in both conductive and non-metallic materials with hardnesses of greater than 40 HRC (Rockwell Hardness measured in the C scale). The USM process can be used to fabricate precision micro-features, round and odd-shaped holes, blind cavities, and OD/ID features. Multiple features can be drilled in parallel, often reducing the total machining time significantly. Ultrasonic waves are sound waves of frequency higher than 20,000 Hz and can be generated using mechanical, electromagnetic, and thermal energy sources. They can be produced in gases (including air), liquids, and solids. One of the first applications for ultrasound was SONAR (an acronym for sound navigation ranging). It was used on a large scale by the US Navy during World War II to detect enemy submarines. Today's ultrasonic applications include medical imaging (scanning the unborn fetus), testing for cracks in airplane construction, and machining.

In ultrasonic machining (USM), also called *ultrasonic grinding*, high-frequency vibrations delivered to a tool tip, embedded in an abrasive slurry, by a booster or sonotrode, create accurate cavities of virtually any shape; that is, "negatives" of the tool. Almost any hard and brittle material, including aluminum oxides, silicon, silicon carbide, silicon nitride, glass, quartz, sapphire, ferrite, fiberoptics, and so on, can be ultrasonically machined. The tool does not exert any pressure on the work piece (drilling without drills) and is often made from a softer material than the work piece, e.g., from brass, cold-rolled steel, or stainless steel, and wears only slightly.[20] Tool material should be tough and ductile (low carbon steels and stainless steels give good performance). Tools are usually about 25 mm long, with a size equal to the "hole size" minus twice the size of abrasives. The mass of tool should be as small as possible so that it does not absorb the ultrasonic energy. The tool, typically vibrating at a low amplitude of 0.025 mm at a frequency of 20–100 kHz, is gradually fed into the work piece to form a cavity corresponding to the tool shape. The vibration transmits a high velocity to fine abrasive grains between the tool and the surface of the work piece. In the process, material is removed by microchipping or erosion with the abrasive particles. The grains are in a water slurry that also serves to remove debris from the cutting area. The high-frequency power supply for the magnetostrictive or piezoelectric transducer stack that drives the tool is typically rated between 0.1 and 40 kW. Piezoelectric transducers use the inverse piezoelectric effect using natural or synthetic single crystals (such as quartz) or ceramics (such as barium titanate), which have strong piezoelectric behavior. Ceramics have the advantage over crystals in that they are easier to shape by casting, pressing, and extruding.

Because this method is nonthermal, nonelectrical, and nonchemical, it produces virtually stress-free shapes even in hard and brittle work pieces. Ultrasonic drilling is most effective for hard and brittle materials; soft materials absorb too much sound energy and make the process less efficient. In Figure 6.20, important components of a typical USM setup are shown.[21] In the case of hard materials, microchipping is possible because of the high stress produced by particles striking the surface.

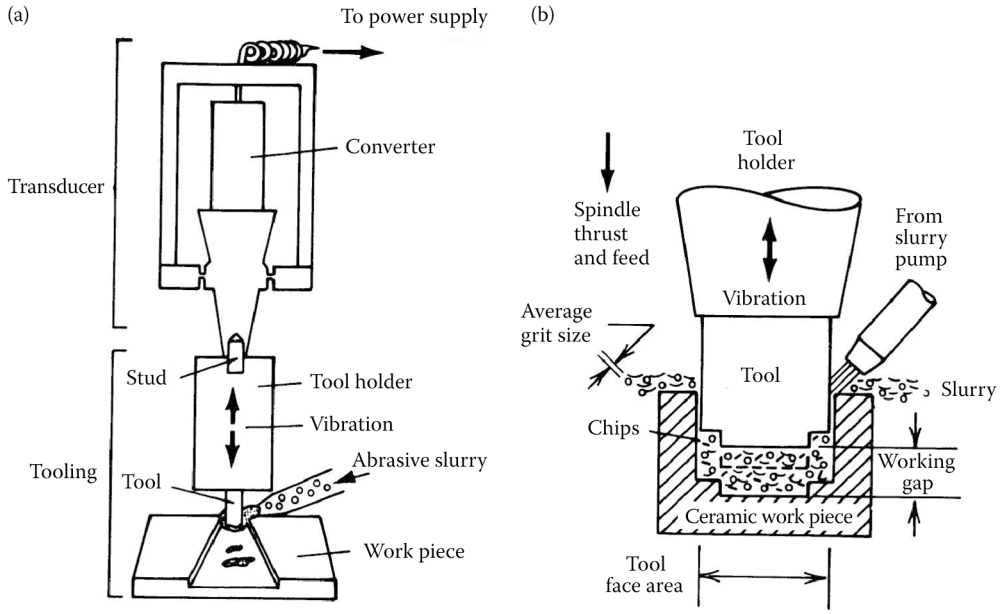


FIGURE 6.20 Schematic showing key components of a typical ultrasonic machining installation. (a) Transducer assembly coupled to tooling assembly. (b) Close-up view of tooling assembly used to machine ceramics. (From Moreland, M.A. 1992. *Engineered materials handbook*. Ed. S.J. Schneider. Metals Park, OH: ASM International. With permission.[21])

In Figure 6.21, SEM photomicrographs are shown of 640-µm holes drilled in alumina by two different techniques: USM (Figure 6.21a) and laser-beam machining (Figure 6.21b). The edges of the hole are much better defined with the ultrasonic technique. In Figure 6.22a, multiple channels and holes drilled all in parallel in a polycrystalline silicon slab are shown,[21] and in Figure 6.22b, a coin grooved with USM is displayed.

The key to USM, besides the tool itself, is the abrasive: a slurry of water or oil and small abrasive particles (large = 0.5 mm average particle diameter or small = 0.008 mm average particle diameter) of boron carbide, silicon carbide, or aluminum oxide. Boron carbide is the hardest abrasive and lasts the longest. The abrasives are suspended in water or oil at a concentration of 20–50 wt% and recirculated constantly to the work piece to supply fresh abrasive at the cutting site and remove abraded particles. The particle size and the amplitude of the vibrations used are typically made about the same. During one strike, the tool is moved down from its uppermost position with a starting speed of zero; then it speeds up to finally reach a maximum speed at the mean tool position. The tool slows down and eventually reaches zero again. When the grit size is close

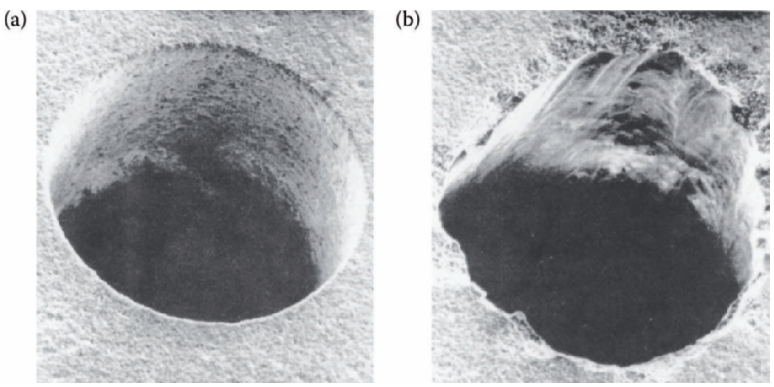


FIGURE 6.21 Scanning electron microscopy photomicrographs of 640-mm holes drilled into alumina with two different techniques. (a) Ultrasonic machining. (b) Laser-beam machining. (From Moreland, M.A. 1992. *Engineered materials handbook*. Ed. S.J. Schneider. Metals Park, OH: ASM International. With permission.[21])

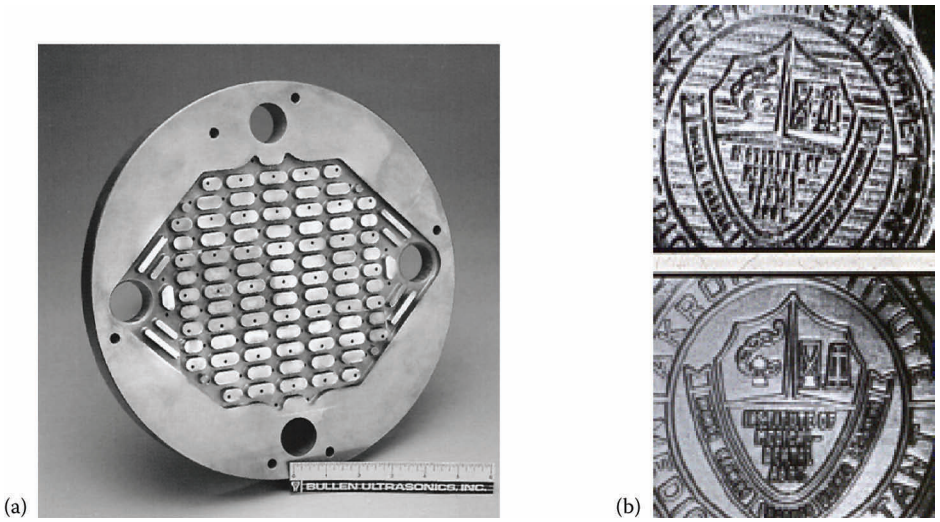


FIGURE 6.22 (a) Channels and holes ultrasonically machined in a polycrystalline silicon wafer. (From Moreland, M.A. 1992. *Engineered materials handbook*. Ed. S.J. Schneider. Metals Park, OH: ASM International. With permission.[21]) (b) Coin with grooving carried out with USM.

to the mean position, the tool hits the grit with its full speed. The smaller the grit size, the lesser the momentum it receives from the tool. Therefore, there is an effective speed zone for the tool, and correspondingly there is an effective size range for the grits. The size of the abrasive particles determines the roughness or surface finish and the speed of the cut. Material removal rates are low, usually less than 50 mm^3/min. The mechanical properties and fracture behavior of the work piece materials also play a large role in both roughness and cutting speed. For a given grit size of the abrasive, the resulting surface roughness depends on the ratio of the hardness (H) to the modulus of elasticity (E). As this ratio increases, the surface roughness increases. For example, under identical ultrasonic cutting conditions, carbide with an H/E ratio of 1.26 has a resulting surface roughness of 0.5 µm, and glass with an H/E of 8.1 ends up with a surface finish of 1.55 µm.[21]

Higher H/E ratios also lead to higher removal rates: 4 mm^3/min for carbide and 11 mm^3/min for glass. Thus, a low fracture toughness will result in a higher material removal rate and an increase in work piece roughness values. For a given material, a larger particle size results in a rougher surface. With a very fine particle size of 0.008 mm and a low H/E ratio (e.g., alumina with H/E 2.36), a 0.5-µm surface roughness can be achieved.[22] Holes as small as 0.076 mm have been drilled, but in production applications 0.25 mm is more the norm. The upper limit on cavity size is approximately 75 mm. With a 250-µm-diameter hole, the aspect ratio is about 2.5; with increasing hole size, the aspect ratio increases. It is safe to assume that a ±10-µm tolerance can be achieved with the finest abrasive powders.

Rotary ultrasonic machining (RUM) enhances material removal rates, finish capabilities, and overall drilling efficiency. A 1/4-in. hole can be drilled 5-in. deep into glass in about 2 minutes with a rotating ultrasonic tool. To make such high-aspect-ratio holes and deep cavities, it is often required that the work piece be removed periodically to ensure uninterrupted abrasive slurry flow to the cutting zone. Conventional USM is not capable of drilling microholes much smaller than 100 µm for lack of corresponding microtools. Using microelectrode discharge machining wire cutting (µ-EDM-WC; see Chapter 5) to make microtools (similar to the pipe structures shown in Figures 5.4 and 5.5), Sun et al. made holes as small as 15 µm in diameter.[23] Surface roughness was typically about 0.2 µm, and ultrasonic drilling speed ranged from 2.0–6.0 µm/min.

Summarizing, with USM, operating costs are reasonably low and required operator skill level is modest (operator must wear eye goggles to protect against abrasive particles and microchips). The method is slow (usually less than 50 mm^3/min), but under ideal conditions, penetration rates of 5 mm/min can be obtained. The required power is usually between 0.1 and 40 kW. Normal hole

TABLE 6.3 Advantages and Disadvantages of Ultrasonic Machining

Advantages	Disadvantages
Machining of any material regardless of conductivity	Low material removal rate
Precision machining of brittle hard materials	Tool wears fast
Does not produce electric, thermal, or chemical defects at the surface	Machining area and depth are restricted
Can drill circular or noncircular holes in very hard materials	
Less stress because of its nonthermal nature	

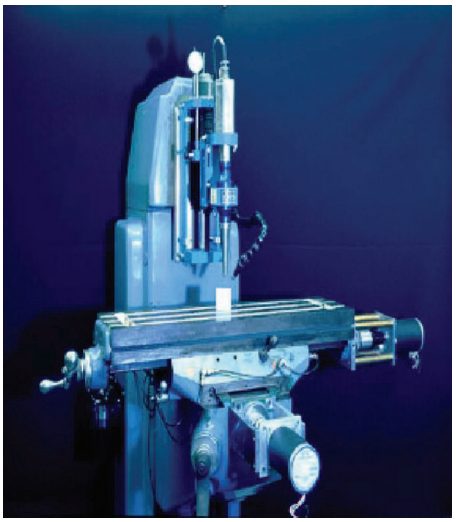


FIGURE 6.23 900-W Sonic Mill ultrasonic mill.

tolerances are 0.007 mm, and a surface finish of 0.02–0.7 µm can be achieved. Computer-controlled standard USM equipment is available with a position resolution on all axes of 0.25 µm (see Figure 6.23).

Machines cost up to $20,000, and production rates of about 2500 parts per machine per day are typical. If the machined part is a complex element (e.g., a fluidic element) of a size >1 cm^2 and the best material to be used is an inert, hard ceramic, this machining method might well be the most appropriate. A typical ultrasonic service provider is Bullen Ultrasonics (Eaton, OH; http://www.bullen-ultrasonics.com).

Advantages and disadvantages of USM are summarized in Table 6.3.

Water Jet Machining

In water jet machining, in industrial use since 1970, material is removed through mechanical impact; that is, erosion effects of a high-velocity (600–900 m/s with pressures up to 400 MPa), small-diameter (in the range of 0.07–0.5 mm) jet of water.[24] It is a nonthermal process that eliminates warping and avoids a heat-affected zone (HAZ). Water is most commonly used, but additives such as alcohols, oils, and glycerol are used when they can be dissolved in water. The exit orifice is often made of sapphire and ranges from 0.05–0.020 in. There are some advantages one can list for water jet machining:

- No heat is produced.
- Cut can be started anywhere without the need for predrilled holes.
- Burr produced is minimum.
- Environmentally safe and friendly manufacturing.

The technique is mainly used to cut soft, nonmetallic materials such as Plexiglas®, cardboard, fabrics, foam rubber, wood, food, and so on. With a water jet nozzle of 0.13 mm and a pressure of 379 MPa, a 3-mm-thick, Plexiglas plate can be cut at speeds of 0.9 m/min. Tolerances are a function of the material type and thickness, and they are usually within ± 0.1–0.2 mm.[25] Edges in water jet machining have a sandblasted appearance. Harder materials exhibit a better edge finish. Finishes range from 63 rms* to very coarse depending on the material. In Figure 6.24, we show a schematic illustration of water jet machining and examples of various nonmetallic parts cut by a water jet machine. When dealing with the choice of using laser machining (a thermal technique covered in Chapter 5) or water jet machining, one may use the following decision tree:

* Rms refers to the root mean square which is an average of peaks and valleys of a materials' surface profile (see Equation 6.5). It is given in micrometers or microinches. Rms is also called R_q.

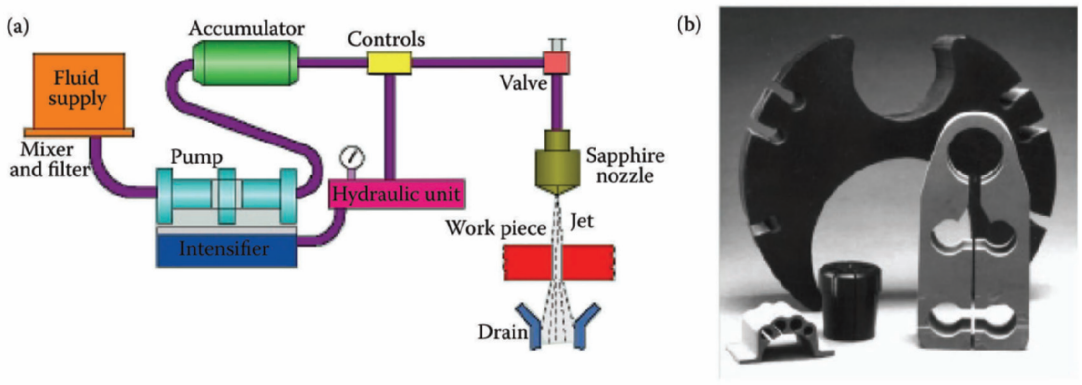


FIGURE 6.24 (a) Schematic illustration of water jet machining. (b) Examples of various nonmetallic parts cut by a water jet machine. (Courtesy of Possis Corporation.)

1. Is the material reflective or heat conductive? Yes →
   No ↓
2. Is material thicker than 3/8"? Yes →
   No ↓
3. Is a heat affected zone (HAZ) a concern Yes →
   No ↓

LASER WATERJET

In Table 6.4 we compare advantages and disadvantages of water jet and laser machining. In water jet-guided laser machining (Chapter 5, Figure 5.23) the two techniques are combined. A thin jet of high-pressure water (between 40 and 100 μm) and water at a pressure between 20 and 500 bars is forced through a nozzle (made of diamond or sapphire). The water jet is not abrasive in this case; the material is solely ablated thermally by the laser. The high speed of the water, up to 300 m/s at 500 bars, ensures fast removal

TABLE 6.4 Advantages and Disadvantages of Water Jet and Laser Machining

Waterjet		Laser	
Advantages	**Disadvantages**	**Advantages**	**Disadvantages**
Water jet diameter 0.001–0.014 in.		Spot size as small as 0.0015 in.	
Higher material thicknesses can be cut	Equipment cost	Narrow cut (0.020 in.)	Equipment cost
No heat-affected zone	Pump maintenance (every 1000 h)	Low maintenance	Requires material that absorbs laser light; material limitations (e.g., laser reflective materials such as Cu and Al)
Does not produce electric, thermal, or chemical defects at the surface	Noise (80 db or more)	More compact nesting	Heat-affected zone (HAZ)
No part distortion	Need for filters	Faster cutting rates	
Not dependent on material property; affected only by resistance to abrasion	Slow cutting rates	Holes as small as 0.003–0.015 in.	
	User education in high-pressure plumbing	Tolerance of ±0.0003 in. through thin material	
	Finishes will range from 63 rms to very coarse	Edge finish 125 rms	
	Water must be highly purified		

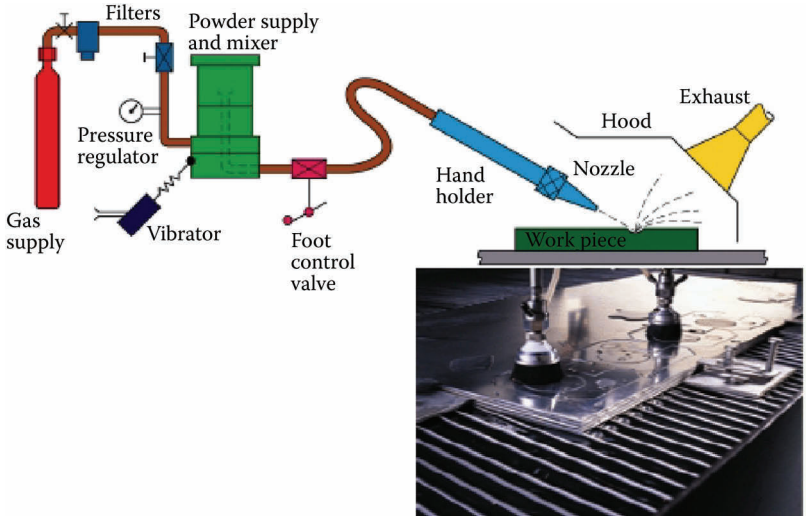


FIGURE 6.25 Abrasive water jet (AWJ) machining setup.

of debris. Water jet-guided laser etching prevents the problem of heat damage (microcracks, structural changes) inherent in any laser ablation process. The result is a heat damage-free material ablation.

Abrasive Water Jet Machining

By adding abrasives to the water jet, a much wider range of materials, including many more metals and ceramics, can be machined. Abrasive water jet (AWJ) has been in industrial use since 1982. The cutting action is a grinding process where the forces and motions are provided by water rather than a solid grinding wheel. The method is principally used for edge finishing, deburring, and polishing and has been increasingly used for treating hard-to-machine and multilayered materials and as an alternative tool for milling, turning, drilling, and polishing. In some applications AWJ is starting to replace electrical discharge machining (EDM).

An abrasive jet uses water that is pressurized up to 400 MPa and then forced through a small sapphire orifice at 1000–2500 ft./s, or up to about 2.5 times the speed of sound. Some abrasive is then pulled into this high-speed stream of water and mixed with the water in a ceramic mixer as shown in Figure 6.25. This jet of water and abrasive is then directed at the material to be machined. The jet drags the abrasive through the material in a curved path, and the resulting centrifugal forces on the particles press them against the work piece. AWJs can machine a wide range of thicknesses and materials, including metals, plastics, glass, and ceramics. Materials cut by the abrasive jet have a smooth, satin-like finish, similar to a fine sandblasted finish. Abrasive jets abrade material at room temperature. As a result, there are no heat-affected zone (HAZ) or structural changes in the work piece. Abrasive jets can also machine hardened metals and materials with low melting points. Abrasive jets typically use garnet as an abrasive. Garnet is a reddish natural crystal, with a Mohs hardness* of 6.5–7.5. No noxious gases or liquids are used in AWJ, nor are there any oils used in the machining process, so this is again a "green" machining approach. A wide range of conventional processes can be performed with this single tool such as drilling, sawing, broaching, profile milling, and so on. No heavy burrs are produced during the AWJ process. Parts can often be used directly without further deburring.

Common applications of AWJ machining include:

- Fast and precise cutting of fabrics
- Vinyl, foam coverings of car dashboard panels
- Plastic and composite body panels used in the interior of cars
- Cutting glass and ceramic tiles

Principles of Abrasive Water Jet Machining provides good further study.[26] For a tutorial on AWJ and

* Mohs hardness is defined by how well a substance will resist scratching by another substance. The scale consists of 10 minerals ranging from 1 for the softest (talc) to 10 for the hardest (diamond).

AJM on the Internet, visit http://www.waterjets.org/about_abrasivejets.html. Common water jet terminology is explained at http://www.omax.com/glossary.php. For a typical equipment supplier, visit Omax at http://www.omax.com.

Abrasive Jet Machining—Powder Blasting

Sandblasting in a pressurized air stream is used to remove paint, clean cathedrals, and decorate glass. On a finer level, powder blasting, using small particles (usually Al_2O_3 powder, particle size <100 μm) is used for device demarking in the electronics industry and in surface preparation before plating. On this level, sandblasting is called *abrasive gas jet machining* (AJM) or *microabrasive blasting*. In this type of machining, the work piece material (hard, brittle materials such as glass, silicon, tungsten, and ceramics) is removed by the mechanical impact of a high-velocity ($v_{particle}$ = 80–300 m/s) stream of abrasive-laden inert gas or air. The impact causes crack generation on the target surface.[27,28] Using masks, the method is also put to use in rapid prototyping[29] and etching of shallow intricate holes in electronic components such as resistor paths in insulators and patterns in semiconductors. The process is fast and directional, and removal rates of up to 25 μm/min have been achieved. Another application is the creation of thousands of holes at once at low cost, high speed, and high accuracy for flat panel displays.[14,28]

Recent efforts are in fine-tuning the technique for mesoscale MEMS applications. The required equipment, as schematically illustrated in Figure 6.26, is relatively inexpensive.

The setup consists of the powder blaster that regulates the flow of abrasive powder accelerated to the nozzle. An air dryer is connected to the powder reservoir to avoid moisture-induced particle agglomeration. To obtain an homogeneous scan of the work piece surface, the sample is placed on an *x,y*-translation stage. A dust collector recovers the used alumina powder. The applied pressure can typically be regulated between 1 and 5 bars.

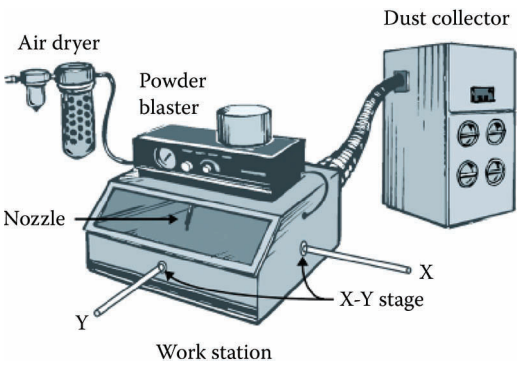


FIGURE 6.26 Schematic of powder blasting machine. (From Yamahata, C. 2005. EPF de Lausanne.[30])

A schematic diagram showing the principle of powder blasting microerosion process is presented in Figure 6.27. The different elements shown are elucidated next. A nozzle (e.g., from tungsten) provides the focusing and acceleration function for the abrasive stream generated by the blaster. A protective mask with an erosion rate much smaller than that of the substrate is also shown. In combination with suitable masks, for example, an electroplated photolithographically patterned Cu layer, minimum feature sizes of less than 50 μm, an aspect ratio of 2.5, and a depth of >1 mm in glass have been demonstrated. Polymer negative resist foils (e.g., ORDYL BF 400, specially developed for powder blasting) can also be used as mask, but in that case feature sizes are limited to 75 μm and an aspect ratio of 1.5 (http://www.el.utwente.nl/tdm/mmd/education/ d-opdrachten/powderblasting.html). Wensink et al. made a comparison study of various mask materials in powder blasting, and the same group is striving to decrease the minimum feature size to less than 10 μm (http://www.el.utwente.nl/tdm/mmd/education/d-opdrachten/powderblasting.html). The substrate in this type of machining may be glass, silicon, tungsten, or a ceramic. Finally, there is the abrasive powder.

Depending on the application aimed for, different types of abrasive powders with various shapes and hardness can be used. The most common powders are:

- Silicon carbide (SiC), because of its hardness, is the most efficient medium used for microabrasive blasting. It has a hardness of more than 9 on the Mohs scale, just below diamond.
- Crushed glass is a mild abrasive medium (Mohs hardness = 5–6), which is used when only a slight degree of abrasion is desired.
- The spherical shape of glass beads keeps them from cutting into the surface of the part. It is

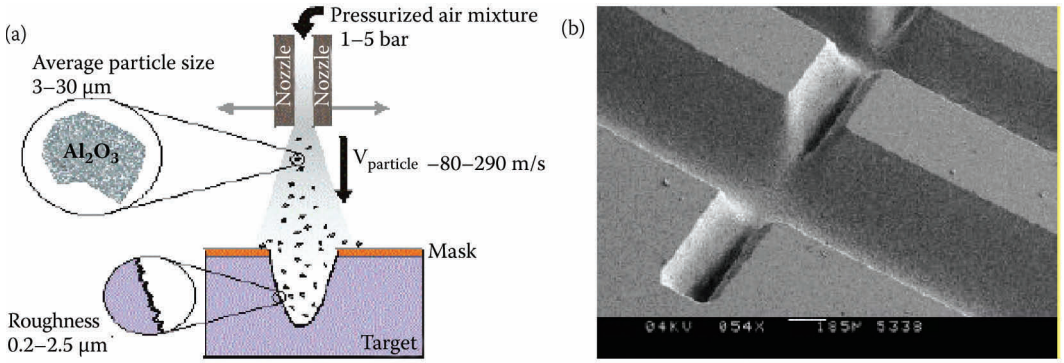


FIGURE 6.27 (a) Schematic diagram demonstrating the principle of the powder blasting microerosion process. (b) 400-μm deep-blasted glass structure. (From Yamahata, C. 2005. EPF de Lausanne[30] and http://www.anteryon.com.)

typically used to perform light deburring or to apply a satin-like finish on a work piece.
- Aluminum oxide is the most commonly used cutting abrasive because of the sharp shape of the particles and good hardness (Mohs hardness = 9).

The rate of material removal in powder blasting is commonly evaluated from the erosion rate E_{rate}, which is defined as:

$$E_{rate} = \frac{\text{weight of removed material}}{\text{weight of impacting particles}} \qquad (6.1)$$

Slikkerveer[31] studied the mechanics of powder blasting in detail. In his analysis, he introduced the erosion efficiency, E_{eff}, as:

$$E_{eff} = \frac{\text{weight of removed material}}{E_k} \qquad (6.2)$$

where E_k is the kinetic energy of the incoming particles. Combining Equations 6.1 and 6.2, one obtains:

$$E_{rate} = \frac{1}{2} E_{eff} v_{particle}^2 \qquad (6.3)$$

Notice that the erosion efficiency is a function of the kinetic energy only, whereas the erosion rate also depends on the particle velocity. E_{eff} was found to be a more useful parameter for comparison of powder blasting processes because the erosion caused by an individual particle impact directly depends on the kinetic energy of the particle. Slikkerveer et al.[32] in particular studied the erosion of brittle materials by hard angular particles impact and proposed a model that was derived from the indentation fracture theory. From their model, they concluded that the main material characteristics influencing the erosion efficiency are:

- Density (ρ) of the particles, which figures in the evaluation of the kinetic energy E_k
- Young's modulus of elasticity (E) of the substrate
- Fracture toughness (K) of the substrate, which is a measure of the ability of a material to resist the growth of an existing crack or flaw
- Indentation hardness (H) of the substrate, which characterizes a material's local resistance to a permanent deformation (e.g., scratching, abrasion, or cutting)

Powder blasting has been shown to be an appropriate micromachining technique for hard, brittle materials. It is a fast, inexpensive, and accurate directional etch technique for brittle materials such as glass, silicon, and ceramics. The time to etch

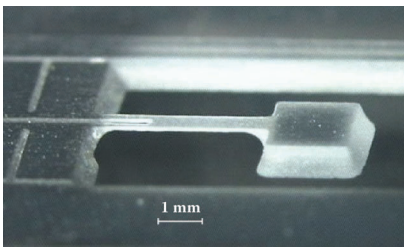


FIGURE 6.28 Powder-blasted accelerometer in glass. A first exposure of the metalized glass substrate to the eroding powder beam defines the freestanding structure and electrical contact geometry. A second exposure from the back is used to thin the cantilever beam. (Belloy, E., A. Sayah, and M. A. M. Gijs. 2000. *Sensors Actuators A* 86:231.[36] Courtesy of Dr. M. Gijs, EPFL.)

through a 500-mm-thick Pyrex wafer with one nozzle is approximately 20 minutes. The erosion process is caused by the impact of accelerated particles on brittle materials, a process that creates cracks. Powder blasting is mostly used with a powder beam at normal incidence angle to the substrate ($\theta = 90°$). Belloy et al.[33] described how, by exploiting oblique powder blasting ($\theta = 80°–40°$) and underetching phenomena of the mask on a glass substrate, a new family of 3D structures can be realized. Using this technique, a glass cantilever accelerometer beam with inertial mass was structured (Figure 6.28). Using the same method, Solignac et al.[34] fabricated capillaries for electrokinetic experiments in soda lime microscope glass slides, and Bornand et al.[35] fabricated magnetic cores from ferrite wafers for millimeter-size transformers for low power applications.

Surface Finish

The desired work piece surface determines the type of finishing operation. To rate the final surface finish of a work piece, one uses one of two criteria: the R_a or the R_q (formerly rms) value. Both are generally given in micrometers or microinches. R_a is known as the arithmetic mean value: with reference to Figure 6.29,

FIGURE 6.29 Surface roughness and center line.

it takes the average of the absolute values of the deviations for the centerline of the surface:

$$R_a = \left(\frac{|a| + |b| + |c| + \cdots}{n} \right) \tag{6.4}$$

The R_a in an older convention is also called the arithmetic average. The R_q value, which was formerly called the rms, calculates the root mean square of the values of the deviations from the centerline:

$$R_q = \sqrt{\frac{a^2 + b^2 + c^2 + \cdots}{n}} \tag{6.5}$$

Other measurements of surface roughness exist, but these two are the most common. In particular, rms or R_q is the most prevalent. In general, R_q values are 1.1–1.4 times the R_a values, but, of course, it depends on the exact surface topology.

In Figure 6.30 we represent a chart of the typical R_a readings associated with different machining

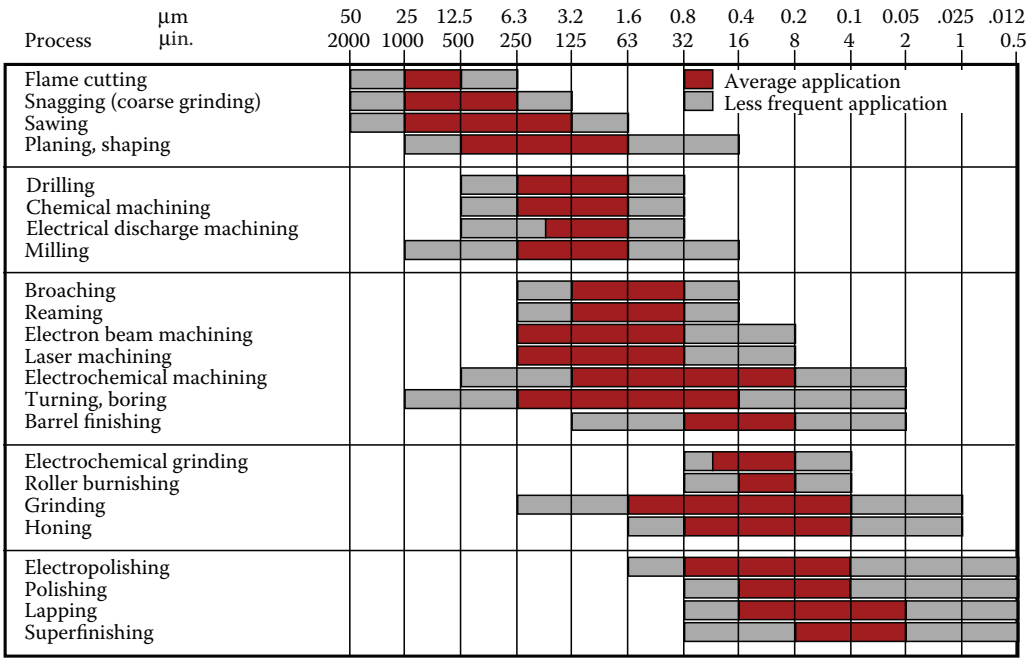


FIGURE 6.30 Typical arithmetic average roughness, R_a, for various machining methods. (Kalpakjian, S. 1984. *Manufacturing processes for engineering materials.* Reading, MA: Addison-Wesley.[37])

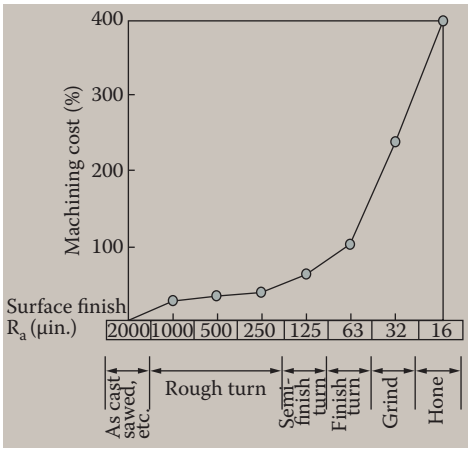


FIGURE 6.31 Increase in the cost of machining and finishing a part as a function of the surface finish required.

processes. Notice the bottom few processes—lapping and superfinishing—that are typical abrasive machining processes. Electropolishing comes close in surface finish to these abrasive methods but is generally not as good a finish as the abrasives. Also notice the very small numbers with which we are dealing. Remember a human hair is about 80 μm. In some cases we are machining to the hundredth of a micrometer! The ELID process, surveyed above, is even better. In Figure 6.31 we illustrate the increase in the cost of a machining process as a function of the surface finish required.

Focused Ion Beam Machining and Fast Atom Beam Machining

Focused Ion Beam Machining Some authors classify focused ion beam (FIB) milling as a thermomechanical technique. It is more widely accepted now as a pure mechanical machining technique in which the drill bit is replaced with a stream of energetic ions—call it ion blasting. A liquid metal ion source (e.g., gallium) is used rather than the wide argon ion beam in dry etching (see Chapter 3), and the ion beam is focused down to a submicrometer diameter. The liquid metal source uses a capillary feed (or needle) system and an acceleration voltage of 5–50 keV. Beam spots smaller than 50 nm are possible. FIB has been used for maskless implantation, metal patterning, and IC repair (see Figure 6.32).

More recently, it has been shown that FIB also constitutes a production technique of micrometer-sized objects in an arbitrary material without the requirements of anisotropy or lithography. Vasile et al., for example, described a 360° rotation mounting device used analogously to a lathe with a 20-keV Ga ion beam as the cutting tool for in-vacuum micromachining.[38,39] The ion beam they used is 0.3 μm in diameter, and they crafted tungsten needles, hooks, tuning forks, and specialized scanning probe tips. Some of these tungsten microstructures are shown in Figure 6.33. The first microstructure (Figure 6.33a and b) is a parasol-shaped scanning probe, which can be used to probe underneath resist lips and other concave features. The disc structure sits on a 1.5-μm square shank and is 4 μm in diameter and about 0.7 μm thick. The shank segment just before the disc is deliberately thinned to 0.4 μm at the thinnest point. The microshovel shown in Figure 6.33c is another illustration of the truly 3D aspect of this new manufacturing technique. The resolution limit to the fabrication process of these various tungsten structures mainly comes from the grain structure of the material. Start-to-finish fabrication time for the objects shown was ~2 hours. It should be noted that this expensive serial technique

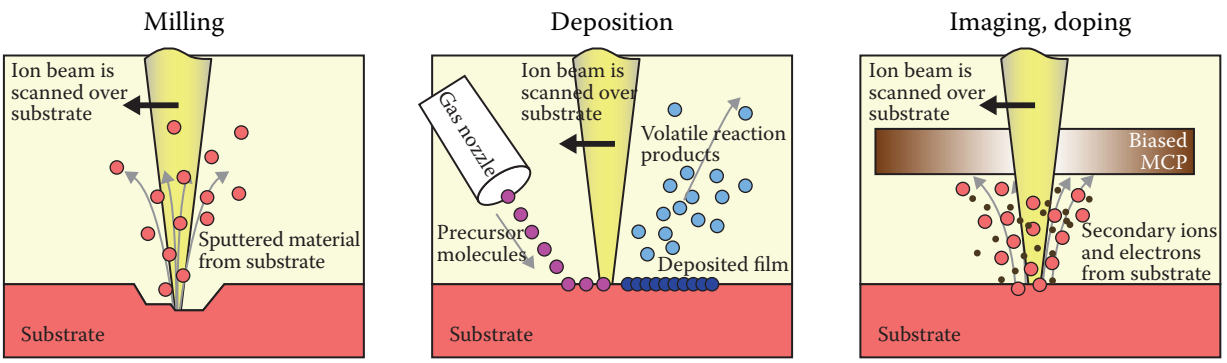


FIGURE 6.32 Uses of focused ion beam (FIB).

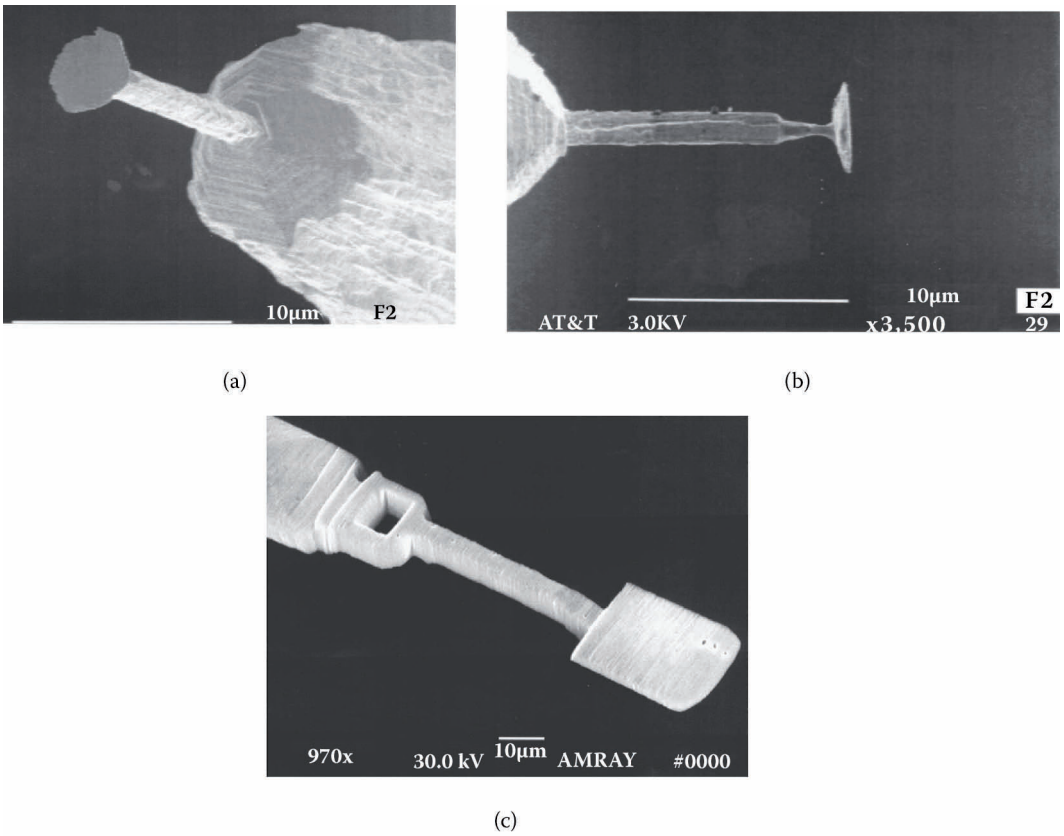


FIGURE 6.33 FIB-manufactured tungsten structures. (a) Specialized scanning tip probe parasol. (b) Side view of a parasol structure. The shank is 1.5 mm^2 and thinned down locally to 0.4 mm. The disk is 4 mm in diameter. (c) Microshovel. (From Vasile, M. J., C. Biddick, and S. A. Schwalm. 1994. *J Vac Sci Technol* B12:2388. With permission.[39] Courtesy of Dr. M. J. Vasile, Louisiana Tech University.)

might make sense for the fabrication of an atomic force microscope tool set, but to make a microshovel as shown in Figure 6.33c is of course merely meant as a fancy demonstration of the technique.

In ion-beam sputtering, a broad surface area is bombarded with energetic ions from inert gases such as argon, and momentum transfer kicks off atoms from the target. The technique was discussed in detail in Chapter 3. In reactive ion-beam etching, also reviewed in Chapter 3, a flux of reactive particles, such as CF_4^+ and F^+, is directed at the work piece in a narrow beam. In the latter process, both momentum transfer and chemical reactions are involved in the etching process, and the process cannot be solely characterized as mechanical. In reactive ion etching, the etching is almost exclusively by chemical reaction with the reactive species formed in the plasma; the method must be regarded as a dry chemical etching technique rather than a mechanical technique. The latter techniques are almost exclusively used in semiconductor settings. It is possible with these ion-machining techniques, in combination with a mask, to etch almost any material in a very controlled fashion (dimensions <100 nm).

Fast Atom Beam Machining A variation on FIB machining is fast atom beam (FAB) machining. In FAB, ionized atoms, accelerated to about 100 km/s, are neutralized and used as the mechanical cutting tool. For example, the etching rate of GaAs with chlorine atoms at 60°C is 0.15 μm/min, and vertical walls are obtained. The work piece is mounted on a lathe as in FIB, and by rotating it with respect to a Ni mask, the different faces of the GaAs work piece are etched, and 3D structures are created. Hatamura et al.[11] produced a multifaced microstructure, a micro-gojyunoto (named after an old Japanese temple tower) of 280 × 100 × 100 μm this way. A hole etched in a polyimide sheet enabled the erection of the temple with the help of the rotational robot holding the temple. More recently the same group has extended the FAB capabilities into the nanodomain

by producing ultrafine stairs, 30 nm wide and 30 nm high. Their results show that these FAB methods are effective in producing 3D microstructures and nanostructures. Combinations of these methods will make it possible to produce functional nanostructures on 3D microstructures. As we will detail in Volume III, Chapter 4 on packaging, assembly, and self-assembly, Hatamura et al. came up with several intriguing microassembly ideas, just as in the erection of the microtemple mentioned above.

Desktop Factory

Despite all the progress in mechanical micromachining reported thus far, the fact that it often takes a two-ton machine tool to fabricate microparts where cutting forces are in the milli- to micro-Newton range is a clear indication that a complete machine tool redesign is required for the fabrication of micro- and nanomachines.[40] Even with the advancement in the precision and tolerances achieved in mechanical machining, ultraprecision machining is still mostly carried out on large equipment. This equipment consumes significant amounts of floor space and energy. Figure 6.34 illustrates the volumetric utilization [V = volume of machine tool (M)/volume of work piece (W)] of conventional and mesoscale machine tools.[41] The ideal equipment should be of the same scale as the final product to constitute an optimized manufacturing system (V = 1).

The size of the equipment producing precision parts has been reduced, but the optimal size has not yet been attained. Along this line, in Japan, the concept for desktop flexible manufacturing systems, or more simply "desktop factory" (DTF), for building micromachines was proposed in the early 1990s.[42] This effort was supported by Japan's MITI (Agency of Industrial Science and Technology). In Korea, the government started a new DTF national effort at the end of 2004, and similar projects are now underway in the EEC.[12] The proposed manufacturing units would be tabletop size and include universal chuck modules to which work pieces would be continuously clamped through most of the manufacturing process. By 1999 the world's first prototype of a DTF (50 cm × 70 cm) combining miniature machine tools and manipulators was developed at AIST in Japan as shown in Figure 6.4. The factory combined a lathe, milling machine, press machine, a transfer arm, and a two-fingered hand. A next portable factory made at AIST, requiring a single 100-V AC outlet only, was packaged in a suitcase (625 mm long, 490 mm wide, and 380 mm high with a weight of about 34 kg). DTFs like these have the potential of becoming the factory of the future: totally self-contained, robotic, desktop-size machine tools that only require materials, power, and water as outside inputs. A first commercial desktop manufacturing unit by Sankyo Seiki is shown as the inset in the header of this chapter. Sankyo Seiki's management believes that DTFs might help manufacturing survive in Japan.[12]

As an example of a thermal micromachine tool, consider the Panasonic hand-held μ-EDM discussed in Chapter 5 (Figure 5.3). Another DTF reviewed in Chapter 5, combining thermal and mechanical machining, is the high-precision four-axes CNC machine developed by Liao et al. at the National Taiwan University. Parts in the size range

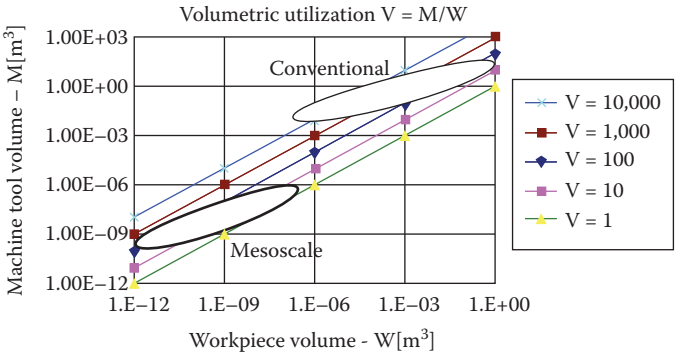


FIGURE 6.34 Volumetric utilization (V = volume of machine tool/volume work piece) of conventional and mesoscale machines. (Vogler, M.P., et al. 2002. *Transactions of NAMRI/SME* 30:653.[41])

of 0.1–2 mm can be produced on this machine tool, and all manufacturing processes are completed on the same machine without moving the work piece. This multifunction high-precision machine (accuracy of 1 μm) is capable of performing four different functions: milling, μ-EDM, μ-EDM-WC, and on-line measuring of the flatness, circularity, concentricity, dimensions, and geometric tolerances of the work piece (Figure 5.8).

In one scenario, DTFs become very automated and require very few human operators, so that exporting these machines—and thus transferring manufacturing from rich to poor countries—does not make sense anymore because no labor costs are saved. In the capitalistic/nationalistic approach, these micromanufacturing stations are owned by big centralized national entities. The other scenario, promoted by, for example, MIT's Neil Gershenfeld, sees desktop manufacturing becoming as popular as desktop publishing and the manufacturing equipment would be owned by as many people as possible.[43] In the latter, more socially responsible approach, manufacturing becomes much less centralized, and many more people get involved. In either case, the impact on society would be dramatic. The dilemma is similar to that of solar energy versus oil; solar energy is widely spread out and difficult to own/control whereas the Bush family and cohorts try to own oil that, because it is more centralized, is easier to control.

Besides high-accuracy cutting, detailed above, there is also significant progress to report in die punching. The miniature die press from Aoki and Takahashi was one of the first examples of progress in the direction of microdie punching machine tools.[44] Aoki et al. pushed the art of die forming with a press for 3D, metallic medical microcomponents to new limits.[44] Medical forceps, for insertion into an endoscope used to remove diseased tissue, have been fabricated as an example of the new capabilities. The forceps have a diameter of 0.6–1 mm and a length of 10 mm, with assembly holes measuring 0.3 mm, and 0.2-mm long teeth in the cutting tool. The final manufacturing time, including assembly, is about 3 min/piece. Some of the other successes in making DTFs are detailed in Table 6.5.

Precision Mechanical Machining versus IC-Based Micromachining

In comparing IC-based micro- and nanomachining with precision mechanically machined structures, Slocum makes some important observations.[4] He notes that, although IC-derived micro- and nanomachines are impressive for their small size, their relative accuracy is two orders of magnitude worse than is typically achieved in mechanical precision machining, which, moreover, makes for much more complex designs (see also Figure 6.3). For more than 100 years, precise mechanical machined devices have been designed and built with mechanical removal techniques achieving part-per-million (ppm) relative tolerances, and recently ultraprecision mechanical machining has entered the realm of 0.1 ppm relative tolerances. Typical IC-derived micro- and nanomachines, on the other hand, Slocum points out, are comparable with the macromachines of the early 1700s with respect to complexity and relative tolerance. Although the surface roughness of IC-derived micro- and nanomachines appears high compared with the specular finishes of precision-machined bearing surfaces, in reality, the absolute roughness is about the same. Another key issue Slocum brings up is that of metrology systems; verification of the fabricated geometry and tolerances is much more difficult for IC-derived micromachines because measurement of a displacement of 1 part in 10,000 will typically get one down to the nanometer level.

Higher-accuracy machining is needed to provide computer memory disks and optical mirrors and lenses with accuracies to a fraction of the wavelength of light. The micro- and nanofabrication techniques that are currently used in IC-derived micro- and nanomanufacturing have serious limitations in terms of the materials that can be used and the creation of complicated shapes with the required surface quality.

A better understanding of the limits of ultraprecision mechanical machining and nontraditional IC-based machining methods would help the mechanical engineer make the best choice in machining tools. At the same time, as IC-based micro- and nanomachine complexity increases, it seems appropriate

Table 6.5 Desktop Factories

Developed by	Equipment Category	Dimension of the Equipment	Features of the Equipment
Kitahara et al.[45]	Microlathe	32 mm length	*x-y* driving unit driven by laminated piezoactuators, main shaft device driven by a micromotor
Lu and Yoneyama[46]	Microlathe	200 mm length	Cutting mechanism directly observed by optical microscope
Nakao and Hatamura[47]	Nanomanufacturing factory	1.6 m × 0.9 m × 1.6 m	Fast atom beam (FAB) etching technology, concentrated motion manipulator, multidirectional SEM and an image-driven operating system
Vogler et al.[41]	Mesoscale machine tool (mMT) systems	25 mm × 25 mm × 25 mm	Air turbine spindle, piezoelectric friction drive, voice coil drives
McKeown[48]	Three-axis ultraprecision aspheric generator		High-stiffness servo drives and hydrostatic bearings, servo resolution of 1.25 nm in the *x*- and z-axes
Hatamura et al.[11]	Integrated manufacturing system for 3D microstructures	1.6 m × 0.9 m × 1.6 m	Fast atom beam etching, 3D assembly using a concentric manipulator, multiview scanning electron microscope (SEM)
Ruiz-Huerta[49]	CNC micromachining tool	130 mm × 160 mm × 85 mm	PC-based control
Mishima et al.[50]	Microfactory	Floor area: 50 × 70 cm	A microlathe, a micropress machine, a micromilling machine, a microtransfer arm, and a micro two-fingered hand
NIST (National Institute for Science and Technology)	Molecular measuring machine (M3)		Images and measures to nanometer accuracy the positions of features located anywhere within the area of 50 mm × 50 mm and up to 100 μm high
Takeuchi et al.[51]	5-axis control ultraprecision machining center	20-mm stroke for vertical movement, 1-mm minimum feature size	Air-borne linear guide and air slide for the translational axes

for the designers of these types of micro- and nanostructures, who are typically electrical engineers, to study ultraprecision machine design.

Acknowledgments

Special thanks to Dr. Rodrigo Martinez and Dr. Chunlei Wang.

Questions

Questions by Dr. Madou assisted by Mr. Chuan Zhang, UC Irvine

6.1: What are some commonly recognized categories of mechanical removal processes?

6.2: What is the meaning of relative and absolute tolerance? Is it sufficient to use only absolute tolerances in describing the precision of a micromachining process?

6.3: What are the merits of single crystal diamond cutting tools?

6.4: Illustrate the importance of material properties in determining machining accuracy.

6.5: List some important trends and their consequences in the development of automated mechanical machining equipment (consider only single-point mechanical removal) for the manufacturing of the future.

6.6: What are the requirements for a material to be used as an abrasive grit? List some conventional abrasives.

6.7: How does the grit size affect the quality of surface finish and the material removal rates? How should the grit size be chosen when dealing with different materials?

6.8: Why must the grinding wheel be dressed periodically during the grinding process? How does electrolytic in-process dressing (ELID) solve this problem?

6.9: Why are soft materials unsuitable to be machined with ultrasonic drilling? How do hardness and modulus of elasticity impact the removal rates?

6.10: What are the advantages and disadvantages of water jet machining? If we have a piece of copper plate 1 in. thick and the heat-affected zone (HAZ) is of great concern, should we choose laser machining, water jet machining, or abrasive water jet machining? Why?

6.11: How can a wider range of materials be machined by adding abrasives to a water jet? List some applications of abrasive water jet machining.

6.12: Show how the erosion efficiency depends on the kinetic energy in powder blasting. What other factors influence the erosion efficiency?

6.13: What are the major applications of focused ion beam (FIB)? How do we operate FIB in different modes for different applications?

6.14: What are the merits of ultrasonic machining? How does a rotary setup enhance its performance?

6.15: Explain the working principle of ultrasonic machining. Why does it produce stress-free shapes?

6.16: What are the differences and similarities between focused ion beam (FIB) machining and fast atom beam (FAB) machining?

6.17: Why are micromanufacturing researchers developing desk top factories (DTFs)? Why should the ideal equipment be of the same size scale as the final product?

6.18: What are the merits and drawbacks of precision mechanical machining compared with IC-based micromachining? List two examples that are better machined with precision mechanical machining and two with IC-based micromachining.

6.19: What are the advantages of electrodischarge wire cutting (EDWC) compared with sinker EDM?

6.20: Fill in the blanks:

Precision machining is defined at a relative tolerance of or less of a feature/part size and covers both and processes.

6.21: Give a history of mankind's manufacturing methods, including major events, rough dates, and examples.

6.22: What are the underlying reasons that both Taniguchi and Moore curves have started showing signs of a slowdown in progress of manufacturing accuracy and of transistor density over time, respectively?

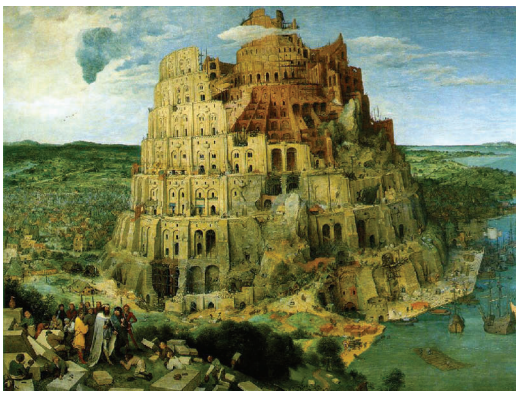
References

1. Hodges, H. 1970. *Technology in the ancient world*. New York: Barnes & Noble.
2. Burke, J. 1978, *Connections*. Boston: Back Bay Books.
3. Zuurveen, F. 1994. *Precisie-Technologie-Jaarboek 1994*. Eindhoven, the Netherlands: NVFT.
4. Slocum, A. H. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92)*. Travemunde, Germany.
5. Slocum, A. H. 1992. *Precision machine design*. Englewood Cliffs, NJ: Prentice Hall.
6. Palmer, C. 2000. *Diffraction grating handbook*, 4th ed. Rochester, NY: Richardson Grating Laboratory.
7. *Compton's interactive encyclopedia, 1994*. Elmhurst, IL: Interactive Multimedia.
8. Taniguchi, N. 1983. Current status in, and future trends of, ultraprecision machining and ultrafine materials processing. *Annals CIRP* 32:S573–82.
9. Gibbs, W. W. 1997. Gordon E. Moore interview. *Scientific American* September 22.
10. Szepesi, D. 1993. *Sensoren en Actuatoren in de Werktuigbouw/Machinebouw, Centrum voor Micro-Electronica*. The Hague, the Netherlands.
11. Hatamura, Y., M. Nakao, T. Sato, K. Koyano, K. Ichiki, H. Sangu, M. Hatakeyama, T. Kobata, and K. Nagai. 1994. Construction of 3D micro structure by multi-face FAB, cofocus rotational robot and various mechanical tools. *IEEE international workshop on micro electro mechanical systems (MEMS '94)*. Oiso, Japan.
12. Ehmann, K. F., D. Bourell, M. L. Culpepper, T. J. Hodgson, T. R. Kurfess, M. Madou, K. Rajurkar, and R. DeVor. 2007. *Micromanufacturing: international research and development*. New York: Springer.
13. Evans, C. 1989. *Precision engineering: an evolutionary view*. Bedford, UK: Cranfield Press.
14. Snoeys, R. 1986. *Advances in non-traditional machining*. Anaheim, CA.
15. McKeown, P. A. 1986. High precision manufacturing and the British economy. *Institute of Mechanical Engineers Proceedings* 200:76.
16. Bier, W., G. Linder, D. Seidel, and K. Schubert. 1991. Kernforschungszentrum Karlsruhe. *KfK Nachrichten* 23:2–3, 165–73.
17. Teshigahara, A., M. Watanabe, N. Kawahara, Y. Ohtsuka, and T. Hattori. 1995. Performance of a 7-mm microfabricated car. *J Microelectromech Syst* 4:76–80.

18. Boothroyd, G., and W. A. Knight. 1989. *Fundamentals of machining and machine tools*. New York: Marcel Dekker.
19. Ohmori, H., and T. Nakagawa. 1990. Mirror surface grinding of silicon wafer with electrolytic in-process dressing. *Annals CIRP* 39:329–32.
20. Bellows, G., and J. B. Kohls. 1982. Drilling without drills. *Am Mach Special Rep* 743:173–88.
21. Moreland, M. A. 1992. *Engineered materials handbook*. Ed. S. J. Schneider. Metals Park, OH: ASM International.
22. Humphrey, E. F., and D. H. Tarumoto, eds. *Fluidics*. Boston: Fluidic Amplifier Associates.
23. Sun, X.-Q., T. Masuzawa, and M. Fujino. *Ninth annual international workshop on micro electro mechanical systems (MEMS '96)*. San Diego.
24. Grass, P., W. Koenig, F. U. Meis, C. Wulf, and H. Willerscheid. 1984. Konturbearbeitung faserverstaerkter Kunststoffe. Wasserstrahlschneiden, laserstrahlschneiden und umrissfraesen im wettbewerb. *VDI-Zeitschrift*. 126(21): S.785ff.
25. Benedict, G. 1987. *Nontraditional manufacturing processes*. New York: Marcel Dekker.
26. Momber, A. W., and Kovacevic, R. 1998. *Principles of abrasive water jet machining*. Berlin: Springer.
27. Dombrowski, T. J. 1983. The how and why of abrasive jet machining. *Modern Machine Shop* 76.
28. Ligthart, H. J., P. J. Slikkerveer, F. H. in't Veld, P. H. W. Swinkels, and M. H. Zonneveld. 1996. Glass and glass machining in ZEUS panels. *Philips J Res* 50:475–99.
29. Kruusing, A., S. Leppävuori, A. Uusimäki, and M. Uusimäki. 1999. Rapid prototyping of silicon structures by aid of laser and abrasive-jet machining. *SPIE Proceedings* 3680:870–78.
30. Yamahata, C. 2005. Magnetically actuated micropumps. Thesis. EPF de Lausanne, Lausanne, Switzerland.
31. Slikkerveer, P. J. 1999. Mechanical etching of glass by powder blasting. PhD diss., Technische Universiteit Eindhoven.
32. Slikkerveer, P. J., P. C. P. Bouten, F. H. in't Veld, and H. Scholten. 1998. Erosion and damage by sharp particles. *Wear* 217:237–50.
33. Belloy, E., P. Q. Pham, A. Sayah, and M. A. M. Gijs. 2000. *Eurosensors XIV*. Eds. R. de Reus and S. Bouwstra. Copenhagen: MIC-Mikroelektronik Centret.
34. Solignac, D., A. Sayah, S. Constantin, R. Freitag, and M. A. M. Gijs. 2000. Powder blasting as a novel technique for the realisation of capillary electrophoresis chips. *Eurosensors XIV*. Eds. R. de Reus and S. Bouwstra. Copenhagen: MIC-Mikroelektronik Centret.
35. Bornand, E., F. Amalou, and M. A. M. Gijs. 2000. *Eurosensors XIV*. Eds. R. de Reus and S. Bouwstra. Copenhagen: MIC-Mikroelektronik Centret.
36. Belloy, E., A. Sayah, and M. A. M. Gijs. 2000. Powder blasting for three dimensional microstructuring of glass. *Sensors Actuators A* 86:231–37.
37. Kalpakjian, S. 1984. *Manufacturing processes for engineering materials*. Reading, MA: Addison-Wesley.
38. Vasile, M. J., C. Biddick, and S. A. Schwalm. 1993. *Micromechanical systems.* Eds A. P. Pisano, J. Jara-Almonte, and W. Trimmer. New York: ASME Press, p. 81.
39. Vasile, M. J., C. Biddick, and S. A. Schwalm. 1994. Microfabrication by ion milling: the lathe technique. *J Vac Sci Technol* B12:2388.
40. Ehmann, K. E., D. Bourell, M. L. Culpepper, T. J. Hodgson, T. R. Kurfess, M. Madou, K. Rajurkar, and D. DeVor. 2007. *Micromanufacturing: international assessment of research and development*. Dordrecht, the Netherlands: Springer. www.wtec.org/micromfg/report/micro-report.pdf
41. Vogler, M. P., X. Liu, S. G. Kapoor, R. E. DeVor, and K. F. Ehmann. 2002. Development of meso-scale machine tool (mMT) systems. *Transactions of NAMRI/SME* 30:653–61.
42. Higuchi, T., and Y. Yamagata. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL.
43. Gershenfeld, N. 2005. *FAB: the coming revolution on your desktop—from personal computers to personal fabrication*. New York: Basic Books.
44. Aoki, I., and T. Takahashi. 1996. *SPIE: micromachining and microfabrication process technology II*. Austin, TX.
45. Kitahara, T., Y. Ishikawa, K. Terada, N. Nakajima, and K. Furuta. 1996. Development of microlathe. *J Mechanical Engineering Lab* 50:5, 117–23.
46. Lu, Z., and T. Yoneyama. 1999. Micro cutting in the micro lathe turning system. *Int J Machine Tools Manufact* 39:1171–83.
47. Nakao, M., Y. Hatamura, and S. Tomomasa. 1996. *Microrobotics: components and applications*. Ed. A. Sulzmann. Boston: SPIE.
48. McKeown, P. 1996. From micro- to nano-machining–towards the nanometre era. *Sensor Review* 16:4–10.
49. Ruiz-Huerta, L., A. Caballero Ruiz, and E. Kussul. 2002. Guidelines for design low cost micromechanics. *Proceedings ASPE*. St. Louis, MO, October 20–25, 2002.
50. Mishima, N., K. Ashida, T. Tanikawa, H. Maekawa, K. Kaneko, and M. Tanaka. 2000. Micro-factory and a design evaluation method for miniature machine tools. *Proc. of Int. Workshop on Microfactories*, pp. 155–58.
51. Takeuchi, Y., H. Yonekura, and K. Sawada. 2003. Creation of 3D tiny statue by 5-axis control ultraprecision machining. *Computer Aided Design* 35:403–09.

Part III

Pattern Transfer with Additive Techniques



Building structures with additive technologies. *The Tower of Babel, c.1563 by Pieter Bruegel the Elder.*

> As an adolescent I aspired to lasting fame, I craved factual certainty, and I thirsted for a meaningful vision of human life—so I became a scientist. This is like becoming an archbishop so you can meet girls.
>
> **Matt Cartmill**

> If any student comes to me and says he wants to be useful to mankind and go into research to alleviate human suffering, I advise him to go into

charity instead. Research wants real egotists who seek their own pleasure and satisfaction, but find it in solving the puzzles of nature.

Albert Szent-Györgi

Introduction to Part III

Chapter 7 Physical and Chemical Vapor Deposition, Thin Film Properties, and Surface Micromachining
Chapter 8 Chemical-, Photochemical-, and Electrochemical-Based Forming
Chapter 9 Thermal Energy-Based Forming
Chapter 10 Micromolding Techniques—LIGA

Introduction to Part III

In Part III we cover additive (forming) processes, where materials are added, usually in a selective manner, to a work piece or a device under construction. In manufacturing "forming" creates an original shape from a molten mass, a solution, the gaseous state, plasma, or solid particles, and during such processes cohesion is created among individual particles. Forming techniques are often accompanied or followed by thermal processing to obtain desired material properties and substrate adhesion. Subtractive manufacturing techniques, covered in Part II, are categorized according to the energy used at the work piece as *mechanical*, *chemical*, or *thermal*. This classification—straightforward in the case of subtractive processes—breaks down quickly when applied to additive processes. Take physical vapor deposition (PVD), covered in Chapter 7, a method classified as a mechanical additive technique because material is deposited onto the work piece by direct line-of-sight impingement deposition. However, in additive techniques the depositing material itself might be generated through the application of thermal energy (e.g., thermal evaporation) or mechanical energy (e.g., sputter deposition) to a source material. In such case, the rule is to categorize the technique with the name of the process that brings the depositing material to the substrate, not for the process that generates the depositing material. For example, laser ablation deposition, also called *laser sputter deposition* (Chapter 7), uses thermal energy to ablate the material to be deposited but must be categorized as a mechanical method because the depositing material reaches the surface in a line-of-sight manner (mechanical). Most complicated to categorize are forming methods where chemical and thermal effects are intertwined in the deposition process on the substrate itself. Along this line, laser-assisted chemical vapor deposition (LCVD) (Chapter 9) is called a *thermal technique* because the heat of the laser beam induces the local material deposition, but because it involves a chemical reaction at the surface it could also have been termed a *chemical additive process*. (Remember that subtractive laser machining is a purely thermal technique; see Chapter 5.) In such cases one could refer to the technique as *thermochemical*, as in the case of the thermal oxidation of Si in a stream of humidified oxygen at 900°C, but in practice one classifies such methods as either *chemical* or *thermal*, depending on the primacy of the chemical reaction or the heat application, with the lower-temperature processes typically classified as *chemical* (e.g., sol-gel, screen printing of organics, and ink-jetting) and the higher-temperature processes as *thermal* (e.g., LCVD, screen printing of inorganics, and spray pyrolysis).

Although deposition methods in MEMS and NEMS, especially in the thin-film arena, are generally the same as in integrated circuits (ICs), additive processes in miniaturization science span a much wider and diverse range from inorganic to organic materials. Besides the typical microelectronic elements (Si, Al, Au, Ti, W, Cu, Cr, O, N, and Ni-Fe alloys), miniaturization science involves deposition of several nontypical elements such as Zr, Ta, Ir, C, Pt, Pd, Ag, Zn, In, Nb, and Sn. Moreover, a plethora of exotic compounds ranging from enzymes to shape memory alloys and from hydrogels to piezoelectrics is used in miniaturization science. In comparison, the number of materials and compounds involved in IC fabrication is very limited. In miniaturization science, particularly in chemical sensors and biomedical devices, modular, thick-film deposition technologies are more prevalent than in the IC industry.

In Table III.1 we list IC and MEMS/NEMS materials, typical deposition techniques, and an example

Table III.1 IC and MEMS/NEMS Materials, Deposition Method, and Typical Application

Material	Deposition Technique	Typical Application
Organic thin films		
Hydrogel	Silk screen	Internal electrolyte in chem. sensors
Photoresist (0.1–2 μm)	Spin-on	Masking, planarization
Thick photoresist (10–1000 μm)	Casting	Mold for metal plating
DNA	Ink-jetting	DNA array for molecular diagnostics
Polyimide	Spin-on	Electrical isolation, planarization, microstructures
Metal oxides		
Aluminum oxide	CVD, sputtering, anodization	Electrical isolation
Indium oxide	Sputtering	Semiconductor
Tantalum oxide	CVD, sputtering, anodization	Electrical isolation
Tin oxide (SnO_2)	Sputtering	Semiconductor in gas sensors
Zinc oxide	Sputtering	Electrical isolation, piezoelectric
ZrO_2	Thermal spray deposition from plasmas or flames	Coatings for aircraft engine parts and ZrO_2 sensors
Compound semiconductors		
Cadmium sulphide (II–VI compound)	Spray pyrolysis (CVD)	Solar cell
Gallium arsenide (III–V compound)	Metalorganic CVD (MOCVD)	Optoelectronic devices
Gallium-aluminum arsenide	Molecular beam epitaxy (MBE)	A semiconductor alloy used as the light confinement layer in both single- and double-heterostructure diode lasers
Noncrystalline silicon compounds		
α-Si-H	CVD, sputtering, plasma CVD	Semiconductors
Polysilicon	CVD, sputtering, plasma CVD	Conductor, microstructures
Silicides	CVD, sputtering, plasma CVD, alloying of metal and silicon	Conductors
Metals (thin films) with ρ in μΩ cm in parentheses		
Silver (Ag) (1.58)	Evaporation, sputtering	Electrochemistry electrode
Aluminum (Al) (2.7)	Evaporation, sputtering, plasma CVD	Electrical interconnects, limited to operations <300°C
Chromium (Cr) (12.9)	Evaporation, sputtering, electroplating	Electrical conduction, adhesion layer (10–100 nm)
Gold (Au) (2.4)	Evaporation, sputtering, electroplating	Electrical interconnects for higher temperatures than Al, optical reflection in the IR
Iridium (Ir) (5.1)	Sputtering	Electrochemistry electrode, biopotential measurements
Molybdenum	Sputtering	Electrical conduction
Platinum (Pt) (10.6)	Sputtering	Electrochemistry electrode, biopotential measurements
Palladium (Pd) (10.8)	Sputtering	Electrical conduction, adhesion layer, electrochemistry electrode, solder wetting layer
Tungsten (W) (5.5)	Sputtering	Electrical interconnects at higher temperatures
Titanium (Ti) (42)	Sputtering	Adhesion layer
Copper (Cu) (1.7)	Sputtering	Low resistivity interconnects
Alloys with ρ in μΩ cm in parentheses		
Al-Si-Cu	Evaporation, sputtering	Electrical conduction
Nichrome (NiCr) (200–500)	Evaporation, sputtering	Thin film laser-trimmed resistor
Permalloy™ (Ni_xFe_y)	Sputtering	Magnetoresistor, thermistor
TiNi (80)	Sputtering	Shape memory alloy

(*Continued*)

Table III.1 IC and MEMS/NEMS Materials, Deposition Method, and Typical Application (*Continued*)

Material	Deposition Technique	Typical Application
Chemically/physically modified silicon		
n/p-type silicon	Implantation, diffusion, incorporation in the melt	Conduction modulation, etch stop
Porous silicon	Anodization	Electrical isolation, light-emitting structures, porous junctions
Silicon dioxide	Thermal oxidation, sputtering, anodization, implantation, CVD	Electrical and thermal isolation, masking, encapsulation
Silicon nitride	Plasma-enhanced CVD (PECVD)	Electrical and thermal isolation, masking, encapsulation

application for the deposit. As in the case of subtractive processes (Part II), only the underlying physical and chemical principles of the various additive processes dealt with in Part III are explored.

Physical vapor deposition (PVD) processes and chemical vapor deposition (CVD) are topics of Chapter 7. In the same chapter we survey thin-film properties and detail surface micromachining, an important application of thin-film deposition techniques. In surface micromachining, features are built up, layer by layer, on the surface of a substrate (e.g., a single-crystal silicon wafer). Dry etching defines the surface features in the x,y-plane, and wet etching releases them from the plane by undercutting. In surface micromachining, shapes in the x,y-plane are unrestricted by the crystallography of the substrate.

In Chapter 8 we deal with additive chemical, photochemical, and electrochemical processes.* In chemical forming, chemical reactions create features by forming new compounds. The energy at work in these additive processes is chemical. Some of the best-known examples, with the chemical reaction dominant over the heat application, are sol-gel, reaction injection molding (RIM), colloidal processes, and self-assembled monolayers (SAMs). With the growth of the bio-MEMS field, techniques for depositing organic materials for chemical and biological sensors, often arranged in some type of an array configuration, are gaining importance. The chemical-forming methods covered in Chapter 8 are geared toward bio-MEMS applications. In photochemical forming, photoenergy solidifies a material into a 3D shape, as illustrated in rapid prototyping and microphotoforming (stereolithography). In the case of electrochemical forming, the energy at work is electrochemical. Electrochemical and electroless metal deposition are gaining renewed interest because of their emerging importance in making mold inserts for replication methodologies in both MEMS and NEMS (see also Chapter 10).

Chapter 9 covers thermal forming techniques where thermal energy provided by a heat source "transforms" a material as in crystallization, sintering, alloying, annealing, decomposition, and pyrolizing. Chemical and thermal forming techniques are often intertwined because reactions are induced by combining different chemicals and are aided by heat and catalysts. Along this line, screen printing is treated as a thermal process when dealing with inorganic material deposition (high temperature) and as a chemical process when depositing organics (low temperature). Similarly, the high-temperature ceramic application of tape casting (also called *doctor blading* or *knife coating* when dealing with paper, polymers, and paints) is treated as a thermal process where it is called *green tape technology*.

Replication methods based on plastic molding and LIGA—a sophisticated type of replication technique combining x-ray lithography with electroplating and plastic molding—are discussed in Chapter 10. Plastic molding techniques include RIM, thermoplastic injection molding (IM), and compression molding or hot embossing. Of the three, only RIM is a chemical-forming technique; the other two are thermal methods. Plastic

* Under subtractive chemical processes (Part II), we grouped purely chemical processes and electrochemical and photochemical machining together; in Part III we treat additive chemical, electrochemical, and photochemical as clearly distinct processes.

molding technology often is the least expensive manufacturing technique available. One only has to look around the home or office to recognize the pervasive nature of highly intricate, very low-cost plastic parts surrounding us. This method has the capability of making truly 3D parts with a wide range of feature sizes. Master molds can be made with both nonlithography- (e.g., CNC-machined master mold) and lithography-based (e.g., LIGA mold) techniques.

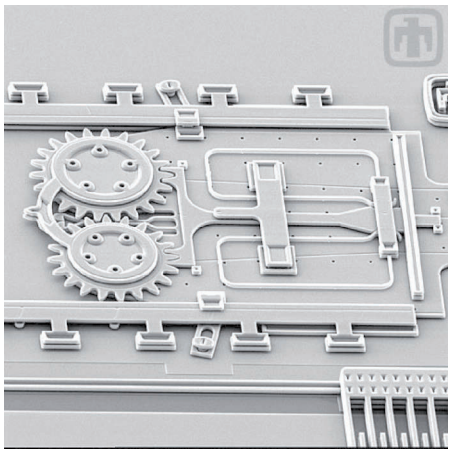
Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

7

Physical and Chemical Vapor Deposition—Thin Film Properties and Surface Micromachining

Outline

Introduction
Physical Vapor Deposition
Chemical Vapor Deposition
Mechanical Properties of Thin Films
Surface Micromachining
Questions
References



Surface micromachining. Clutch mechanism made in polysilicon. Gears are 50 μm across. Courtesy Sandia National Lab.

Introduction

Chapter 7 introduces physical vapor deposition (PVD) and chemical vapor deposition (CVD). PVD constitutes a group of direct line-of-sight impingement deposition techniques, and CVD represents a class of diffusive-convective mass transfer deposition techniques. Evaporation, sputtering, laser ablation deposition, ion plating, cluster deposition,

molecular beam epitaxy (MBE), and aerosol deposition (AD) represent the PVD techniques discussed. CVD techniques covered are plasma-enhanced CVD (PECVD), high-density plasma CVD (HDPCVD), atmospheric pressure CVD (APCVD), low-pressure CVD (LPCVD), very low-pressure CVD (VLPCVD), electron cyclotron resonance CVD (ECRCVD), metalorganic CVD (MOCVD), and atomic layer deposition (ALD).

Epitaxial techniques arrange atoms in single-crystal fashion on a crystalline substrate that acts as a seed crystal so that the lattice of the newly grown film duplicates that of the substrate. If the film is of the same material as the substrate, the process is called *homoepitaxy*, *epitaxy*, or simply *epi*. Epi deposition represents one of the cornerstone techniques for building micromachines, and a separate section reviews progress in epitaxial methods. Epitaxial techniques are usually chemical (CVD) in nature but may also be physical (e.g., MBE is a PVD technique).

The mechanical properties of thin CVD and PVD films are discussed next, and their applications in surface micromachining are detailed. We review thin film material properties in general, focusing on significant differences with bulk properties of the same material. Because of the complexity of the many parameters influencing thin film properties, we present the deposition of polysilicon as a case study. This is followed by a review of the main surface micromachining processes. Bulk micromachining means that three-dimensional (3D) features are etched into the bulk of crystalline and noncrystalline materials (see Chapter 4). In contrast, surface micromachined features are built up, layer by layer, on the surface of a substrate (e.g., a single crystal silicon wafer). Dry etching defines the surface features in the *x*,*y*-plane, and wet etching releases them from the plane by undercutting. In surface micromachining, shapes in the *x*,*y*-plane are unrestricted by the crystallography of the substrate. For illustration, in Figure 7.1 we compare typical bulk micromachining etching (both dry and wet) with polysilicon surface micromachining. Not reflected in this figure is the fact that the surface micromachined devices typically end up a bit smaller than their bulk micromachined counterparts.

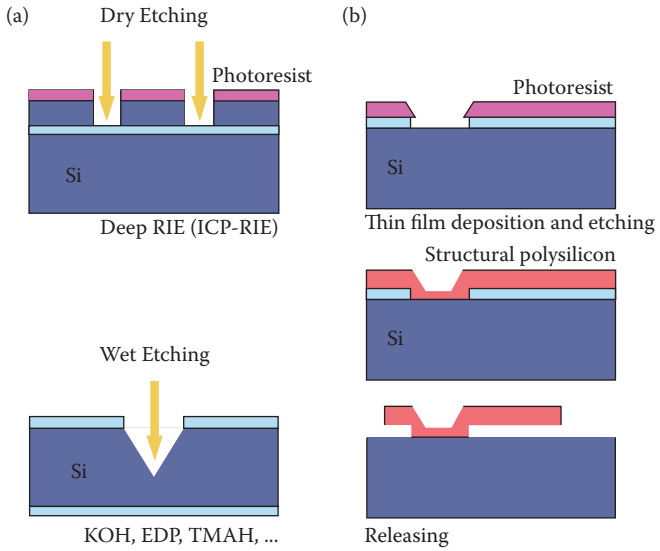


FIGURE 7.1 Typical bulk micromachining (dry and wet) (a) versus polysilicon surface micromachining process steps (b). (Figure by Dr. Han Xu.) Acronyms such as RIE, ICP-RIE, EDP and TMAH are explained in this chapter.

Physical Vapor Deposition

Introduction

Many different kinds of thin films in integrated circuits (ICs) and micro- and nanomachines are deposited by evaporation and sputtering, both of which are examples of physical vapor deposition (PVD). The technique represents a very versatile coating technology applicable to an almost unlimited combination of coating substances and substrate materials. PVD reactors may use solid, liquid, or vapor as raw source material in a variety of configurations and are used to apply a wide variety of metals, alloys, ceramics, and other inorganic compounds, and even certain polymers. However, in practice, the technique is mostly used to deposit metals. Possible substrates include semiconductors, metals, glass, and plastics. At the low pressures used in a PVD reactor, the vaporized material encounters few intermolecular collisions while traveling to the substrate. Consequently, modeling of deposition rates is a relatively straightforward exercise in geometry.

Most PVD processes consist of the following elementary steps: 1) the synthesis of the coating vapor from the source material, 2) transport of the vapor to the substrate, and 3) condensation of vapors onto the surface of the substrate. These steps are generally

Table 7.1 Principal Types of PVD Techniques

PVD Process	Features and Coating Materials
Thermal evaporation	Features: Equipment is relatively low cost and simple; deposition of compounds is difficult; coating adhesion not as good as for other PVD processes Typical coating materials: Ag, Al, Au, Cr, Cu, Mo, W
Sputtering	Features: Better throwing power and coating adhesion than vacuum evaporation; can coat compounds; slower deposition rates and more difficult process control than vacuum evaporation Typical coating materials: Al_2O_2, Au, Cr, Mo, SiO_2, Si_3N_4, TiC, TiN
Ion plating and cluster deposition	Features: Best coverage and coating adhesion of PVD processes; most complex process control; higher deposition rates than sputtering Typical coating materials: Ag, Au, Cr, Mo, Si_3N_4, TiC, TiN
Laser sputter deposition or laser ablation deposition	Features: Best for complex compound deposition Typical coating materials: high-temperature superconductor $YBa_2Cu_3O_{7-x}$ and biocompatible calcium hydroxylapatite, or $Ca_{10}(PO_4)_6(OH)_2$
Aerosol deposition	Features: Lowest temperature deposition Typical coating materials: ceramic coatings on all types of substrates

carried out in a vacuum chamber; therefore, evacuation of the chamber precedes the PVD processes. The substrates are held upside down to minimize particulate contamination. The PVD techniques may be grouped into five principal types: 1) thermal evaporation, 2) sputtering, 3) ion plating and cluster deposition, 4) laser sputter deposition or laser ablation deposition, and 5) aerosol deposition as listed in Table 7.1. Ion plating and cluster deposition are based on a combination of evaporation and plasma ionization and offer some of the advantages inherent to both techniques. In aerosol deposition, solid particles are impacted at high speed on a substrate. This promising new process, we will learn, deviates significantly from all other PVD processes.

PVD is applied in the fabrication of electronic devices, microelectromechanical systems (MEMS), and nanoelectromechanical systems (NEMS), principally for depositing metal to form electrical connections. However, applications also include thin decorative coatings on plastic and metal parts such as trophies, toys, pens and pencils, watchcases, and interior trims in automobiles. The coatings in this case are thin films of aluminum (around 150 nm) coated with clear lacquer to give a high-gloss silver or chrome appearance. Another use of PVD is to apply antireflection coatings of magnesium fluoride (MgF_2) onto optical lenses. Ion plating is widely used to coat titanium nitride (TiN) onto cutting tools and plastic injection molds for wear resistance, and pulsed laser deposition (PLD) is particularly useful when dealing with complex compounds, as in the case of the deposition of high-temperature superconductor films (HTSC), for example, $YBa_2Cu_3O_{7-x}$. Another example application of PLD is the deposition of the biocompatible calcium-phosphate-based ceramic, calcium hydroxylapatite, or $Ca_{10}(PO_4)_6(OH)_2$. Aerosol impact deposition excels at making ceramic films at low temperature on all types of substrates.

Thermal Evaporation

Certain materials (mostly pure metals) can be deposited onto a substrate by transforming them first from the solid source to the vapor state in a vacuum and then letting them condense on the substrate surface. Thermal evaporation represents one of the oldest thin film deposition techniques. Evaporation is based on the boiling off (or sublimating) of a heated material onto a substrate in a vacuum. From thermodynamic considerations, the number of molecules leaving a unit area of evaporant per second, i.e., flux F, is given by:

$$F = N_0 e^{\frac{\phi_e}{KT}} \tag{7.1}$$

where N_0 is a slowly varying function of temperature (T) and Φ_e is the activation energy (in eV) required to evaporate one molecule of the material. The activation energy for evaporation is related to the enthalpy of formation of the evaporant, H, as $H = \Phi_e \times e \times N$ (Avogadro's number) J/mol.

Table 7.2 suggests the need for a good vacuum during evaporation; even at a pressure of 10^{-5} Torr,

Table 7.2 Kinetic Data for Air as a Function of Pressure

Pressure (Torr)	Mean Free Path (cm)	Number Impingement Rate (s^{-1}. cm^{-2})	Monolayer Impingement Rate (s^{-1})
10^{1}	0.5	3.8×10^{18}	4400
10^{-4}	51	3.8×10^{16}	44
10^{-5}	510	3.8×10^{15}	4.4
10^{-7}	5.1×10^{4}	3.8×10^{13}	4.4×10^{-2}
10^{-9}	5.1×10^{6}	3.8×10^{11}	4.4×10^{-4}

Source: Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press. With permission.[1]

4.4 contaminating monolayers per second will impinge on the substrate.[1] For reference purposes, the number of atoms per unit area corresponding to a monolayer for a metal is about 10^{15} atoms/cm^2. Moreover, to avoid reactions at the source (e.g., oxide impurities being formed), the oxygen partial pressure needs to be less than 10^{-8} Torr.

In a laboratory setting, metals are often evaporated by passing a high current through a highly refractory metal. A refractory metal (for example, W or Mo) is formed into a suitable container [e.g., a tungsten crucible (boat) or filament] to hold the source material. Current is applied to heat the container, which then heats the material in contact with it. One problem with this heating method is possible alloying between the holder and its contents so that the deposited film becomes contaminated with the metal of the container. This method is called *resistive heating* (Figure 7.2a). Resistive evaporation is simple but easily spreads contaminants present in the filament, and the small size of the filament limits the thickness of the deposited film. Because heating is accomplished in a vacuum, the temperature required for vaporization is significantly below the corresponding temperature required at atmospheric pressure. Also, the absence of air in the chamber prevents oxidation of the source material at the heating temperatures.

In industrial applications, resistive heating has been surpassed by electron beam (e-beam) and radiofrequency (RF) induction evaporation. In the e-beam mode of operation, a high-intensity e-beam gun (3–20 keV) is focused on the target material that is placed in a recess in a water-cooled copper hearth. As shown in Figure 7.2b, the e-beam is magnetically directed onto the evaporant, which melts locally. In this manner, the metal forms its own crucible, and the contact with the hearth is too cool for chemical reactions, resulting in fewer source-contamination problems than in the case of resistive heating. E-beam evaporation not only results in higher quality films, but it also provides a higher deposition rate

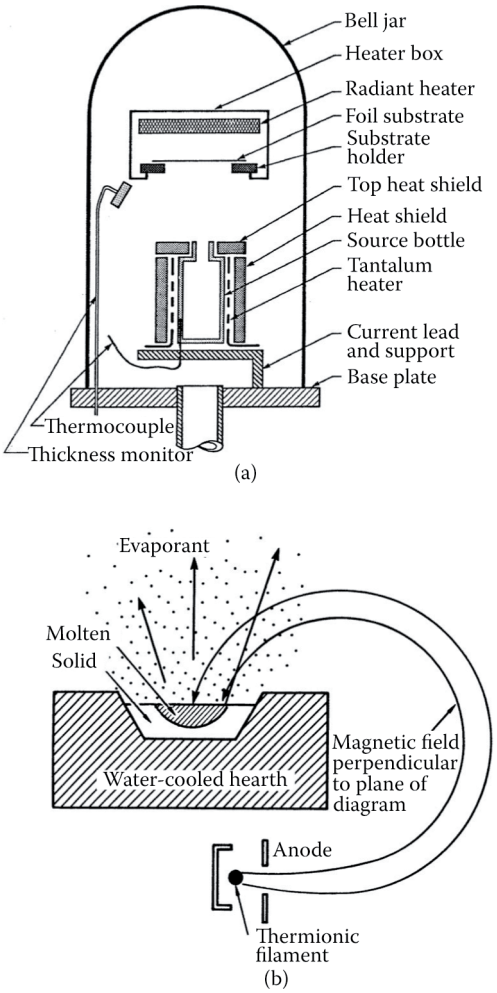


FIGURE 7.2 Thin film deposition by evaporation. (a) Typical evaporation setup. (b) Diagram of a magnetized deflection electron-beam evaporation system. (From Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press. With permission.[1])

(50–500 nm/min). Two disadvantages of e-beam evaporation are that the process might induce x-ray damage and possibly even some ion damage to the substrate (at voltages >10 kV, the incident e-beam causes x-ray emission) and that the deposition equipment is more complex. X-ray damage may be avoided by using a focused, high-power laser beam instead of an e-beam. However, this technique is expensive and has not yet penetrated commercial applications.

In RF induction heating, a water-cooled RF coupling coil surrounds a crucible with the material to be evaporated. Because about two-thirds of the RF energy is absorbed within one skin depth of the surface, the frequency of the RF supply must decrease as the size of the evaporant charge increases. With a charge of a few grams of evaporant, frequencies of several hundred kilohertz are sufficient.

In Table 7.3, we compare the different heat sources available for thermal evaporation.

The substances used most frequently for thin-film formation by evaporation are elements or simple compounds whose vapor pressures range from 1–10^{-2} Torr in the temperature interval from 600–1200°C. Refractory metals such as platinum, molybdenum, tantalum, and tungsten do not easily heat to the temperatures required to reach that vapor pressure range. To obtain a deposition rate high enough for practical applications, the vapor pressure at the source must be above the background pressure and reach at least 10^{-2} Torr. An evaporation source operating at a vapor pressure of 10^{-1} Torr, for example, can deliver rates of about 1000 atomic layers per second. The rate of atoms or molecules lost from the source as a result of evaporation, flux *F*, in molecules/unit area/unit time (Equation 7.1), can also be expressed in the following relationship derived from kinetic considerations on how the vapor pressure of the evaporant relates to the evaporation rate:

$$F = \frac{P_v(T)}{(2\pi MKT)^{\frac{1}{2}}} = 3.513\ 10^{22}\ \frac{P_v(T)}{(MT)^{\frac{1}{2}}} \qquad (7.2)$$

where M = molecular weight of the evaporant
T = source temperature in Kelvins
$P_V(T)$ = vapor pressure of the evaporant in Torr

Metal deposition rates of 0.5 µm/min are typical when operating in the 10-mTorr range. Whatever the vaporization technique, evaporated atoms leave the source and follow straight-line paths until they collide with other gas molecules or strike the substrate to be coated. If a high vacuum is established, most atoms/molecules will deposit on the substrate without suffering intervening collisions with other gas molecules. The fraction of particles scattered by collisions (N/N_0) with atoms of residual gas is proportional to:

$$\frac{N}{N_0} \propto 1 - e^{\frac{d}{\lambda}} \qquad (7.3)$$

where d = distance between source and substrate (see Figure 7.3a)
λ = mean free path of the particles

At 10^{-5} Torr, the mean free path in air hovers about 5 m, and at about 0.5 m at 10^{-4} Torr (see Table 7.2). The source-to-wafer distance must be smaller than the mean free path of the residual gas. Typically, this distance is 25–70 cm. With $\lambda > d$, Equation 7.3 approaches 1, i.e., no particles are scattered ($N = N_0$).

On contact with the relatively cool substrate surface, the energy of the impinging atoms is suddenly reduced to the point where they cannot remain in the vapor state; they condense and become attached to the solid surface, forming a deposited thin film. Reaching the substrate, atoms, depending on their energy, still may diffuse laterally over the surface according to:

$$D = D_0 e^{-\frac{E_D}{kT}} \qquad (7.4)$$

Table 7.3 Comparison of Heat Sources for Evaporation

Heat Sources	Advantages	Disadvantages
Resistance	No radiation	Contamination
Electron beam	Low contamination	Radiation
RF	No radiation	Contamination
Laser	No radiation, low contamination	Expensive

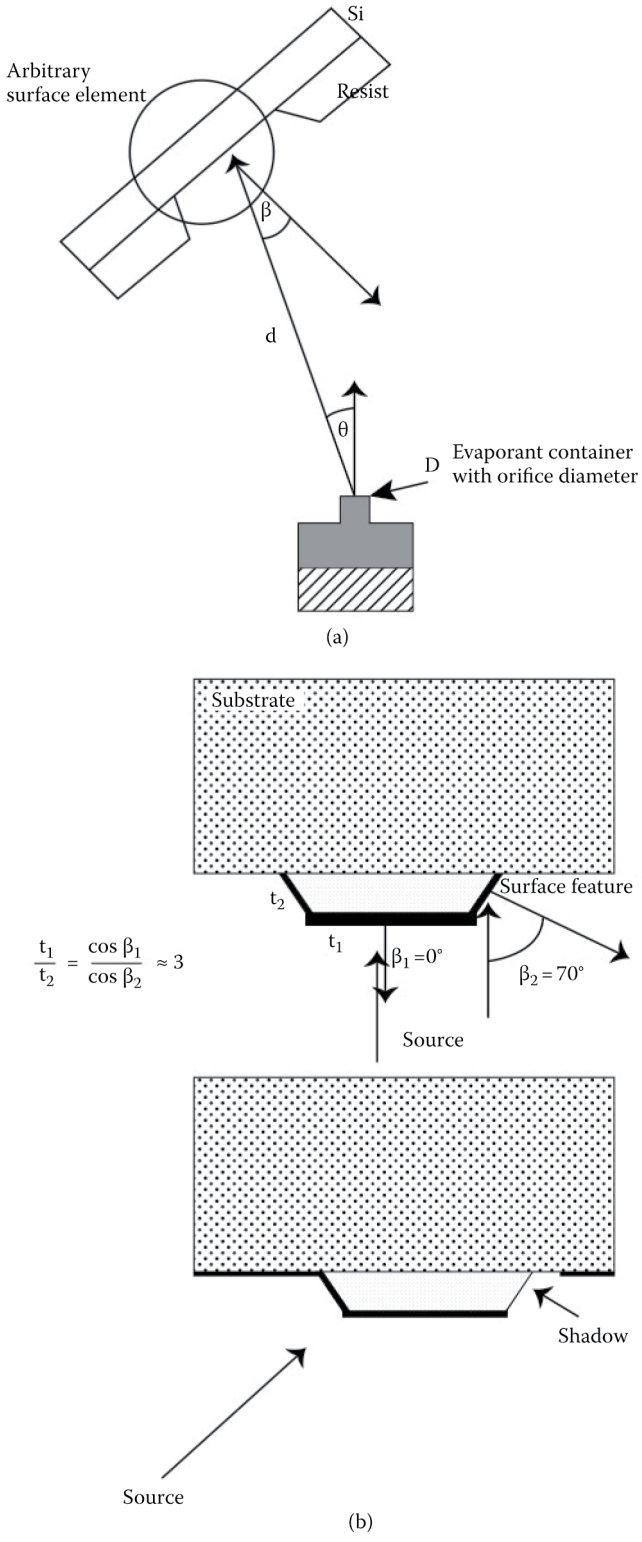


FIGURE 7.3 Geometric considerations in evaporation. (a) Geometric considerations of arrival rate A at an arbitrary surface element in an evaporation experiment. (b) Nonuniform thickness of deposits over varying topography with θ = 0° (top) and θ ≠ 0°C (bottom).

where E_D is the activation energy for diffusion (~2–3 eV) and kT is the energy of the diffusing species. For atoms, after arrival on the surface, to find the best energetic site one must either use energetic atoms or heat the substrate. In a low energy deposition system, such as molecular beam epitaxy (MBE) (see Ion Cluster Beam Deposition—Nanoparticle Deposition section), arriving atoms (~0.1 eV) may form islands unless one picks the right substrate or the substrate is heated to high temperatures. On the other hand, in a high energy deposition system, such as in sputtering (~1 eV), one gets smoother films at lower substrate temperatures, but intermixing may occur.

In a configuration where the evaporant is held in a container with a small hole (i.e., in an *effusion* or *Knudsen cell* as shown in Figures 7.2a and 7.3a) rather than in an open containment structure, the flow through the exit orifice of the evaporant source can range from free molecular to viscous. If the Knudsen number (*Kn*, ratio of the mean free path, λ, to the orifice diameter, *D*) is greater than 1:

$$\mathrm{Kn} = \frac{\lambda}{\mathrm{D}} > 1 \tag{7.5}$$

an atom or molecule passes the orifice in a single, straight track and the flow is free molecular. If, on the other hand, the Knudsen number is <0.01, the flow is viscous, that is, the atom or molecule, before emerging from the orifice, bounces several times into the orifice's sidewalls. A transition region exists between the two.

By kinetic molecular theory the mean free path relates to the total pressure as:

$$\lambda = \left(\frac{\pi \mathrm{RT}}{2\mathrm{M}}\right)^{\frac{1}{2}} \frac{\eta}{\mathrm{P_T}} \tag{7.6}$$

where P_T = pressure at the point of interest
R = ideal gas constant
η = gas viscosity

Low pressure, low molecular weight, and high viscosity translate into a large mean free path. A high vacuum prevents oxidation of, e.g., aluminum and also ensures a purer material deposit with very little

gas inclusions. The arrival rate, *A*, at distance *d* from a small evaporation source follows the cosine law of deposition:

$$A \sim \frac{\cos\beta \cos\theta}{d^2} \tag{7.7}$$

where β = angle between the normal to the substrate and the radial vector joining the source to the arrival point being considered

θ = angle between the normal to the evaporation source and the same radial vector (see Figure 7.3a)

The deposition, according to Equation 7.7, is not spherical; maximum deposition occurs in directions normal to the evaporation source, where $\cos\theta$ is maximum ($\theta = 0$). With the substrate positioned at a distance *d* from the aperture and directly in line with the aperture, the arrival rate, *A*, or total number of molecules per second striking a unit area of substrate, is:

$$A = 3.513\ 10^{22}\ \frac{P_v(T)}{d^2(MT)^{\frac{1}{2}}} \tag{7.8}$$

(See also Equation 7.2 for the evaporation rate *F*.) The straight line or directional deposition of evaporant leads to difficulties in obtaining a continuous coating over topographical steps on a wafer, a problem known as *shadowing*. This is of particular concern in micromachining, where high-aspect-ratio features are common. Figure 7.3b illustrates a typical case. Because the thickness of the deposited film, *t*, is proportional to the $\cos\beta$, the ratio of the film thickness shown in Figure 7.3b, at $\theta = 0°$, is given as:

$$\frac{t_1}{t_2} = \frac{\cos\beta_1}{\cos\beta_2} \tag{7.9}$$

In case of a steep wall, $\beta_2 \sim 70°$ and $\beta_1 \sim 0°$, the ratio of t_1/t_2 is approximately 3. The thinner section, t_2, is susceptible to cracking at the extreme ends of the interval. Two methods can help overcome this problem. The first requires heating of the substrate during deposition to 300–400°C to increase the surface mobility of the metal atoms. The second relies on rotating the wafers in planetary wafer holders so that the angle β varies during deposition (see Figure 7.4).[2] Shadowing prevents the formation of a conformal coat over a step. Mounting the substrates on rotating planetaries aids in making the coat more conformal. A conformal coat is uniform in thickness independent of the underlying topography (see Figure 7.5).

Evaporation is fast (e.g., 0.5 µm/min for Al) and comparatively simple; it registers a low-energy impact on the substrate (~0.1 eV); that is, no surface damage results except when using e-beam evaporation. Under proper experimental conditions, evaporation can provide films of extreme purity and known structure. In cases where the purity of the deposited film is of prime importance, evaporation is the preferred technique. For example, for electrochemical sensors, where the electrocatalytic activity of the top monolayers of the sensing electrode determines proper operation, material purity

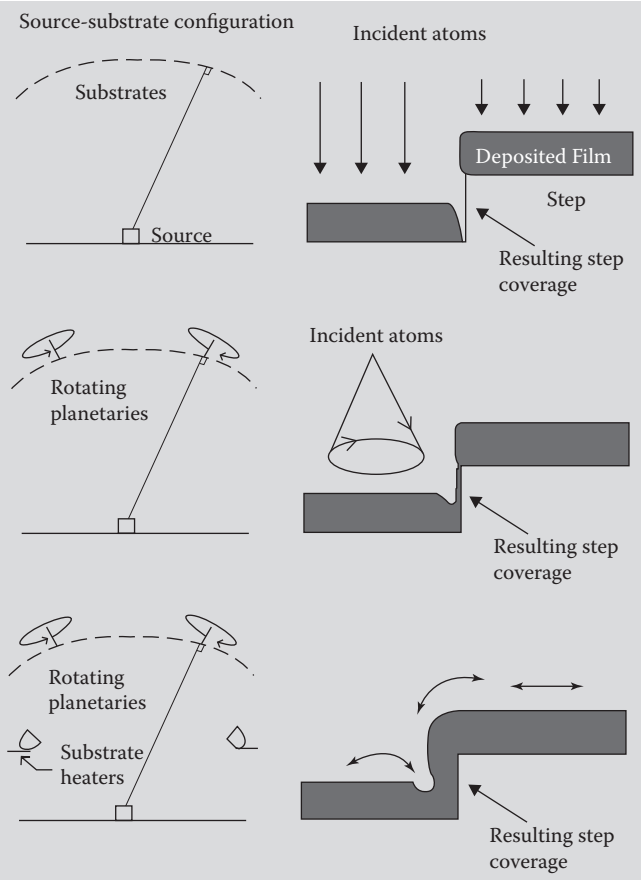


FIGURE 7.4 Shadowing effect may be overcome by using rotating planetaries and heating of the substrate.

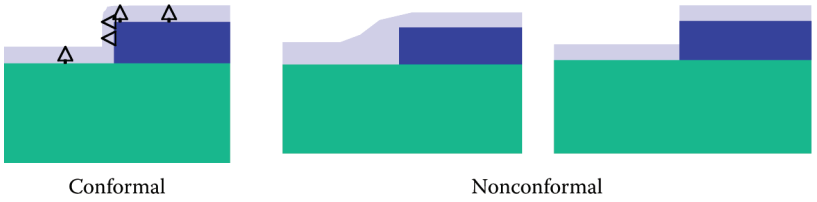


FIGURE 7.5 Illustration of conformal and nonconformal step coverage.

is a priority consideration. In chemical sensors, the surface purity of films is more important than the bulk resistivity (a criterion often used to evaluate IC films). As we just learned, evaporators emit material from a point source (e.g., a small tungsten filament), resulting in shadowing and sometimes causing problems with deposition, especially on high-aspect structures. Difficulties also arise for large areas where highly homogeneous films are required, unless special setups are chosen. A different problem to overcome arises from the source materials decomposing at high evaporation temperatures. Although this risk does not exist when evaporating pure elements, it becomes a problem in the evaporation of compounds and substance mixtures. The e-beam heating systems have an advantage here because only a small part of the metal source evaporates. The initial evaporant stream is richer in the higher-vapor-pressure component; however, the melt depletes that constituent locally, and eventually an equilibrium rate is established. The best way to deposit complex metal alloys is then to e-beam evaporate the metallic elements from different sources because modern quartz-crystal deposition rate monitors enable excellent control in alloy composition. To form oxides of the deposited metals, evaporation is performed in a low-pressure oxygen atmosphere. This process is known as *reactive evaporation*. The oxygen supply comes from a jet directed at the substrate during deposition. To obtain the correct stoichiometry, the deposition needs to take place on a heated substrate.

Evaporated thin films are usually under tensile stress, and the higher the material's melting point, the higher the stress. Tungsten and nickel, for example, can have stress in excess of 500 MPa, which may lead to curling or peeling.[3] Increasing the temperature of the substrate tends to lower the internal stresses in the thin film.

Theoretical film thickness profiles for deposition into an arbitrary-shaped cavity or trench can be calculated using a computer program named SAMPLE, which gives the thin film profile for particular input parameters (http://www.eecs.berkeley.edu/~neureuth).[4,5] In Figure 7.6a we show a SAMPLE result for the line-of-sight deposition in the case of reentrant side-wall angles. The program correctly predicts a discontinuity in the thin film lines in Figure 7.6b. In Figure 7.6c it is shown how, for the case of wafers rotating on a planetary system, step coverage is significantly improved. However, the solution is not perfect because a keyhole effect might result if too much material is deposited (Figure 7.6d).

Sputtering

During sputtering, the target (a disc of the material to be deposited), at a high negative potential, is bombarded with positive argon ions (other inert gases such as Xe can be used as well) created in a plasma (also glow discharge). The target material is sputtered away mainly as neutral atoms by momentum transfer, and ejected surface atoms are deposited (condensed) onto the substrate placed on the anode. The substrate must be placed close to the cathode and is usually heated to improve bonding of the coating atoms. A typical arrangement is shown in Figure 7.7.

Sputtering is preferred over evaporation in many applications because of a wider choice of materials to work with, better step coverage, and better adhesion to the substrate. Whereas vacuum evaporation is generally limited to metals, sputtering can be applied to nearly any material, both metallic and nonmetallic elements, alloys, ceramics, and polymers. Films of chemical compounds can also be deposited by using reactive gases that form oxides, carbides, or nitrides with the sputtered metal. Films of alloys and compounds can be sputtered without changing their chemical compositions. Sputtering

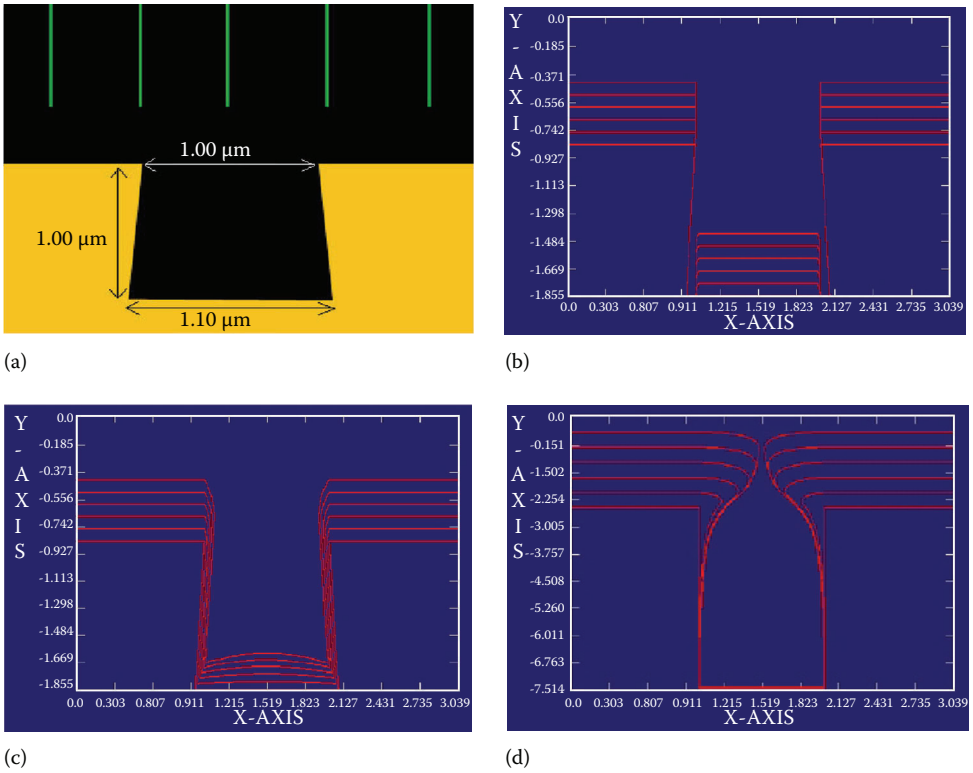


FIGURE 7.6 The SAMPLE computer program gives the thin film profile for particular input parameters (http://www.eecs.berkeley.edu/~neureuth). (a) dimensions of feature to be coated. (b) predicted line-of-sight deposition profile in the case of re-entrant side-wall angles. Notice the discontinuity in the thin film lines. (c) for wafers rotating on a planetary system, step coverage is improved. (d) the solution is not perfect, however, because a keyhole effect might result.

is used in laboratories and production settings, whereas evaporation mainly remains a laboratory technique. Other reasons to choose sputtering over evaporation can be concluded from a comparison of the two techniques in Table 7.4.

Because most aspects pertaining to the physics and chemistries of DC and RF plasmas were discussed in Chapter 3, we will only amplify these topics here as they apply to material deposition. During ion bombardment, the source is not purposely

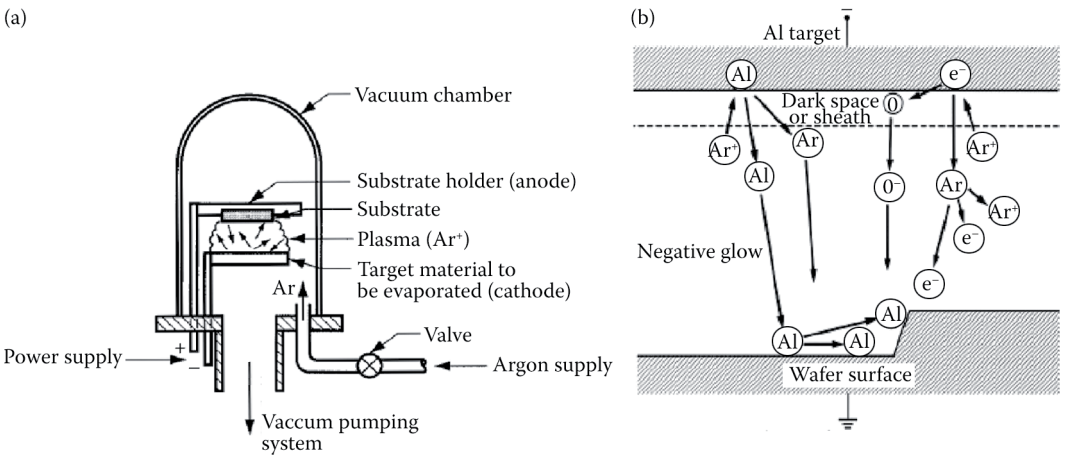


FIGURE 7.7 (a) One possible setup for sputtering, a form of physical vapor deposition. (b) Gas ions are accelerated by a high voltage, producing a glow discharge or plasma. A source (the *cathode*, also called the *target*) is bombarded in high vacuum by gas ions. Atoms from the target (Al in this case) are ejected by momentum transfer and move across the vacuum chamber. Atoms are deposited on the substrate to be coated and form a thin film.

Table 7.4 Comparison of Evaporation and Sputtering Technology

	Evaporation	Sputtering
Rate	Thousand atomic layers per second (e.g., 0.5 μm/min for Al)	One atomic layer per second
Choice of materials	Limited	Almost unlimited
Purity	Better (no gas inclusions, very high vacuum)	Possibility of incorporating impurities (low-medium vacuum range)
Substrate heating	Very low	Unless magnetron is used substrate heating can be substantial
Surface damage	Very low, with e-beam x-ray damage is possible	Ionic bombardment damage
In situ cleaning	Not an option	Easily done with a sputter etch
Alloy compositions, stoichiometry	Little or no control	Alloy composition can be tightly controlled
X-ray damage	Only with e-beam evaporation	Radiation and particle damage is possible
Changes in source material	Easy	Expensive
Decomposition of material	High	Low
Scaling up	Difficult	Good
Uniformity	Difficult	Easy over large areas
Capital equipment	Low cost	More expensive
Vacuum path	High: few collisions, line-of-sight deposition, little gas in film	Low: many collisions, less line-of-sight, gas inclusions
Number of depositions	Only one deposition per charge	Many depositions can be carried out per target
Thickness control	Not easy to control	Several controls possible
Adhesion	Often poor	Excellent
Shadowing effect	Large	Small
Film properties (e.g., grain size and step coverage)	Difficult to control, larger grain size, fewer grain orientations	Control by bias, pressure, substrate heat, smaller grain size, many grain orientations

heated to high temperature, and the vapor pressure of the source is not a consideration as it is in vacuum-evaporation. The amount of material, W, sputtered from the cathode is inversely proportional to the total gas pressure, P_T, and the anode-cathode distance, d[1]:

$$W = \frac{kVi}{P_T d} \tag{7.10}$$

where V = working voltage
i = discharge current
k = proportionality constant

Other energetic particles, such as secondary electrons, secondary ions and photons, and x-rays are created at the target and can be incorporated in the growing film and/or influence its properties through heating, radiation, or chemical reactions. Deposition rate is roughly proportional to sputter yield, S, for a given plasma energy. The sputter yield S, we saw in Chapter 3, stands for the average number of atoms removed per incident ion (S = number ejected/number incident) and is a function of the bombarding species (for example, with a heavier sputtering gas, S increases), the ion energy of the bombarding species, the target material, the incident angle of the bombarding species, and its electronic charge. Target atoms come off with a nonuniform distribution as shown in Figure 3.16, where we plot the sputter yield S as a function of θ, the angle at which ions are directed at the surface (0° is perpendicular and 90° is glancing). Sputtering is seen here to be most efficient at about 20–30° off glancing. In sputtering at energies <50 keV, the incoming ions are modeled as hard spheres colliding with the substrate, with 95% of the incident energy going into the substrate (that is why in sputter deposition cooling of the target is required!) and only 5% of the incident energy carried off by substrate/target atoms. Some yield figures for commonly used metals are tabulated in Table 7.5.[6] Sputtering yield is

Table 7.5 Sputter Yields of Several Commonly Used Metals with 500 eV of Argon

Element	Symbol	Sputter Yield S
Aluminum	Al	1.05
Chrome	Cr	1.18
Gold	Au	2.4
Nickel	Ni	1.33
Platinum	Pt	1.4
Titanium	Ti	0.51

Source: Vossen, J. L., and W. Kern, eds. 1978. *Thin film processes.* Orlando, FL: Academic Press. With permission.[6]

low, usually from less than 1% up to 4% depending on the materials and ion energy level. Because of the low sputtering yield, sputter deposition rates are low (slow process). Another drawback is that because the ions bombarding the surface are embedded in a gas, traces of that gas can usually be found in the coated films; the entrapped gases sometimes adversely affect mechanical properties.

A typical sputter yield as a function of ion energy was shown in Figure 3.15.[1] In the low-energy region, the yield increases rapidly from a threshold energy. The sputter threshold, usually in the range of 10–130 eV, is essentially independent of the bombarding ion species used and equals about four times the activation energy for evaporation. This threshold voltage decreases with increasing atomic number within each group of the periodic table. Above a few hundred volts, a changeover region takes hold. Its value depends on the ion species and the target used, but after its occurrence the sputter yield increases more slowly. Ion energies in the range of 0.5–3 kV are typically used for sputter deposition because nuclear collisions are predominant in this range. In this region, the sputter yields typically range from 0.1–20 atoms per ion, and the yield of most metals is about 1 (see Table 7.5). An ion can only penetrate up to a certain depth into the target and still effect a recoil collision that can reach the surface with sufficient energy to eject atoms (e.g., 1 nm/kV penetration for argon ions in copper). Consequently, at ion bombardment energies in the range of 10 keV to 1 MeV, the sputter yield reaches a maximum and then gradually decreases in value as a result of deep ion implantation. Average ejection energies of ions from the target range between 10 and 100 eV. At those energies, the incident ion can penetrate a substrate one to two atomic layers into the surface on which it lands. As a result, the adhesion of sputtered films is superior to films deposited by other methods.[7] Although sputtering is basically a low-temperature process, considerable amounts of energy are dissipated at the target surface. To liberate one atom requires 100–1000 times the activation energy needed for evaporation, limiting useful sputtering rates to about one atomic layer per second (1000 times less than evaporation from a source operating at 10^{-1} Torr). The target needs water cooling because most of the energy, estimated at about 95%, dissipates into it, and excessive heating might result. Only 1% of the remaining energy goes into sputtering, and the rest is dissipated by secondary electrons that bombard and heat the substrate.

Figure 7.7 exemplifies a typical sputtering setup. Sputtering systems often have a load-lock connected to the sputtering chamber used for loading and unloading the substrates. With the use of a load-lock system, the sputtering station can be kept under continuous high vacuum; the target and chamber can be kept free of contamination from gaseous atmospheric components. Instead of pumping the large sputtering chamber, only the small load-lock chamber is pumped, resulting in a significant shorter cycle time.

To prevent sputtering of the structural elements of the cathode assembly, a shield of metal at anode potential is placed around all the surfaces to be protected at a distance less than that of the dark space at the cathode in Figure 7.8.[7] No discharge will

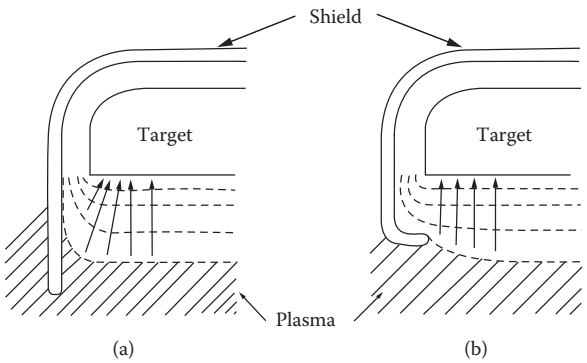


FIGURE 7.8 Sputtering: use of cathode shield. (a) Reducing rim effect by extending cathode shield. (b) Reducing rim effect by wrapping shield around the cathode. (From Maissel, L. I., and R. Glang, eds. 1970. *Handbook of thin film technology.* New York: McGraw-Hill.[7])

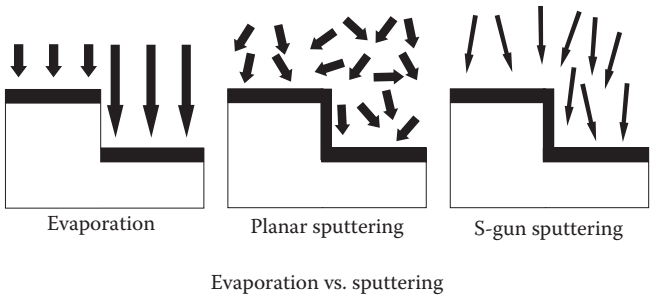


FIGURE 7.9 Sputtering leads to better step coverage than evaporation. (By Dr. Han Xu.)

take place between two surfaces that have a separation less than the Crookes cathode dark space (see Chapter 3).

Crucial to the formation of sustained DC plasma, as discussed in Chapter 3, is the production of ionizing collisions between secondary electrons released from the cathode and the gas in the sputtering chamber. In effect, this requires a relatively high working pressure of greater than 1×10^{-2} Torr. However, if the pressure is too high, significant numbers of the sputtered atoms cannot pass through the sputtering gas and are reflected back to the cathode by collisions. At 10^{-1} Torr, the mean free path of the sputtered metal atoms is about 1 mm, which is near the practical limit. Because of the multiple collisions of the metal atoms in the path between the cathode and anode, the metal atoms arrive at the anode at random incident angles. This leads to a better step coverage compared with evaporation (Figure 7.9). Furthermore, because the sputtering target is very broad compared with an evaporation point source, evidence of shadowing decreases drastically. Sputtering at low gas pressure leads to improved film adhesion because the sputtered atoms have a higher energy. Reduced pressure also reduces contamination of the film by trapped gas molecules, resulting in films of higher density and purity. The higher material density causes a slower chemical etch of the deposited film.

For conductors, a DC sputtering setup can be used. Insulating materials require an RF power supply. The two types of setups are schematically shown in Figure 7.10. Before sputter deposition, the substrate may be sputter etched by connecting it to the negative pole of the power supply. Sputter cleaning further promotes adhesion of subsequent metallizations. In situ cleaning can also be done efficiently with an ion gun.

As in dry etching, the plasma can be confined and densified by using electron cyclotron resonance (ECR) in a planar or cylindrical magnetron or an S-gun using a conical magnetron, which produces higher deposition rates than evaporation produced by electron beam (e-beam) or RF induction. Higher ion densities permit higher sputter rates at lower gas pressures (down to 0.5 mTorr). In a planar magnetron sputtering apparatus, the crossed electric and magnetic fields (magnets of 200 Gauss are put behind the target) contain the electrons and force them into long, helical paths, thus increasing the probability of an ionizing collision with an argon atom (see Figure 7.11). Furthermore, secondary electrons emitted by the cathode as a result of ion bombardment are bent by the crossed fields and are collected by ground shields. This eliminates the secondary electron bombardment of the substrates, which is one of the main sources of unplanned substrate heating. The deposition rate reaches hundreds of angstrom per minute. For some materials such as aluminum, the deposition rate can be as high as 1 μm/min. Magnetron sputter enhancement can be

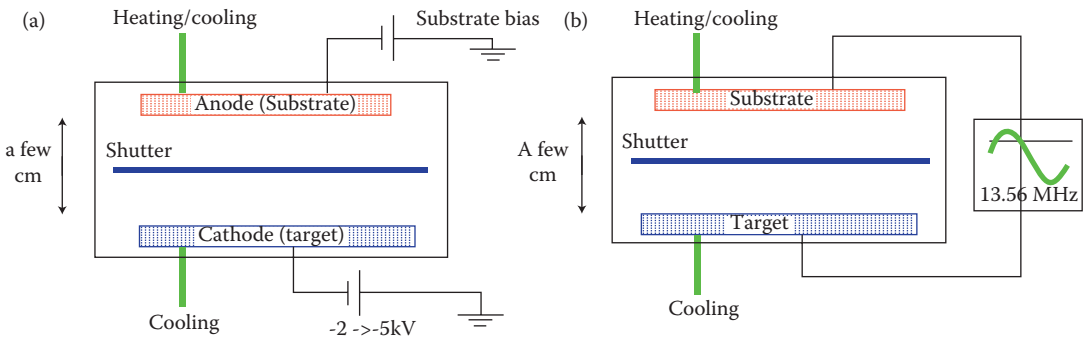


FIGURE 7.10 The DC (a) and RF (b) sputter deposition setups.

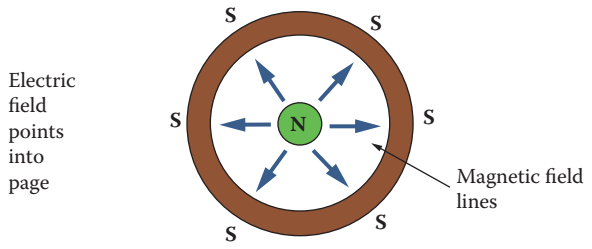


FIGURE 7.11 Magnets behind the target in a magnetron trap electrons near the cathode, causing more ionization near the cathode (10×) with fewer electrons reaching the substrate (less heating).

used in high-frequency sputtering systems, as well as in DC sputtering systems.

The substrate is also often heated to promote film adhesion and reduce film stress. Stress levels vary from compressive at low pressures (0.1–1 Pa) to tensile at high pressures (1–10 Pa). The transition regime between compressive and tensile is often sharp (more than a few tenths of Pa), making the crossover, an ideal point for zero-stress deposition, difficult to control.[3] In general, it is preferable to have ions strike the substrate while growing a film. The plasma ions add energy to the film and keep the surface atoms mobile enough to fill virtually every void. The influence of plasma on a depositing film is considered in more detail in the section on plasma-enhanced chemical vapor deposition (PECVD) further below.

Sputtering often is carried out in the presence of a reactive gas, ensuring control or modification of the properties of the deposited film in so-called *reactive sputtering*. A sensor material produced this way is IrO_X, a low-impedance pH-sensing material arrived at by sputtering from a pure Ir target with argon ions in the presence of oxygen.[8] The exact mechanism for compound formation in reactive sputtering still eludes researchers. It is conjectured that at low pressures the reaction takes place at the substrate as the film is being deposited. At high pressure, the reaction is believed to occur at the cathode with the compound transported to the substrate.

Composite films can also be made by cosputtering or by sputtering from a single composite target. Because the various elements in a target have different sputtering rates and sticking probabilities, the composition of the target might differ somewhat from that on the substrate. However, the decomposition is less significant than in evaporation because no thermal equilibrium is established at the target, and atoms are ejected by momentum transfer. High substrate temperatures are needed to obtain the right composition and to promote adhesion.

TABLE 7.6 Unique Characteristics of Sputter Deposition

Characteristics
Uniform thickness over large area
Simple thickness control (because it is so slow!)
The alloy composition maintains stoichiometry with the original target composition
Deposition rates do not differ a great deal from one material to another
Sputtering cleaning before initiating film deposition is possible; the surface is not again exposed to ambient after such cleaning
The lifetime of a sputtering target may be as long as hundreds of runs and is seldom less than 20

In Table 7.6 we list some of the salient characteristics of sputter deposition.

Nowadays, sputtering equipment applies films to compact discs, computer disks, large area active-matrix liquid crystal displays, and magneto-optic disks.[9] Bearing gears, saw blades, and so on also can be coated with a number of hard, wear-resistant coatings such as TiN, TiC, TiAlN, NbN, CrN, TiNbN, or CrAlN.[10] Commercial in-line sputtering equipment is available, enabling, in a series of evacuated chambers and load locks, depositions of insulators such as SiO_2, Si_3N_4, and Ta_2O_5, metals, and semiconductors such as indium tin oxide (ITO) on substrate with sizes of 830 × 1500 mm. Some example materials deposited by sputter deposition are listed in Table 7.7.

In one particular setup, one can perform continuous sputter deposition (Figure 7.12) on rolls

TABLE 7.7 Example Materials Deposited by Sputtering

Type of Material: Examples
Metals: Al, Cu, Zn, Au, Ni, Cr, W, Mo, Ti
Alloys: Ag-Cu, Pb-Sn, Al-Zn, Ni-Cr
Nonmetals: graphite, MoS_2, WS_2, PTFE
Refractory oxides: Al_2O_3, Cr_2O_3, Al_2O_3-Cr_2O_3, SiO_2, ZrO_2-Y_2O_3
Refractory carbides: TiC, ZrC, HfC, NbC
Refractory nitrides: TiN, Ti_2N, ZrN, HfN, TiN-ZrN, TiN-AlN-ZrN
Refractory borides: TiB_2, ZrB_2, HfB_2, CrB_2, MoB_2
Refractory silicides: $MoSi_2$, WSi_2, Cr_3Si_2

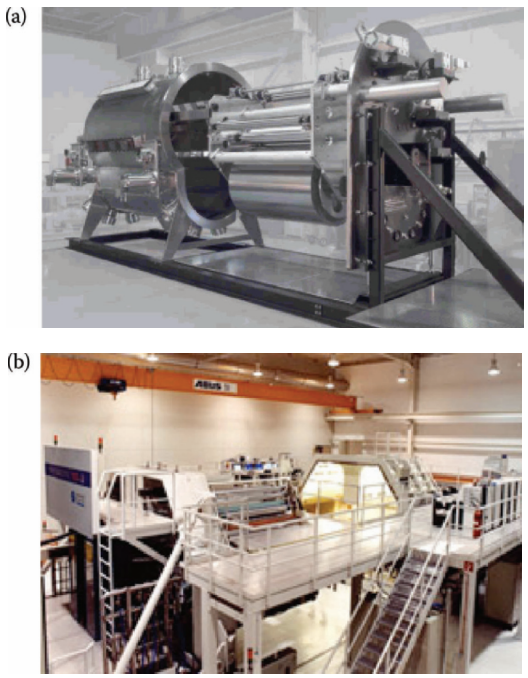


FIGURE 7.12 Continuous deposition. (a) Mill Lane offers systems for continuous vacuum web coating (http://www.mill-lane.com). (b) Innovative Sputtering Technology (I.S.T.), Belgium. Southwall Technologies (http://www.southwall.com) uses two-story high sputtering machines to manufacture transparent sputtered films on 2 m wide × 5,000 m long rolls of polyethylene terephthalate (PET) substrate that is 25 μm or 50 μm thick. One roll allows fabrication of up to 10,000 square meters of film.

of substrate, including foils and fabrics. In a high-vacuum chamber, the rolled substrate is unwound and passes over the cathode cylinder where thin layers of target material are deposited. The substrate is continuously rewound on the other side of the deposition site. The sputter target is a long cylinder, and reactive and nonreactive processes are available.[11] Roll-to-roll (R2R) manufacturing is currently used to print, for example, newspapers and labels. One envisions that many other products could be manufactured the same way. For example, organic light-emitting diodes (OLEDs) could be manufactured by patterning all layers with continuous deposition processes (web vs. batch). This would be potentially low cost and be scalable to large areas (see also Volume III, Chapter 1). Innovations like these might help revive the potential of micromachined chemical sensors for which Si batch processes are often too expensive. When manufacturing microfabricated chemical sensors, one often competes with continuous fabrication processes (e.g., continuous printing of glucose sensor paper strips), and classical Si micromachining cannot beat the cost advantages these offer. A process as described here, complemented by continuous lithography, might enable chemical microsensor fabrication at an acceptable cost (see Volume III, Chapter 1). A good reference on continuous, roll-to-roll thin film coating is *Vacuum deposition onto webs, films, and foils* by Charles A. Bishop (William Andrew Publishing, 2007).

The most negative aspects of sputtering are its complexity compared with an evaporation process, the excessive substrate heating resulting from secondary electron bombardment, and, finally, the slow deposition rate. In a regular sputtering process, the rate is one atomic layer per second versus thousand atomic layers per second available from a typical evaporation source at a vapor pressure of 10^{-1} Torr. Controls in a sputtering setup include argon pressure, flow rate, substrate temperature, sputter power, bias voltage, and electrode distance.

Example 7.1: Sputtered Polysilicon

The quest for low-temperature polysilicon deposition processes is driven in part by the need for compatibility with prefabricated aluminum-metallized complementary metal oxide semiconductor (CMOS) circuitry. This makes the 350°C sputter deposition of polysilicon especially interesting.

Honer et al.[12] introduced sputtered polysilicon to the MEMS community. Sputtered silicon can be used to fabricate polysilicon microstructures (the key mechanical structural element in surface micromachining; see section on Doped Polysilicon) atop standard, aluminum-metallized CMOS at temperatures less than 350°C. A drawback for sputtered boron doped Si films is that to increase their conductivity (to 25 Ω/□) they need to be clad in symmetric, 50-nm-thick layers of titanium-tungsten. Films with stress values less than 100 MPa are routinely obtained.

An important benefit of this low-temperature polysilicon process is that it is not only compatible with conventional oxide sacrificial layers (sacrificial layers, we shall see below, are another key ingredient in surface micromachining) but also with certain organic sacrificial layers (e.g., polyimide). Organic layers can be removed in a dry

oxygen plasma etch, thus avoiding the stiction and selectivity problems associated with wet etch releases and at the same time alleviating the concerns over chemical attack on structural elements by HF.

Ion Plating

In ion plating, evaporation of a material is combined with ionization of the atom flux by an electron filament or plasma. The principle of ion-plating in argon plasma is illustrated in Figure 7.13.[13] The substrate is set up to be the cathode in the upper part of the chamber, and the source material is placed below it. A vacuum is then established in the chamber. Argon gas is admitted, and an electric field is applied to ionize the gas (forming Ar^+) and establish plasma. This results in ion bombardment (sputtering) of the substrate so that its surface is atomic clean (interpret this as "very clean"). Next, the source material (Ti in the example case) is heated sufficiently to generate coating vapors. The heating methods used here are the same as those used in vacuum evaporation: resistance heating, e-beam bombardment, and so on. The Ti vapor passes through the plasma and coats the substrate. Sputtering is continued during deposition; thus, the ion bombardment of the substrate consists not only of the original argon ions but also of source material ions (Ti) that have been energized while being subjected to the same energy field as the argon. The effect of these processing conditions is to produce films of uniform thickness and excellent adherence to the substrate. Ion plating is applicable to parts with irregular geometries because of the high degree of scattering of the depositing material in the plasma. An example of interest here is TiN coating of high-speed cutting tools (for example, drill bits). As shown in Figure 7.13, the addition of a gas (nitrogen in this case) to the reactor enables one to make new compounds (such as TiN) on the substrate surface, with the gas reacting with the ionized atoms from the evaporation source (Ti). Because of the high kinetic energy of the impacting ions, a very well-adhering, dense TiN film with extraordinarily low friction coefficient and high hardness coefficient (Vickers hardness of 50,000) forms. Because of the thermal nature of the process, very high deposition rates can be achieved. In addition to coating uniformity and good adherence, other advantages of the process include high deposition rates, high film densities, and the capability to coat the inside walls of holes and other hollow shapes.

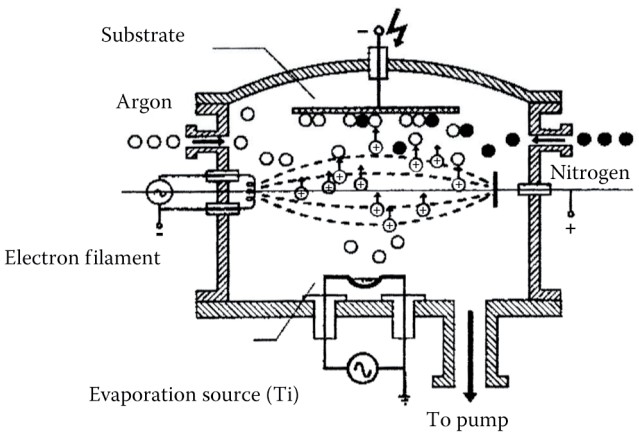


FIGURE 7.13 Principle of ion plating. (From Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure.* Weinheim, Germany: VCH Publishers.[13])

The NTH-1000 from Nanotec Corp. (http://nanotec-jp.com/www_nanotec/english/nth1000.html) uses a tandem hollow cathode discharge technique to deposit high-quality TiN, CrN, TiCN, and TiCrN thin films on cutting tools, machine parts, and molds.

Ion Cluster Beam Deposition—Nanoparticle Deposition

Clusters are aggregations of atoms or molecules, intermediate in size between individual atoms and aggregates that are large enough to be called *bulk matter.* Atom clusters are also called *nanoparticles* because their size is on the order of nanometers or tens of nanometers. The ultrasmall size of building blocks leads to dramatically different properties, and it is anticipated that such atomically engineered materials can be tailored to perform as no previous material could. T. Takagi, from Kyoto University, first proposed the idea of ionized cluster beam (ICB) thin film deposition in 1972. Conditions for formation of ion cluster beams, suitable for thin film deposition, were only established following an additional 20 years of effort. When applying cluster beam technology, ionized atom clusters (e.g., 100–1000 atoms) are deposited on a substrate in a high vacuum (10^{-5}–10^{-7} mbar). Those atom clusters typically

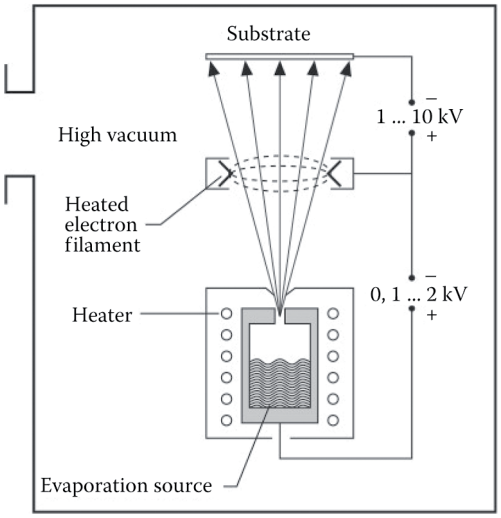


FIGURE 7.14 Setup for ion-cluster beam deposition. (From Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure.* Weinheim, Germany: VCH Publishers.[13])

carry one elementary charge per cluster and therefore achieve the same energy in an electrical field as would a single ion.

A special evaporation cell must be used to generate these atom clusters. As shown in Figure 7.14,[13] the heating of the evaporant in an evaporation cell with a small opening causes an adiabatic expansion of more than 100–10^{-5} or 10^{-7} mbar on exiting that cell. The expansion causes a sudden cooling, inducing the formation of atom clusters. These clusters are then partially ionized by an electron bombardment from a heated filament. Low-energy neutral clusters (0.1 eV) and somewhat higher energy ionized clusters (a few eV) arrive at the surface where they flatten and form a film (Figure 7.15)[13] of excellent adhesion and purity, with a relatively low number of defects. Cluster beam epitaxy is possible at temperatures as low as 250°C, and no charge buildup occurs when depositing on an insulator. Ion cluster beam technology qualifies as a gas-phase nanoparticle synthesis method. In gas-phase synthesis of nanoparticles supersaturation is achieved by vaporizing material into a background gas. Condensation to form particles is induced by cooling, e.g., by mixing with a cool gas or expansion through a nozzle. In ion cluster beam technology the cooling occurs when the vapor exits the evaporation cell. Other types of gas-phase nanoparticle fabrication techniques are discussed in Chapter 9 on thermal energy-based forming techniques.

The cluster beam technology is also applied to shallow ion implantation, high-yield sputtering, smoothing, surface cleaning, and low-temperature thin film formation. Cluster ion beam processing is now expanding into new industrial fields presently limited by the available atomic and molecular ion beam processes. With the wide variety of selective layers needed in chemical sensors and the frequent need for low-temperature deposition of thick, well-adhering layers, ion plating, laser sputter deposition (see below), and cluster beam technology seem destined to play a more crucial role in chemical sensor development in the future. Also, building devices from atomic constituents (bottom-up approach), with proximal probes for example, is too time consuming when performed one atom at the time. To speed up the process, massive parallelism is required; for example, by using an army of coordinated proximal probes. Cluster beam technology better enables nanomachining because the size of the building blocks lends itself to being built faster.

Epion (http://www.epion.com) is an example of a company selling ion cluster beam equipment, and their units feature surface smoothing to less than 3 Å, surface cleaning, and high-yield sputtering.

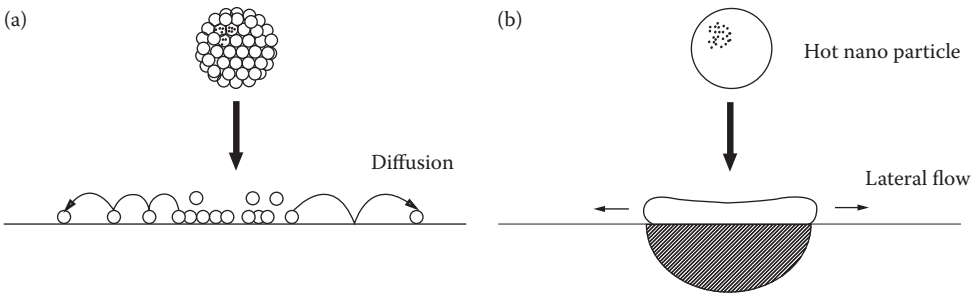


FIGURE 7.15 Film forming with atom clusters. (a) Old model based on atom diffusion. (b) New model based on particle flow. (From Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure.* Weinheim, Germany: VCH Publishers.[13])

Pulsed Laser Deposition, Laser Sputtering, or Laser Ablation Deposition

Pulsed laser deposition (PLD) uses intense laser radiation to erode a target and deposit the eroded material onto a substrate. A high-energy focused laser beam avoids the x-ray damage to the substrate encountered with e-beam evaporation. A high-energy UV excimer laser pulse (10 ns) coming from, for example, a KrF laser at 248 nm with a pulse energy in the focus of 2 J/cm^2, is directed onto the material to be deposited, as sketched in Figure 7.16. As shown here, the laser ablates material from the target, forming a high-temperature vapor plume incident on the sample. The target may be just about anything, and one can operate at atmospheric pressures or in a very high vacuum.

The PLD process can be split into the following basic steps:

1. The energy of the very short wavelength radiation is absorbed in the upper surface of the target, i.e., absorption of the laser pulse in the material (for a metal, the absorption depth is ~10 nm, depending on λ).
2. Energy relaxation to the lattice through electron-phonon interaction (~1 ps).
3. Heat diffusion (microseconds), melting (tens of nanoseconds), and evaporation of a small amount of material.
4. Plasma creation.
5. Absorption of light by plasma and ionization.
6. Interaction of target and ablated species with plasma.
7. Cooling and resolidification between pulses.

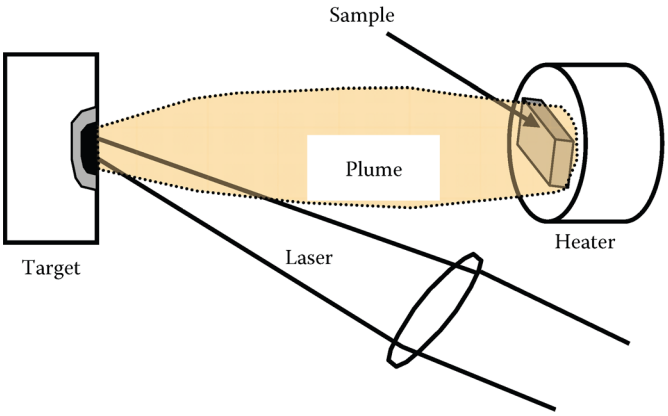


FIGURE 7.16 Schematic of pulsed laser deposition (PLD).

Laser ablation is a very nonequilibrium process. At the peak of the laser pulse, temperatures on the target can reach >10^5 K (>40 eV!) with electric fields >10^5 V/cm (accompanied by high magnetic fields) and plasma temperatures of 3000–5000 K. The ablated species come with energies in the 1–100-eV range.

Laser ablation deposition was first mentioned in Chapter 5, under thermal removing techniques, where we learned that laser subtractive machining leaves an undesirable heat-affected zone (HAZ) on a surface that is being machined. The size of the HAZ is a function of the laser pulse duration and the thermal conductivity and specific heat of the work piece. The HAZ constitutes a problem in the case of laser subtractive machining because it degrades the quality of the cut. Moreover, during the ablation process, the expulsion of material in the plasma jet redeposits material onto the surrounding region, further compromising the fidelity of the process. In the case of PLD, we take advantage of this deposition of materials from the plume onto an adjacent substrate surface.

Splashing is the bane of PLD. The laser superheats the subsurface layer before the surface reaches its evaporation point, and, as a consequence, the surface breaks apart into large (micrometer-sized) globular particles when the subsurface expands. This leads to globular deposits. Another problem with PLD is the expansion of the plume, which causes a sudden decrease in the pressure just above the surface. This shock wave pulls droplets of liquid of the surface. The thermal shock also causes irregularities in the surface to break off (exfoliation). Advantages and disadvantages of PLD are summarized in Table 7.8.

As we saw in Chapter 5 with ultrashort laser pulses (<1 ps), so much energy is deposited in the material at such a fast rate that the material is forced directly into a plasma state; the material goes from a solid to a gas phase without first going through a melt phase. The reason behind this is that pulse width is now shorter than the electron lattice-relaxation time; thus, the heat diffusion and melting are significantly reduced. This plasma then expands away from the material as a highly energetic gas, taking almost all the heat away with it. This leads

TABLE 7.8 Advantages and Disadvantages of Pulsed Laser Deposition (PLD)

Advantages	Disadvantages
Flexible, easy to implement	Uneven coverage
Growth in any environment	High defect or particulate concentration
Exact transfer of complicated materials (YBCO)	Not well suited for large-scale film growth
Variable growth rate	Mechanisms and dependence on parameters not well understood
Epitaxy at low temperature	Uneven coverage
Resonant interactions possible (i.e., plasmons in metals, absorption peaks in dielectrics and semiconductors)	Splashing as a result of subsurface boiling, shock wave recoil, exfoliation
Atoms arrive in bunches, allowing for much more controlled deposition	
Greater control of growth (e.g., by varying laser parameters)	

Source: Based on *Pulsed laser deposition of thin films.* New York: John Wiley and Sons, 1994.[230]
The table does not take into account pulsed laser deposition with ultrafast pulses improving the technique a lot, for example, smoother films result.

not only to cleaner cutting but also smoother film deposition (fewer particulates, less splashing) and the potential for ultrahigh quality films, circuit writing, isotope enrichment, new materials, and nanoparticle production.

The ablated material, partially ionized in the laser-induced plasma, is deposited onto a substrate almost without decomposition. This technique is particularly useful when dealing with complex compounds, as in the case of the deposition of high-temperature superconductor films (HTSCs), for example, $YBa_2Cu_3O_{7-x}$. PLD faithfully replicates the atomic ratios present in the hot isostatically pressed target disc onto the thin film coating. Achieving complex stoichiometries presents more difficulties with any other deposition technology. Approximately 10,000 pulses (pulse length of 20 ns and a repetition rate of 15 pulses/s) are needed to achieve a film thickness of 0.1 µm on the substrate. Normally, the laser-deposited films are amorphous. The energy necessary to crystallize the film comes from heating the substrate 700–900°C and from the energy transferred from the intense laser beam to the substrate via atomic clusters.

Figure 7.17a shows a schematic setup of a laser deposition system.[14] Here, a pulsed excimer laser is used to deposit superconductor materials such as those based on YbaCuO.[14] Figure 7.17b shows an improved laser deposition setup where substrate heating can be replaced in part or completely by

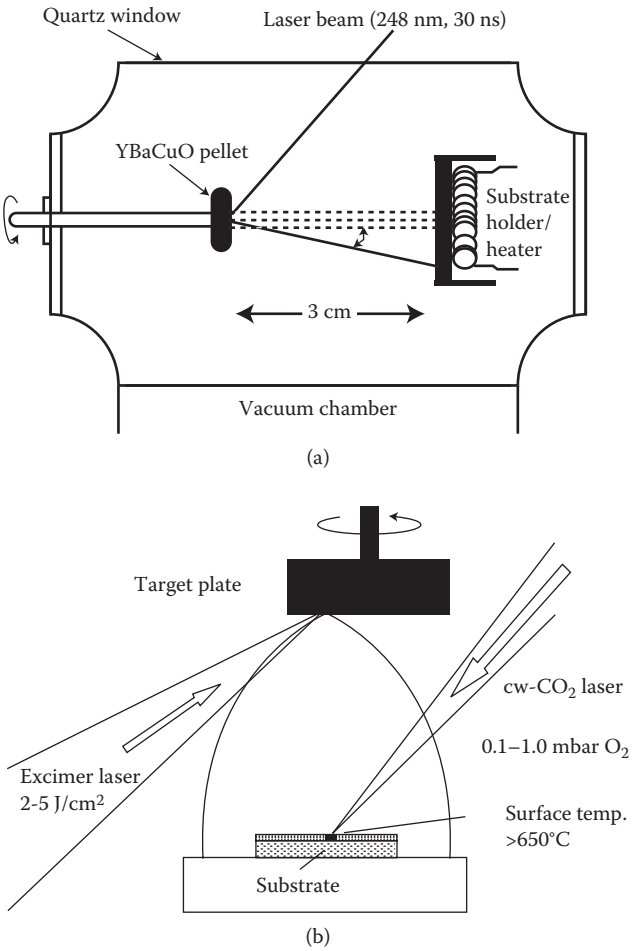


FIGURE 7.17 Laser deposition system. (a) Traditional laser ablation system. (From Dutta, B., X. D. Wu, A. Inam, and T. Venkatesan. 1989. Pulsed laser deposition: a viable process for superconducting thin films? *Solid State Technol* February:106–10. With permission.[14]) (b) Laser ablation with additional laser for surface heating.

additional laser radiation, in this case, a continuous wave-CO_2 (cw-CO_2) infrared laser. With this setup, using two crossed laser beams, it is possible to deposit and induce the correct crystalline structure in the growing HTSC film while it is being deposited.[15] Superconductive Components Inc. manufactures superconducting and nonsuperconducting ceramic targets for use in sputtering and laser ablation systems (http://www.superconductivecomp.com/targets2.html).

A sensor-related application of PLD is the protection of components in contact with body fluids using the biocompatible calcium-phosphate-based ceramic calcium hydroxylapatite, or $Ca_{10}(PO_4)_6(OH)_2$. This material exemplifies the most stable calcium phosphate in contact with body fluids. Deposition methods of this material include sputtering, plasma spraying, electrophoretic deposition, and combinations of these techniques. In all cases, a postdeposition treatment is needed to crystallize the partially or wholly amorphous and/or dehydrated films. With the laser pulse technique, in water-vapor-enriched inert gas environments, deposition of the right hydroxyl apatite was observed at temperatures between 400 and 700°C. Adhesion of this material to Si and Ti-6Al-4V was found to be excellent.[16] Laser sputtering also could help prepare complex and stoichiometry-sensitive coatings such as electrochromic devices.

Pulsed laser ablation, like ion cluster beam deposition, can also be used as a nanoparticle gas-phase synthesis method. In this approach the solid precursor is heated by absorption of laser energy. Typically an infrared laser (CO_2) is used, and laser energy may be directly absorbed by precursors or by an inert photosensitizer (SF_6). The use of a pulsed laser shortens the reaction time and allows for the preparation of the smallest particles. In 1995, at Rice University it was reported that synthesis of carbon nanotubes could be accomplished by laser ablation.[17] A pulsed or continuous laser is used to vaporize a graphite target placed in an oven at 1200°C (Figure 7.18). The oven is filled with argon gas to keep the pressure at 500 Torr. A very hot vapor plume forms, which then expands and cools rapidly. Contact with a cooled copper substrate causes the carbon atoms to be deposited in the form of nanotubes.

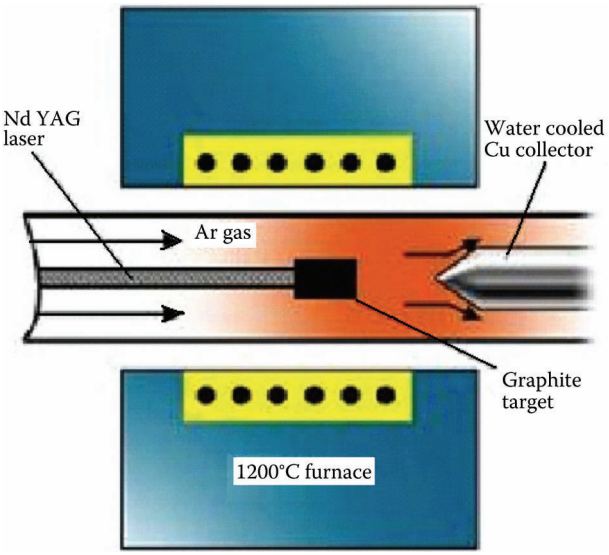


FIGURE 7.18 Laser ablation system for carbon nanotube production.

Laser vaporization produces a higher yield of single-walled nanotubes (SWNTs) with better properties and with a narrower size distribution than nanotubes produced by arc discharge. Laser ablation also produces nanotubes that are more pure (up to 90% purity) than those produced by arc discharge.

Pulsed laser deposition (PLD) with ultrafast pulses (<1 ps) is opening up all types of new research and application opportunities as evidenced by the Defense Advanced Research Projects Agency's (DARPA's) mesoscopic integrated conformal electronics (MICE) program. The goal of that program is to create electronic circuits and materials on any surface, e.g., to print electrical circuits on the frames of eyeglasses or interwoven with clothing (see Figure 7.19). MICE efforts have demonstrated the ability to print metal lines on curved surfaces, feature sizes as small as 5 μm, and print speeds close to 1 m/s. Other exciting developments include the demonstration of 1) printed zinc-air batteries that have four times more volumetric power density than commercial batteries, 2) direct-write antennas, and 3) direct-write solar cells. A single laser does surface pretreatment, spatially selective material deposition, surface annealing, component trimming, ablative micromachining, dicing, and via drilling.

To sum up, laser ablation is a good technique for preparing thin films of any desired stoichiometry,

FIGURE 7.19 Direct writing in air with pulsed laser deposition (PLD).

but, because of the small source size, it is not useful for large-scale coatings. A small source size might make it essential, especially when working with high-aspect-ratio micromachined structures, to rotate the sample at an angle with respect to the material flux so as to obtain coverage of the vertical features in the device. The possibility of using PLD for nanoparticles and carbon nanotubes and the use of PLD with ultrashort laser pulses (<1 ps) has given the technique an added impetus. For further reading on laser sputtering, refer to references 14–16,18,19.

Other good references are found in "Pulsed laser vaporization and deposition," Wilmott and Huber, *Reviews of Modern Physics*, vol. 72, 315 (2000), *Pulsed laser deposition of thin films*, Chrisey and Hubler (Wiley, New York, 1994), and *Laser ablation and desorption*, Miller and Haglund (Academic Press, San Diego, 1998). Epion, for example, sells large-area PLD systems (http://www.pvdproducts.com/products/pld_systems.aspx).

Molecular Beam Epitaxy

Epitaxial techniques arrange atoms in single-crystal fashion on a crystalline substrate acting as a seed crystal so that the lattice of the newly grown film duplicates that of the substrate. If the film is of the same material as the substrate, the process is called *homoepitaxy*, *epitaxy*, or simply *epi*. Epi deposition represents one of the cornerstone techniques for building micromachines. Of special importance is that Si membranes of a predetermined thickness and doping level can be fashioned. The growth rate of an epilayer depends on the substrate crystal orientation. Si (111) planes have the highest density of atoms on the surface, and the film grows most easily on these planes. Important epi applications are Si on Si substrates and GaAs on GaAs substrates. If the deposit is made on a chemically different substrate, usually of closely matched lattice spacing and thermal expansion, the process is termed *heteroepitaxy*. One important heteroepitaxy application is the deposition of silicon on insulator (SOI), for example, Si on SiO_2 or Si on sapphire (Al_2O_3). Various sapphire orientations such as $<01\bar{1}2>$, $<10\bar{1}\,2>$, and $<1\,\bar{1}02>$ have been used to grow <100>-oriented silicon layers. The lattice mismatch between sapphire and silicon crystals limits the thickness of the Si to about 1 μm. Another example of heteroepitaxy is that of gallium phosphide on gallium arsenide.

Various types of epitaxy techniques exist. Chemical vapor phase and liquid epitaxy are described further below. Here, we briefly discuss PVD epitaxy, that is, molecular beam epitaxy (MBE), invented in late 1960s at Bell Laboratories by J. R. Arthur and A. Y. Cho. In MBE, the heated single-crystal sample (e.g., between 400 and 800°C) is placed in an ultrahigh vacuum (10^{-11} Torr) in the path of streams of atoms from heated cells that contain the materials of interest. These atomic streams impinge, in a line-of-sight fashion, on the surface-creating layers with a structure controlled by the crystal

structure of the surface, the thermodynamics of the constituents, and the sample temperature. This technique is the most sophisticated form of PVD. The deposition rate of MBE is very low (i.e., about 1 μm/h or 1 monolayer/s), and considerable attention is devoted to in situ material characterization to obtain high-purity epitaxial layers. Once on the substrate, the atoms or molecules move around until they find an atomic site to chemically bond to. Fast-acting shutters control the deposition. One or two atomic layers of material lie between the shutter action. This becomes important when an ultrasharp profile is called for. The low deposition rate gives the operator better control over the film thickness. Ultrasharp profiles are needed, for example, when making quantum well devices. A quantum well might be 40 Å thick. To uniquely define its energy levels it must be 40 ± 2 Å. Figure 7.20 represents a schematic of a molecular beam deposition setup.[1]

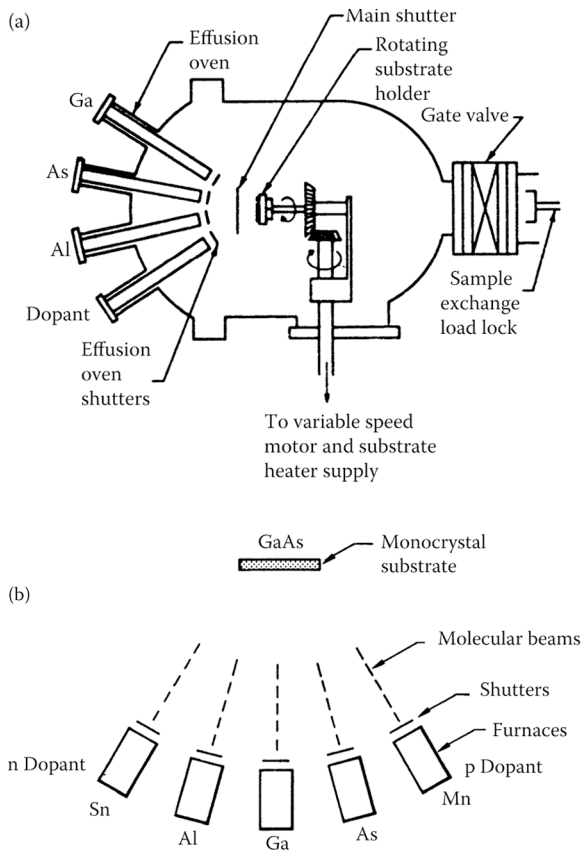


FIGURE 7.20 (a) Schematic of molecular beam epitaxy growth chamber. (b) Example is the growth of GaAlAs epi on GaAs single crystal. (From Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press. With permission.[1])

TABLE 7.9 Key Points in MBE

Characteristics
Ultrahigh vacuum
High purity layers with good crystal structure
No chemical byproducts created at growth surface
High uniformity (<1% deviation)
Often uses multiples sources to grow alloy films
High control of composition
Growth rates are very low (0.1–10 μm/h)
In situ monitoring and feedback
Expensive
Mature production technology

The technique has several potential advantages over CVD epitaxy (see next section): for example, the relatively low growth temperatures reduce diffusion and autodoping effects. Precise control of layer thickness and doping profile on an atomic layer level is possible. Novel structures such as quantum devices, silicon/insulator/metal sandwiches, and superlattices can be made. Today, half of the lasers that are used in CD players, as well as those in pen-sized laser pointers, are grown by MBE. Figure 7.20 shows the deposition of n- and p-type GaAlAs on a GaAs single-crystal surface.[20] With MBE, virtually any device structure can be made. The limitations to consider lie in volume manufacturing and cost. The ultrahigh vacuum requirements make operation very expensive. From the available epi techniques, MBE exemplifies one that is perhaps the most complicated.[21] In Table 7.9 we review the key points concerning MBE, and in Figure 7.21 we show a typical commercial MBE station.

FIGURE 7.21 Typical MBE system photo.

Aerosol Deposition—Impact Deposition

Aerosol deposition (AD) is a mechanical film fabrication method invented in 1997 by Dr. Jun Akedo of the National Institute of Advanced Industrial Science and Technology (AIST) in Japan.[22] In Figure 7.22 we illustrate the overall configuration for film fabrication using AD, together with electron micrographs of raw powder and a resulting AD film.

Raw powders with an average powder diameter of 600 nm are mixed and vibrated with a carrier gas to form an aerosol. By using the pressure difference between the aerosol and a vacuum chamber, the aerosol is sprayed from a nozzle and impacted onto a substrate. Using this simple process, AD films with ultrafine crystalline particles of several tens of nanometers can be formed as shown in the cross-sectional image of the AD film in Figure 7.22. The impact of solid-state particles creates a strongly adherent, high-density nanocrystalline film by gas blasting the nanosized particles onto the surface. The deposition rate is 30 times faster (10 μm/min) than traditional deposition rates and can be carried out even at room temperature. The formation mechanism is assumed to be room-temperature shock-compaction. The impacting 600-nm particles are deformed and fractured into nanocrystalline structures of 10–30-nm size. Uniform deposition rates over a 200-mm square area have been achieved. The raw powders may be manufactured by conventional processing methods for most materials.

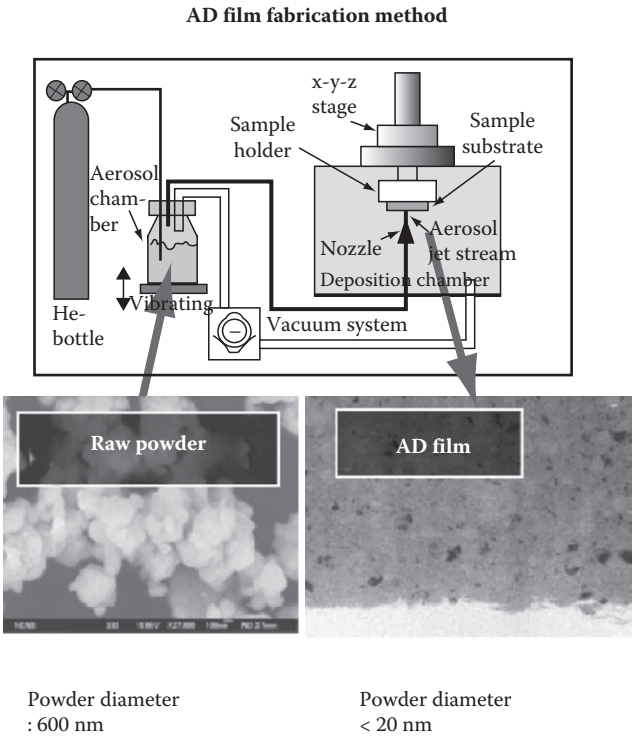


FIGURE 7.22 Schematic illustration of aerosol deposition (AD), powder, and a resulting AD film [solid solution of PNN (lead nickel niobate) and PZT (lead zirconate titanate)]. (http://www.nec.co.jp/techrep/en/journal/g07/n01/070115.html)

The two most important distinguishing characteristics of the AD process are that 1) the deposited films maintain the crystalline properties of the raw powder even at low deposition temperatures, and 2) relatively thick films can be produced fast. This stands in stark contrast to the thin film-forming methods such as sputtering, molecular beam epitaxy (MBE), and chemical vapor deposition (CVD). When these latter methods are used to deposit a material at low temperatures, for example, a piezoelectric material (typically an oxide with a perovskite crystalline structure) onto a substrate such as glass, an amorphous layer results, and all piezoelectric properties are lost. To maintain the piezoelectric function of this material, it is essential to retain the perovskite crystalline structure. To achieve this, one relies on the surface energy of the substrate and a temperature high enough for the atoms to be rearranged to form the desired perovskite oxide film. In this type of deposition process, restructuring the particles in the crystallization process occurs at the atomic level and far from equilibrium, inducing many lattice defects that result again in the degradation of the piezoelectric effect.

For many piezoelectric actuator applications, a film, several micrometers thick, is essential to output sufficient mechanical energy for a given electrical input. When a thick film of more than several micrometers is formed onto a substrate with the conventional thin film-forming technologies, various problems often occur: the speed of film forming is not fast enough, composition tends to vary, too much heat is required to crystallize the film limiting the choice of substrates, and so on. Because of these concerns, the conventional thin film-forming technologies have been used only for limited applications, such as for thin film sensors. In case thicker films are required, bulk-type piezoelectric ceramics

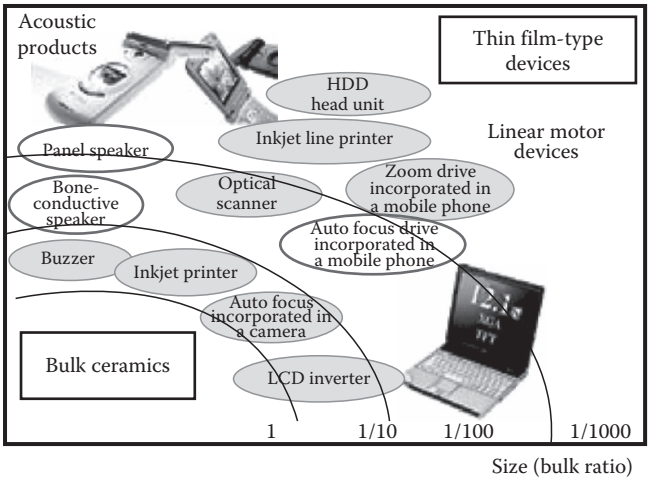


FIGURE 7.23 Use of piezoelectrics in consumer products. (http://www.nec.co.jp/techrep/en/journal/g07/n01/070115.html.)

made of sintered powders have usually been adopted (see Figure 7.23).

Screen-printing of piezoelectrics is a thick-film technique that, in principle, could cover the meso-range between bulk ceramics and thin films. In screen-printing one uses a mixture of ceramic powder and binder to make an ink. After screen-printing, the ink is fired at 1000°C or more (screen-printing of ceramics is a thermal additive technique covered in Chapter 9). This limits this technology to ceramic substrate materials.

As suggested in Figure 7.23, thick film-formed piezoelectric devices integrated onto a substrate will be essential for many piezoelectric actuators of the future, such as in sound equipment and for products that require a high-precision positioning motion. The AD method is a film fabrication method that can fulfill that need. Piezoelectric AD results in nanocrystalline films that preserve a large dielectric and piezoelectric effect and that are little influenced by the underlying layers and/or substrates. For example, by using ceramic particles that feature chemically unstable surface characteristics, AD technology has been used to form ceramic films at room temperature on resin circuit boards to which ceramic is usually difficult to adhere. The maximum dielectric constant of the newly created ceramic film was 400, i.e. 10 times the dielectric constant of ceramic combined with resin, and is approximately at the same level as that of ceramic film made with the sputtering method that uses heat treating at a high temperature of 600°C. A condenser composed of three layers was also successfully created on an FR4 printed circuit board. The capacitance of the newly created condenser was 300 nF/cm^2, a high level that is ideal for practical use (http://www.nec.co.jp/techrep/en/journal/g07/n01/070115.html).

From the above, the AD method is an excellent candidate for the fabrication of piezoelectric thick film actuators for the manufacture of micropumps, ultrasonic mixers in microelectromechanical systems (MEMS) and micrototal analysis systems (μ-TAS), micromanipulators for medical applications, ink-jet printer heads, flapper actuators for high-density hard disk drives, and other applications where large strains and high-speed response are necessary (see Figure 7.23).

An AD of a solid solution of PNN (lead, nickel, and niobate) and PZT (lead zirconate titanate) was attempted on zirconate ceramics and stainless steel substrates. Zirconate ceramics have good high-temperature resistance and stainless steel features, superior corrosion resistance, and processability that lead to low-cost manufacturing of practical piezoelectric devices. The AD film-forming process was carried out at room temperature with a deposition rate of 10 μm/min or higher over an area of 10×10 mm^2. These process conditions confirm that this process is amenable to mass production. Figure 7.24 shows a comparison of the film properties between the piezoelectric film formed by the AD and the conventional film deposition method. The piezoelectric film fabricated by AD showed a piezoelectric constant (–d31) of 158 mp/V on a stainless substrate, which is almost the same value as that of film fabricated by conventional film-forming technology. On a zirconate substrate, 360 pm/V was achieved, which is similar to that for bulk ceramics. As shown in Figure 7.24, where we plot the piezoelectric constant as a function of film thickness of the actuator, AD covers a broad range of thicknesses—between bulk ceramic material and thin film—with a better piezoelectric constant than screen printing (http://www.nec.co.jp/techrep/en/journal/g07/n01/070115.html).

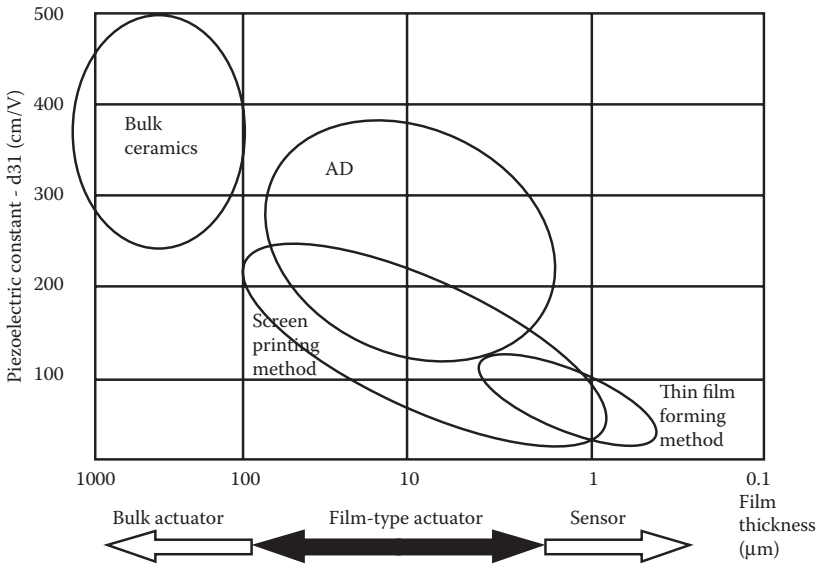


FIGURE 7.24 Plot of the piezoelectric constant (–d31) as a function of film thickness of the actuator [solid solution of PNN (lead nickel niobate) and PZT (lead zirconate titanate)] and manufacturing means. AD covers the thickness range between bulk and thin film. (Based on http://www.nec.co.jp/techrep/en/journal/g07/n01/070115.html by Dr. Han Xu.)

Chemical Vapor Deposition

Introduction

Chemical vapor deposition (CVD) is one of the principal thin film manufacturing options. In Figure 7.25 we show how thin film processes fit in the overall IC fabrication scheme. In the case of thin film of silicon dioxide, if the surface of the wafer is silicon (e.g., at the start of IC fabrication), then thermal oxidation is the appropriate process by which to form a layer of SiO_2. If the oxide layer must be grown over materials other than silicon, such as aluminum or silicon nitride, then an alternative technique must be used, such as CVD. During CVD, the constituents of a vapor phase, often diluted with an inert carrier gas, react at a hot surface (typically higher than 300°C) to deposit a solid thin film. In CVD, the diffusive-convective transport to the hot substrate surface involves many intermolecular collisions, and mass and heat transfer modeling of deposition rates is consequently much more complex than in PVD. Chemical vapor deposition of SiO_2 is accomplished by reacting a silicon compound such as silane (SiH_4) with oxygen onto a heated substrate. The reaction is carried out at around 425°C and can be summarized by:

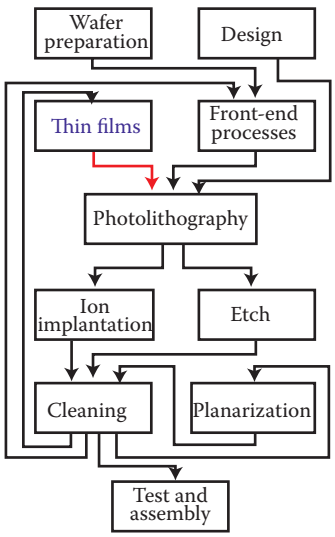


FIGURE 7.25 Thin films. CVD is a broad class of processes using controlled chemical reactions to create thin films on substrates. A variant of CVD, called *plasma-enhanced chemical vapor deposition* (PECVD), uses gas plasma to lower the temperature required to obtain a chemical reaction and achieve film deposition.

$$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2 \qquad \text{Reaction 7.1}$$

The density of a CVD silicon dioxide film and its bonding to the substrate are generally poorer than that achieved by thermal oxidation. Consequently, CVD is used only when the preferred process is not feasible: when the substrate surface is not silicon or when the high temperature used in thermal oxidation cannot be tolerated.

In the CVD reaction chamber, the reactants are adsorbed on the heated substrate surface, and the adatoms undergo a series of processes resulting in film formation. When a molecule has reached the surface, the required reaction analysis is the same, regardless of deposition method. The molecular phenomena at the surface to be considered include sticking coefficient, surface adsorption, surface diffusion, surface reaction, desorption, and film or crystal growth. Gaseous byproducts are desorbed and removed from the reaction chamber. The reactions forming a solid material do not always occur on or close to the heated substrate (heterogeneous reactions) but can also occur in the gas phase (homogeneous reactions). As homogeneous reactions lead to gas-phase cluster deposition and result in poor adhesion, low density, and high defect films, heterogeneous reactions are preferred. The slowest of any of the CVD steps mentioned, the gas phase or surface process determines the rate of deposition. The sample surface chemistry, its temperature, and thermodynamics determine the compounds deposited. The most favorable end product of the physical and chemical interactions on the substrate surface is a stoichiometric-correct film. Several activation barriers need to be surmounted to arrive at this end product. Some energy source, such as thermal, photon, or ion bombardment, is required to achieve this.

The CVD method is very versatile and works over a wide range of temperatures and pressures. Amorphous, polycrystalline, epitaxial, and uniaxially oriented polycrystalline layers can be deposited with a high degree of purity, control, and economy. CVD is used extensively in the semiconductor industry and has played an important role in transistor miniaturization, for example, by introducing very thin film deposition of silicon. More recently, CVD copper and low dielectric insulators ($\varepsilon < 3$) also have become important CVD applications. CVD embodies the principal building technique in surface micromachining (see section on Doped Polysilicon). Modern interest in CVD is also focused on coating applications, such as coated carbide tools, solar cells, depositing refractory metals on jet engine turbine blades, and other applications where resistance to wear, corrosion, erosion, and thermal shock

Table 7.10 Example CVD Reactions

Process Description Reaction
1. The Mond process includes a CVD process for decomposition of nickel from nickel carbonyl [$Ni(CO)_4$], which is an intermediate compound formed in reducing nickel ore: $$Ni(CO)_4 \xrightarrow{400\,°F\,(200\,°C)} Ni + 4CO$$
2. Coating of titanium carbide (TiC) onto a substrate of cemented tungsten carbide (WC-Co) to produce a high-performance cutting tool: $$TiCl_4 \xrightarrow{1800\,°F\,(1000\,°C),\ excess\ H_2} TiC + 4HCl$$
3. Coating of titanium carbide (TiC) onto a substrate of cemented tungsten carbide (WC-Co) to produce a high-performance cutting tool: $$TiCl_4 + 0.5N_2 + 2H_2 \xrightarrow{1650\,°F\,(900\,°C)} TiN + 4HCl$$
4. Coating of aluminum oxide (Al_2O_3) onto a substrate of cemented tungsten carbide (WC-Co) to produce a high-performance cutting tool: $$2AlCl_3 + 3CO_2 + 3H_2 \xrightarrow{900\,°F\,(500\,°C)} Al_2O_3 + 3CO + 6HCl$$
5. Coating of silicon nitride (Si_3N_4) onto silicon (Si), a process in semiconductor manufacturing: $$3SiF_4 + 4NH_3 \xrightarrow{1800\,°F\,(1000\,°C)} Si_3N_4 + 12HF$$
6. Coating of silicon dioxide (SiO_2) onto silicon (Si), a process in semiconductor manufacturing: $$2SiCl_3 + 3H_2O + 0.5O_2 \xrightarrow{1600\,°F\,(900\,°C)} 2SiO_2 + 6HCl$$
7. Coating of the refractory metal tungsten (W) onto a substrate, such as a jet engine turbine blade: $$WF_6 + 3H_2 \xrightarrow{1100\,°F\,(600\,°C)} W + 6HF$$

Table 7.11 Comparison of a Typical PVD Method (Sputtering) with a CVD Method

	PVD	CVD
Example processes	Sputtering	PECVD
Advantages	Clean process Easy to control composition Easy mechanism Good adhesion Safe process High deposition rate	Good step coverage
Disadvantages	Poor step coverage	Expensive High impurity using toxic gas hardware complexity
Example of deposited films	Ti, TiN, Co, W	W, TiN (MOCVD), Ti, TiN, … SiO_2, Dielectrics

are important. Table 7.10 presents some examples of CVD reactions that result in deposition of a metal or ceramic coating onto a suitable substrate.

A few words about the difference between physical vapor deposition (PVD) and chemical vapor deposition processes follow. PVD constitutes a group of direct line-of-sight impingement deposition techniques, and CVD represents a class of diffusive-convective mass transfer deposition techniques. Whereas CVD includes a wide range of pressures and temperatures and can be applied to a great variety of coating and substrate materials, PVD is mostly limited to metal film deposition. With PVD techniques we saw that sputtering leads to a better conformal coating than evaporation, but CVD is better than sputtering in this respect. Unlike CVD, which operates at elevated temperature, PVD operates at room temperature; consequently, CVD often produces better crystalline films than PVD. Advantages typically cited for CVD include that 1) it is possible to deposit refractory materials at temperatures below their melting or sintering temperatures, 2) control of grain size is possible, 3) the process can be carried out at atmospheric pressure (it does not require vacuum equipment); and 4) there is good bonding of coating to substrate surface.

Disadvantages include that 1) the corrosive and/or toxic nature of chemicals generally necessitates a closed chamber and special pumping and disposal equipment, 2) certain reaction ingredients are relatively expensive, and 3) material utilization is low. Many gases used in CVD systems are toxic (hazardous to humans), corrosive (causes corrosion to stainless steel and other metals), flammable (burns when exposed to an ignition source and an oxygen source), and explosive and/or pyrophoric (spontaneously burn or explode in air, moisture, or when exposed to oxygen). In Table 7.11 we compare a specific PVD technique (sputtering) with a specific CVD technique, i.e., plasma-enhanced chemical vapor deposition (PECVD). With CVD, layer thicknesses range from 1 monolayer to 100 µm.

Reaction Mechanisms and CVD Reactor

Figure 7.26 illustrates the various transport and reaction processes underlying CVD schematically[23]:

- Mass transport of reactant and diluent gases (if present) in the bulk gas flow region from the reactor inlet to the deposition zone (1)
- Gas-phase reactions (homogeneous) leading to film precursors and byproducts (often unselective and undesirable) (2)
- Mass transport of film precursors and reactants to the growth surface (3)
- Adsorption of film precursors and reactants on the growth surface (4)
- Surface migration of film formers to the growth sites (5)
- Surface reactions (heterogeneous) of adatoms occurring selectively on the heated surface, and incorporation of film constituents into the growing film, that is, nucleation (island formation) (6)
- Desorption of byproducts of the surface reactions (7)
- Mass transport of byproducts in the bulk gas flow region away from the deposition zone toward the reactor exit (8)

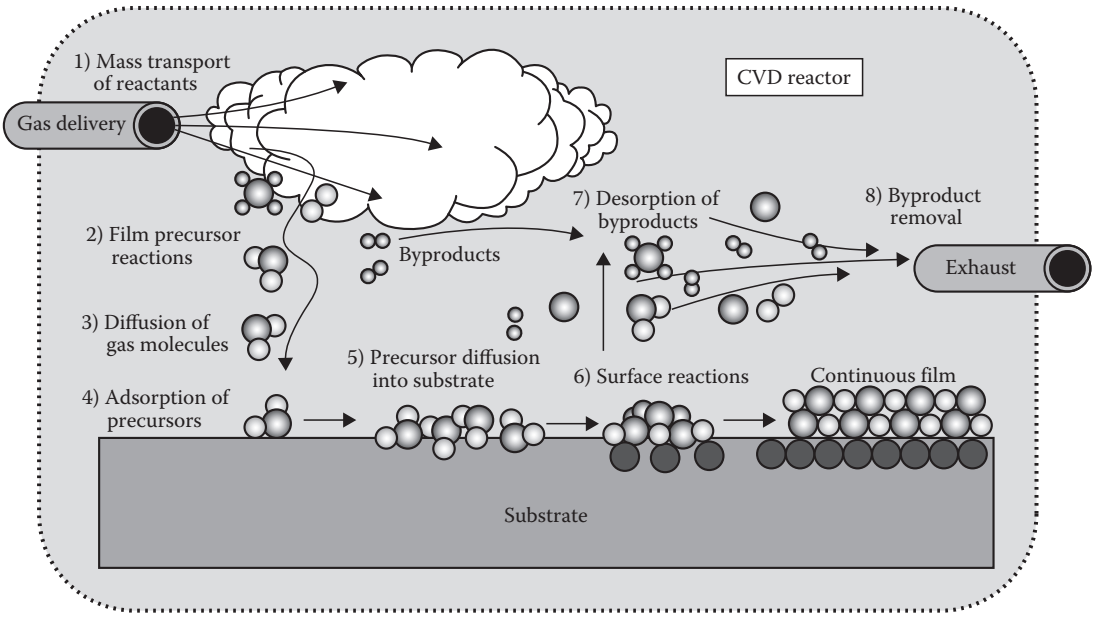


FIGURE 7.26 Schematic of transport and reaction processes underlying CVD. (Based on Jensen, K. F. 1989. *Microelectronics processing: chemical engineering aspects.* Eds. D. W. Hess, and K. F. Jensen. Washington, DC: American Chemical Society.[23])

CVD processes are carried out in a reactor as shown in Figure 7.27, which consists of 1) reactant supply system, 2) deposition chamber, and 3) recycle/disposal system. The purpose of the reactant supply system is to deliver reactants to the deposition chamber in the proper proportions. The deposition chamber contains the substrates and the chemical reactions that lead to the deposition of reaction products onto the substrate surfaces. Deposition occurs at elevated temperatures, and the substrate may be heated by induction heating, radiant heat, or other means. Deposition temperatures for different CVD reactions can range from around 250–1950°C; therefore, the chamber must be designed to meet these temperature demands. The recycle/disposal system is to render harmless the byproducts of the CVD reaction.

In the case of a thermally driven CVD reaction, a temperature gradient is imposed on the reactor; the gas-phase species (e.g., SiH_4) forms in a hot region; and the equilibrium shifts toward the desired solid (e.g., Si) in a slightly colder region. Either the gas phase or the surface processes can determine rate. Thermal energy is the sole driving force in high-temperature CVD reactors; for lower-temperature deposition, an additional energy source is needed. A radiofrequency (RF) plasma, photoradiation, or laser radiation can be used to enhance the process,

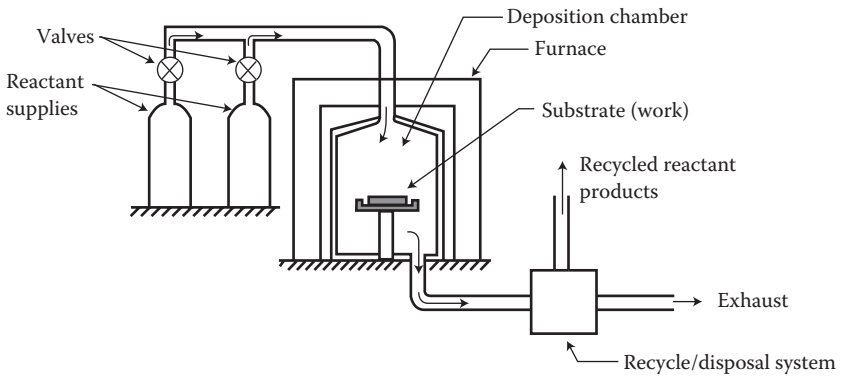


FIGURE 7.27 Schematic of a CVD reactor.

known as *plasma-enhanced chemical vapor deposition* (PECVD), photon-assisted CVD,[24] and laser-assisted CVD (LCVD),[25] respectively.

With photon- and laser-assisted CVD systems, part of the energy needed for deposition is provided by photons. This method fills the need for lower-temperature deposition processes. With a laser source, it is possible to write a pattern on a surface directly by scanning the micrometer-size light beam over the substrate in the presence of the suitable reactive gases. By adjusting the focal point of the laser continuously, it is even possible to grow 3D microstructures such as fibers and springs in a wide variety of materials such as boron, carbon, tungsten, silicon, SiC, Si_3N_4, and so on. Laser CVD is treated as a thermal additive technique in Chapter 9.[25]

In PECVD, which also encompasses high-density plasma CVD (HDPCVD), plasma activation provides the radicals that result in the deposited films, and ion bombardment of the substrate provides the energy required to arrive at the stable desired end products. Operational temperatures are lower because part of the activation energy needed for deposition now comes from the plasma.

Because CVD processes involve carrier gas flow over the substrate's surface, the issue of the "fluid/solid interface" must be dealt with in assessing the effectiveness of the CVD process. Whenever a fluid is flowing over a solid surface, a "boundary layer" is created, as shown in Figure 7.28. The boundary layer acts as a barrier to transfer of heat or transport of medium. Transport in the gas phase takes place through diffusion proportional to the diffusivity of the gas, *D*, and

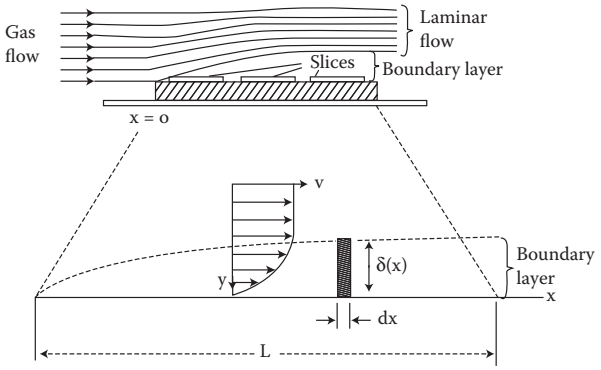


FIGURE 7.28 Development of a boundary layer in gas flowing over a flat plate. The figure shows an expanded view of the boundary layer. (Granger, R. A. 1995. *Fluid mechanics.* New York: Dover Publications.[26])

the concentration gradient across the boundary layer that separates the bulk flow (source) and substrate surface (sink). The flux of depositing material is given by Volume I, Equation 4.20 (Fick's first law):

$$J = -D\frac{\delta N(x,t)}{\delta x}$$

The boundary layer thickness, $\delta(x)$, as a function of distance along the substrate, *x* (see Figure 7.28), can be calculated from (x = 0 is the leading edge):

$$\delta(x) = \left(\frac{\eta x}{\rho V}\right)^{\frac{1}{2}} \tag{7.11}$$

where η = gas viscosity
ρ = gas density
V = gas stream velocity parallel to the substrate

The average boundary layer thickness, in the boundary layer model from Prandtl,[26] can then be calculated over the whole plate as follows:

$$\delta = \frac{1}{L}\int_0^L \delta(x)dx = \frac{2}{3}L\left(\frac{\eta}{\rho VL}\right)^{\frac{1}{2}} \tag{7.12}$$

where *L* stands for the length of the plate receiving the deposit.

The dimensionless Reynolds number, *Re*, is a most important figure in fluid dynamics characterizing the nature of a fluid flow (gas or liquid) (see Volume I, Chapter 6). It is given by:

$$Re = \frac{\rho VL}{\eta} \qquad \text{(Volume I, Equation 6.71)}$$

It equals the ratio of the magnitude of inertial effects to viscous effects in fluid motion. For low values (<2000–3000), the flow regime is called *laminar*, whereas for larger values the regime is *turbulent*. Substituting the expression for the Reynolds number in Equation 7.12, we obtain:

$$\delta = \frac{2L}{3\sqrt{Re}} \tag{7.13}$$

Control of δ and its uniformity over the substrate is an important design factor. In a horizontal CVD reactor (see Figure 7.29a) the substrate is often tilted

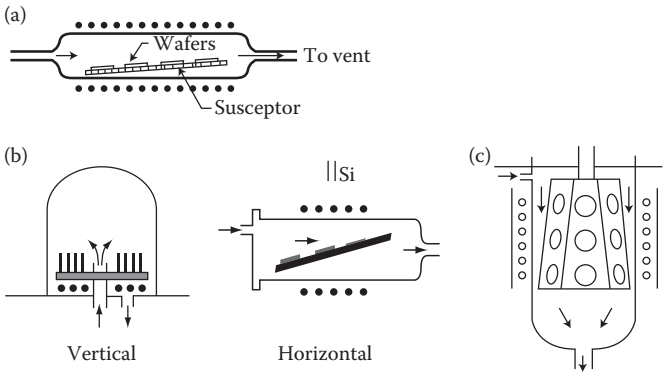


FIGURE 7.29 CVD reactor configurations: horizontal (a), vertical reactor (b), and barrel CVD reactor (c).

1–3° for Si and 5–15° for GaAs. The susceptor is tilted so that the cross-sectional area of the chamber is decreased, increasing the gas velocity along the susceptor. This compensates for both the boundary layer and depletion effects.

In a vertical reactor (see Figure 7.29b and c) one may use a baffle to deflect the gas from the center of the susceptor in addition to substrate rotation, again leading to a more uniform δ. The stability of the flow in a CVD reactor is crucial in achieving uniform deposition. The criterion for flow stability depends on whether the flow is fully developed (i.e., laminar) before it reaches the susceptor. Gas flows in containment structures (internal flows), e.g., a cylindrical tube CVD reactor, differ from external flows in that there is a boundary layer and a uniform free stream in the entry region of an internal flow that accelerates according to the rate of growth of the boundary layer as shown in Figure 7.30. The velocity is uniform at the entrance, and the boundary layer grows with distance from the entrance until the flow becomes fully developed at X_L. From the continuity equation (see Volume I, Chapter 6) it follows that the frictionless core must accelerate, whereas from applying the Bernoulli equation (see Volume I, Chapter 6) to a streamline in the free stream region, we learn that the pressure must decrease. At X_L the velocity varies over the entire channel, and there is no free stream anymore. The laminar development length X_L for the flow to become fully developed can be empirically determined as:

$$X_L = 0.056\,Re\,D \tag{7.14}$$

where D is the diameter of the tube. The thermal entrance length X_T is the length beyond which the dimensionless temperature profile is constant with x and before which it varies with x (i.e., is developing). For laminar flows in a tube, the thermal entrance length is a function of the Reynolds number and the Prandtl number, and for uniform surface temperature it is given as:

$$X_T = 0.033\,Pr\,Re\,D \tag{7.15}$$

where the Prandtl number is defined as $Pr = \nu/\lambda_d$ (see Volume I, Equation 6.61), i.e., the ratio of kinematic viscosity and thermal diffusivity. The interpretation of this number is the ratio of the thickness of the fluid velocity boundary layer to the fluid temperature boundary layer (for water Pr is 10, and for air it is 0.71). Both velocity and thermal entrance length are considerably longer when the flow is turbulent.

By substituting the value for the average boundary layer thickness from Equation 7.12 in Volume I, Equation 4.20 (i.e. Fick's first law with dx = δ), the following expression for the material flux to the substrate results:

$$J = -D\frac{\Delta N}{2L}3\sqrt{Re} \tag{7.16}$$

According to Equation 7.16, the film growth rate in the mass flow-controlled regime should depend on the square root of the gas velocity, V (because the Reynolds number is proportional to gas velocity V).

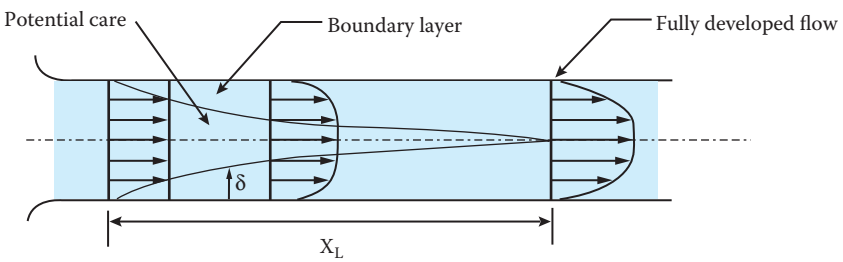


FIGURE 7.30 Flow in the entry region of a pipe for laminar flow.

Table 7.12 Mass Transport: Fundamental and Experimental Parameters

Fundamental Parameters	Experimental Parameters
Reactant concentration	Pressure
Diffusivity	Gas velocity
Boundary layer thickness	Temperature distribution
	Reactor geometry
	Gas properties (viscosity…)

Because gas-phase mass transport (see Table 7.12) is not an activated process, it is not strongly temperature-dependent. In Figure 7.31A we show a plot of silicon growth rate (in micrometers per minute) as a function of the gas flow rate (in liters per minute and proportional to the gas velocity in a fixed volume reaction chamber) with a substrate temperature of 1270°C, and the predicted square root dependence is clearly observed over a wide range of flow rates.[27]

At high flow rates, the growth rate reaches a maximum and then becomes independent of flow. In this regime, the reaction rate controls the deposition, evidenced by the exponential dependence of the growth rate on temperature observed at those flow rates (see Figure 7.31Ba).[27]

Surface reactions can be modeled by a thermally activated phenomenon proceeding at a rate, k, given by:

$$k = k_0 e^{\frac{E_a}{kT}} \tag{7.17}$$

where k_0 = frequency factor

E_a = activation energy in electronvolts

In Figure 7.31Bb, the growth rate (in micrometers per minute) of Si as a function of substrate temperature and a variety of precursor gases is illustrated in an APCVD reactor. From the slope of an Arrhenius plot, as demonstrated here, the activation energy of the rate-determining surface process can be deduced (for most surface reactions, E_a ranges from 1–4 eV).

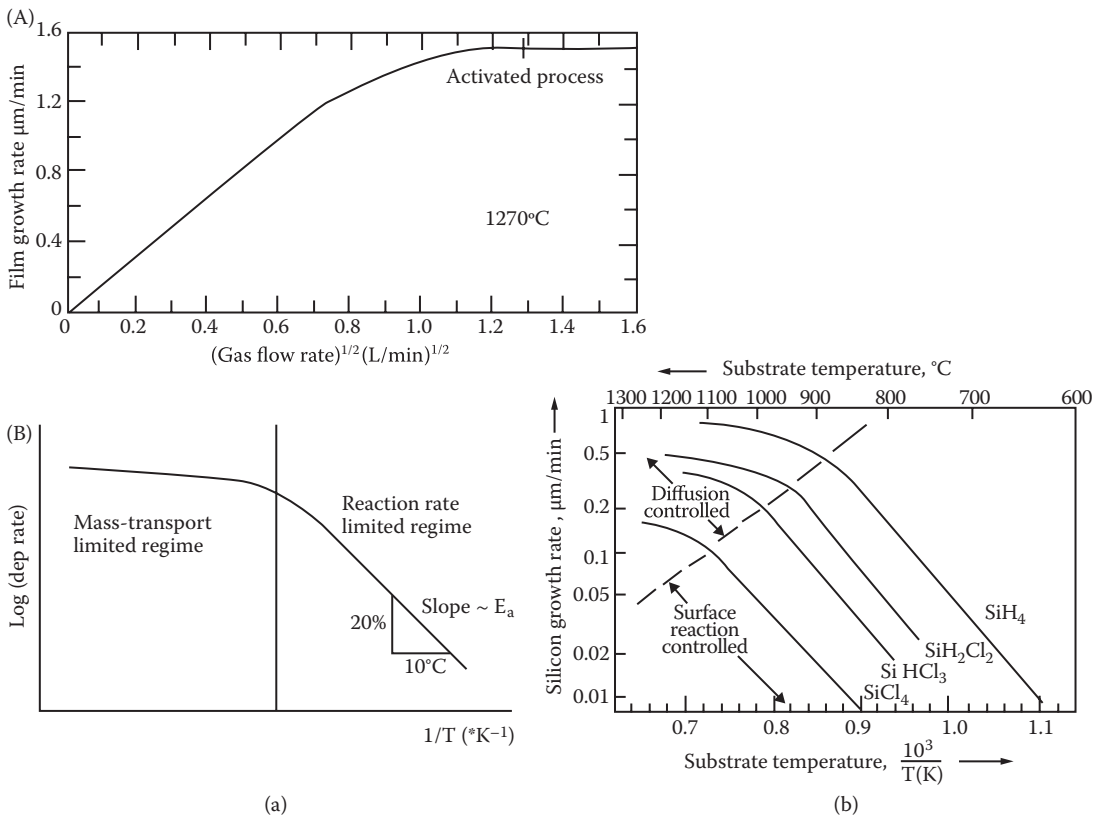


FIGURE 7.31 Growth rate dependence in a Si CVD process as a function of gas flow rate (A) and temperature (B). At high flow rates, the growth rate reaches a maximum and then becomes independent of flow. In this regime, the reaction rate controls the deposition, evidenced by the exponential dependence of the growth rate on temperature observed at those flow rates (see a). The growth rate (in micrometers per minute) of Si as a function of substrate temperature and a variety of precursor gases is illustrated in an APCVD reactor (see b). (Bb) Various precursor gases are used. (Wolf, S., and R. N. Tauber. 1987. *Silicon processing for the VLSI era.* Sunset Beach, CA: Lattice Press.[27])

We want the deposition reaction to occur at the gas/wafer boundary (heterogenous reaction). However, the higher temperatures (650–800°C for silane) used to enhance surface migration also result in gas-phase decomposition of silane, resulting in particles (homogeneous reaction). This is undesirable because particles tend to fall onto the wafers. Higher-quality (fewer defects) material is obtained at temperatures higher than 800°C. To be able to operate at higher temperatures, instead of silane, chlorosilanes that are stable up to higher temperatures are commonly used. Dichlorosilane (SiH_2Cl_2) is the most common and allows growth at ~800–1050°C. For a certain rate-limiting reaction, the temperature may increase high enough for the reaction rate to exceed the rate at which reactant species arrive at the surface. In such a case, the reaction rate cannot proceed any faster than the rate at which the reactant gases are supplied to the substrate by mass transport, no matter how high the temperature is increased (see plateau in Figure 7.31Ba and b). This situation is referred to as a mass transport-limited deposition process. Temperature is less important in this regime than it is in the reaction rate-limited (Arrhenius) one. In the latter case, the arrival rate of reactants is less important because their concentration does not limit the growth rate.

A direct practical application of these two possible rate-limiting processes is the way substrates are stacked in low-pressure CVD (LPCVD) (<10 Pa) versus atmospheric pressure to slightly reduced pressure CVD (APCVD) (±100–10 kPa) reactors. In an LPCVD reactor (e.g., at ~1 Torr), the diffusivity of the gas species is increased by a factor of 1000 over that at atmospheric pressure, resulting in a one order of magnitude increase in the transport of reactants to the substrate. The rate-limiting step becomes the surface reaction. LPCVD reactors enable wafers to be stacked vertically at very close spacings as the rate of arrivals of reactants is less important (Figure 7.32). On the other hand, APCVD, operating in the mass transport-limited regime, must be designed such that all locations on the wafer and all wafers are supplied with an equal flux of reactant species. In this case, the wafers often are placed horizontally or at a slight incline (Figure 7.29a). In APCVD, reactor walls typically are cooled in a so-called *cold wall reactor type* (where only the susceptor holding the substrates is heated) to minimize particulate and impurity problems caused by deposition on them. Horizontal tube, hot wall reactors are the most widely used LPCVD reactors. In this case, not only the wafers but also the reaction chamber walls get coated. Such systems require frequent cleaning to avoid serious particulate contamination. They find wide application because of their economy, throughput, and uniformity.

FIGURE 7.32 Vertical stacking of substrates in LPCVD reactors.

The Reynolds number in a typical LPCVD reactor is smaller than that for APCVD. Consequently, the thermal entrance length for APCVD is longer than that for LPCVD reactors (see Equation 7.15).

Step Coverage

Thin deposited films grow more readily on horizontal, exposed surfaces than within trenches or holes, resulting in poor coverage of the surface with the film. The gaps that are formed can lead to shorts, reliability problems, etching failures, and other generally undesirable phenomena. Poor step coverage results from different deposition rates in different parts of a microstructure. The profile of the thin films deposited by any CVD or PVD techniques depends on:

1. Equipment configuration
2. Deposition method (e.g., LPCVD, PECVD, PVD)
3. Reaction chemistry
4. Reactant transport mechanism

TABLE 7.13 Mean Free Path in Various Gases at 20°C

Gas	P_T (Torr) λ (m) (cm·Torr) at 20°C	
O_2	4.9×10^{-3}	Atmospheric pressure (760 Torr) ~40 nm (smaller than most chip features)
N_2	4.6×10^{-3}	
H_2	9×10^{-3}	Low vacuum (0.76 Torr) ~40 μm (larger than most chip features)
He	13.6×10^{-3}	
Ar	4.8×10^{-3}	Medium vacuum (7.6 Torr) ~4.0 mm (larger than working gap)
Ne	9.3×10^{-3}	
Xe	2.7×10^{-3}	High vacuum (7.6 μTorr) ~4.0 m (larger than chamber)
CO_2	3.0×10^{-3}	
H_2O	3.0×10^{-3}	Evaporated material travels straight to wafer (line of sight; thus, is very directional)
Cl	2.3×10^{-3}	
NH_3	3.5×10^{-3}	

Here we emphasize step coverage in CVD processes. The mean free path of a molecule, λ, based on a slightly modified Equation 7.6, is given by:

$$\lambda = \frac{kT}{2^{\frac{1}{2}} P_T \pi a^2} \tag{7.18}$$

where *a* is the molecular diameter.[28] If we assume a = 3 Å and T = 300 K, then $P_T(\text{torr})\lambda(\text{m}) = 7.810^{-5}$. Therefore, if λ = 30 cm, then a pressure of 2.6×10^{-4} Torr is required (see also Table 7.13).

Equation 7.18 is of crucial importance in understanding CVD coating of micromachined features. CVD films may electrically isolate underlying features from subsequent layers or protect them against the atmosphere, determined by the degree at which edges and pits of the underlying features can be covered uniformly. As demonstrated in Figure 7.33A, three cases can be distinguished. Ideally, a uniform, dense, and conformal coating should form (Figure 7.33Aa). This can occur in instances where the reactants, after first hitting the solid, have enough energy left for surface migration before a bond is established with the underlying substrate. Coatings in which equal film thickness exists over all substrate topography, regardless of its slope, are called *conformal* (see also Figure 7.5). In a second case, as shown in Figure 7.33Ab, the mean free path of the molecules is large enough to reach the bottom of the trench, but little energy remains for surface migration. Finally, in the third case, shown in Figure 7.33Ac, the mean free path length is too short to reach the bottom, and there is little surface migration (temperature too low). Whereas sputtering is better than evaporation, for the best conformal coating CVD is preferred. Evaporated and sputtered metal films often have trench profiles as shown in the Figure 7.33Ab, whereas CVD (especially LPCVD) deposited films are often more uniform and conformal, as demonstrated in Figure 7.33Aa.

The value of the integral of the material flux in Equation 7.16 (∫Jdθ), and thus the CVD film thickness, is directly proportional to the range of feasible angles of arrival, θ, of the depositing species (in the absence of surface migration). Different

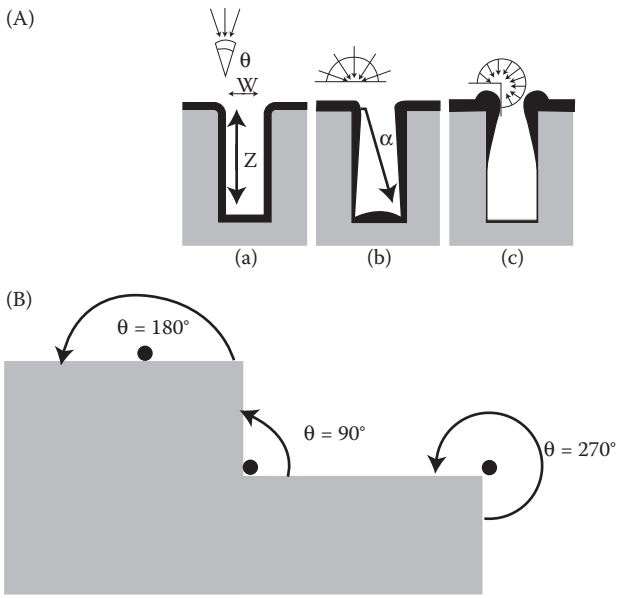


FIGURE 7.33 Step coverage cases of deposited film. (Aa) Uniform coverage resulting from rapid surface migration. (Ab) Nonconformal step coverage for long mean free path and no surface migration. Distance a is the longest path a molecule travels to reach the corner of the trench. (Ac) Nonconformal step coverage for short mean free path and no surface migration. (B) Different arrival angles in two dimensions.

arrival angles in two dimensions are illustrated in Figure 7.33B. The arrival angle at a planar surface is 180°. At the top of a vertical step, the arrival rate is nonzero over a range of 270°; the resultant film thickness is 270/180, or 1.5 times greater than that for the planar case. At the bottom corner of a trench, the arrival angle is only 90°, and the film thickness is 90/180 or one-half that of the planar case. The CVD profile in Figure 7.33Ac, where the mean free path is short compared with the trench dimensions and there is no surface migration, reflects the 180°, 90°, and 270° arrival angles. The thick cusp at the top of the step and the thin crevice at the bottom combine to give a re-entrant shape that is particularly difficult to cover with evaporated or sputtered metal. Gas depletion effects also are observed along the trench walls. Along the vertical walls, the arrival angle, θ, is determined by the width of the opening, and the distance from the top and can be calculated from:

$$\theta = \arctan\frac{w}{z} \tag{7.19}$$

where w is the width of the opening and z the distance from the bottom of the trench to the top surface (Figure 7.33Aa). This type of step coverage, as shown in Figure 7.33Ab, thins along the vertical walls and may have a crack at the bottom of the step. For uniformity of deposition, in the case of Figure 7.33Ab, where the mean free path is longer than the distance α (the longest path a molecule must travel to reach the corner of the trench), the rate of surface migration of adspecies should exceed the rate of adsorption of adspecies. The condition of $\lambda > \alpha$ can be met by working at low pressures, based on Equation 7.18:

$$\frac{kT}{2^{\frac{1}{2}} P_T \pi a^2} > \alpha \tag{7.20}$$

For a given gas, Equation 7.20 gives us the pressures and temperatures at which $\lambda > \alpha$. The most important variable influencing the requirement for large surface diffusion is the reactor temperature. The mobility of surface species is largest on metallic and semiconducting surfaces, where bonding is not very directional, and rather limited on dielectrics, where highly directional covalent bonds tend to hold a molecule in one place once it is chemisorbed. There is another factor influencing conformality, namely, the reactive stiction coefficient, S_c, which might be even more important than surface diffusion. To appreciate the importance of this parameter we need to reintroduce some simple facts about the mean free path in various deposition systems. Remember that the mean free path for APCVD is 40 nm at 760 Torr (see Table 7.13), smaller than most features on an IC or a micromachine. However, the mean free path is 10 μm in typical working conditions for LPCVD and PECVD, and for sputtering it is a few centimeters; therefore, in this case one can ignore collisions between the gas particles when they are inside a via or a trench-like structure. The only collisions are those with the surface of the wafer. The particular flow characteristics around an IC, MEMS, or NEMS feature on a substrate are given by the value of the Knudsen number (*Kn*), defined in this case as the ratio of the mean free path of the molecules (λ) to the characteristic dimension of the structure to be coated (e.g., the width D of the structure) (see also Equation 7.5). If the Knudsen number is greater than 1, the flow is free molecular. If, on the other hand, the Knudsen number is <0.01, the flow is viscous, that is, the atom or molecule bounces several times into the feature's sidewalls. A transition region exists between the two. The flow around micromachined features is typically in this transition regime ($0.1 < Kn = \lambda/w < 10$) or in the free molecular regime ($Kn > 10$).

The transport of reactants to the surface at low pressures is sketched in Figure 7.34. The three processes shown here are 1) direct deposition, 2) re-emission, and 3) surface diffusion. In most applications, it has been found that the surface diffusion (3) is less important than re-emission (2), which is controlled

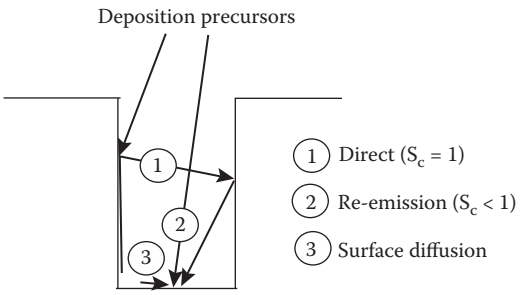


FIGURE 7.34 The transport of reactants to the surface at low pressures.

by the reactive stiction coefficient, S_c. The stiction coefficient is the ratio of incident precursor species minus re-emitted ones, i.e., all the species that remain on the surface react, over the total of incident precursor species:

$$S_c = \frac{\Gamma_{\text{incident}} - \Gamma_{\text{re-emission}}}{\Gamma_{\text{incident}}} = \frac{\Gamma_{\text{reaction}}}{\Gamma_{\text{incident}}} \qquad (7.21)$$

The reactive sticking coefficient (RSC) S_c depends on the nature of the incoming species (molecule, ion, neutral, radical, size of molecule) and on the surface conditions. A radical or other unstable precursor molecule is likely to react readily with a surface site and stick wherever it strikes ($S_c = 1$). A more stable precursor may not react readily, particularly at higher temperatures or when the surface is already saturated with adsorbents ($S_c < 1$). Particularly in dielectric deposition, it has been shown that an important mechanism for transport of the material into gaps and holes is through multiple adsorption and re-emission from the surface: that is, from precursors having a low RSC. When such "reflections" occur many times, the flux of precursor to the surface inside a hole may differ very little from the flux onto the horizontal surfaces. In other words, the likelihood of reaction is important because successive adsorption and re-emission is one of the key mechanisms determining the effect of topography of a surface on the thickness of a film deposited on that surface, and is one of the key advantages of CVD over evaporation or sputtering of thin films. Therefore, for a direct deposition (1) $S_c = 1$ and with any degree of re-emission (2) $S_c < 1$. Shadowing of the direct flux by the walls reduces the flux at the bottom corners of a via or a trench, resulting in thinner deposition. Lower reaction stiction coefficients increase the number of bounces in the trench or via and hence increase the probability of deposition precursors reaching the bottom. Thus, step coverage is better for lower S_c (see Figure 7.35).

To illustrate the effect of a lower S_c, imagine that we have a couple of metal lines (e.g., aluminum) on a substrate that are being used as wires, and that we need to cover them up with an insulating dielectric so that a subsequent layer of metal can be deposited without being shunted out. When leaving voids

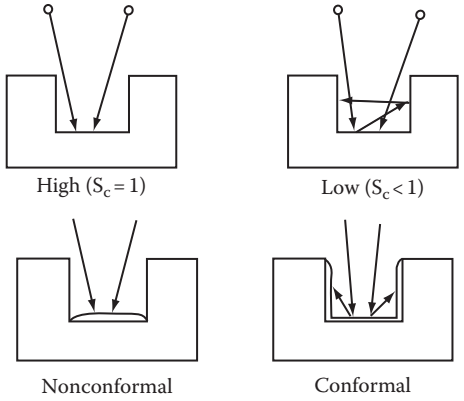


FIGURE 7.35 A low S_c can produce more conformal coverage because of redeposition (usually more important than surface diffusion)

in a fill they trap chemicals that may lead to cracks (dielectric) or large contact resistance and sheet resistance (metallization). In Figure 7.36 we show how a decreasing S_c fills the gap between the two metal lines better with the dielectric material and that at a value of $S_c = 0.007$ the gap is filled completely. The simulation model used here is SPEEDIE,

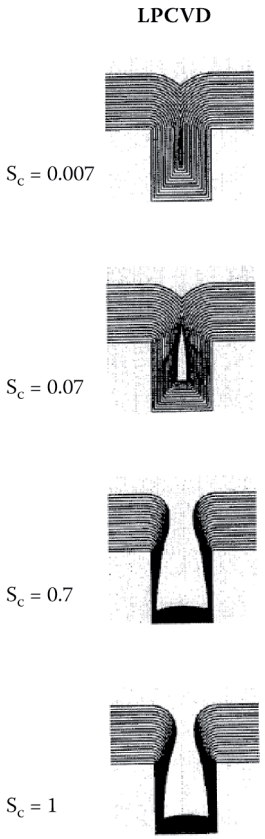


FIGURE 7.36 Effect of varying S_c on filling a gap between two metal lines. The simulation model used here is SPEEDIE, the Stanford etching and deposition profile simulator.

the Stanford etching and deposition profile simulator. Scaling of interconnects in future ICs will require more metal layers with reduced interconnect pitch. The height of the metals also has to be increased. The larger aspect ratio between the metal lines will increase dramatically the problem of step coverage. For a small gap or opening in a film, the aspect ratio is the ratio of its depth to width. Filling high-aspect-ratio gaps (>3:1) is critical for advanced wafer fabrication. It is anticipated that the thickness of deposited films in the IC industry will reach the order of 10 nm rather than a few micrometers, and thickness uniformity will be harder to maintain.

Although LPCVD reactors have lower deposition rates than APCVD reactors, it is found that this is more than compensated for by the high wafer capacity and coating uniformity. The future trend is toward single-wafer processing systems instead of the batch processes. This is because of the increasing wafer sizes (8 in. or more), and the development in technology to achieve better film uniformity in modern single-wafer reactors. The tendency is toward large single wafers that are transferred from station to station in so-called *cluster tools*.

Besides SPEEDIE, simulation models for nonplanar CVD were presented, for example, by Coronell et al.[29] The latter model again provides a picture of the evolution of the depositing film profile. The parameters investigated include the sticking coefficient, the surface mobility of the adsorbed reactants, the Knudsen number (the ratio of the mean free path to the feature size), the feature aspect ratio, and feature geometry.

Overview of CVD Process Types

In Table 7.14[23,30] we review some important CVD processes, listing applications and operational pressures and temperatures for the different types.

Atmospheric Pressure CVD

Atmospheric pressure to slightly reduced pressure CVD (±100–10 kPa) is used primarily to grow epitaxial (i.e., single-crystalline) films of Si and to deposit, at high rates, SiO_2, for example, from the reaction of SiH_4 and oxygen, at low temperatures of 300–450°C [low-temperature oxide (LTO)].[23] The epitaxy processes [also vapor phase epitaxy (VPE)] involve high growth temperatures (>850°C for Si). As we saw above, an APCVD reactor operates in the mass-transport regime so that wafer access, in contrast to the LPCVD process, becomes more important and temperature control less so. APCVD is susceptible to gas-phase reactions, and step coverage is often poor. High gas dilutions help avoid gas-phase nucleation. APCVD, reactor walls typically are cooled in a so-called *cold wall reactor type* (where only the susceptor holding the substrates is heated) to minimize particulate and impurity problems caused by deposition on them. In the early days of semiconductor processing, atmospheric pressure reactors were used mostly to deposit silicon and dielectric films. The advantage of deposition at atmospheric pressure is the simplicity of the reactor design and the high deposition rates (700 Å/min for SiO_2). The wafers are put on a hot susceptor, and reactant gases flow over the susceptor. The susceptor is heated by using high-intensity lamps, radiofrequency (RF) induction, or DC electric current (resistive) heating. These atmospheric pressure reactors have low throughput, require excessive wafer handling during wafer loading and unloading, and have poor thickness uniformity (>10%) across the wafer. To overcome some of these disadvantages, continuous throughput atmospheric pressure reactors were developed as shown in Figure 7.37. In these reactors, the wafers are carried through the reactor on a conveyor belt and are heated by convection. Also, there is a more uniform flow of the reactant gases across the wafer surface. These reactors have high throughput, good uniformity, and the ability to handle large-diameter wafers. The main disadvantages of this type of reactor are that they required frequent cleaning and that particles formed on the reactant dispenser head wind up on the wafers, which impaired process yield and device reliability.

Example 7.2: Epitaxial Si

As a first example of an APCVD application, consider silicon epitaxy at 1200°C (see also Figure 7.31B):

$$SiCl_4\,(gas) + 2H_2 \rightarrow Si\,(solid) + 4HCl\,(gas)$$

Reaction 7.2

Table 7.14 Review of CVD Processes

Process	Advantages	Disadvantages	Applications	Remark	Pressure/ Temperature
APCVD	Simple, high deposition rate (700 Å/min for SiO_2), low temperature	Poorer step coverage, particle contamination	Doped and undoped, low temperature, thick oxides (mainly dielectrics)	Mass transport-controlled	100–10 kPa 250–450°C
LPCVD	Excellent purity and uniformity, conformable step coverage, large wafer capacity	High temperature and lower deposition rate	Doped and undoped, high-temperature oxides, silicon nitride, polysilicon, W, WSi_2	Surface reaction-controlled	100 Pa 550–650°C
VLPCVD			Single-crystalline Si and compound semiconductor superlattices	Surface reaction-controlled	1.3 Pa
MOCVD	Excellent for epi on large surface areas	Safety concerns (highly toxic), very expensive source material	Compound semiconductors for solar cells, laser, photocathodes, LEDs, HEMTs, and quantum wells; also W and Cu	High volume, large surface area production	
PECVD	Lower substrate temperatures, fast, good adhesion, good step coverage, low pinhole density	Chemical (e.g., hydrogen) and particulate contamination, plasma damage	Low-temperature insulators over metals, passivation (nitride)	Tends to have more pinholes than LPCVD	2–5 Torr (pressure higher than in sputter deposition, more gas-phase collisions, less ion bombardment on surface) more collision in gas phase, less ion bombardment on substrate –200–400°C
Atomic layer deposition (ALD)	ALD is the deposition method with the greatest potential for producing very thin, conformal films with control of the thickness and composition of the films possible at the atomic level.	A major limitation of ALD is its slowness; usually only a fraction of a monolayer is deposited in one cycle.	A major driving force for the interest in ALD is for scaling down microelectronic devices.	Many technologically important materials such as Si and Ge, several multi-component oxides and certain metals cannot currently be deposited by ALD in a cost-effective way.	RT to 400°C
Spray pyrolysis	Inexpensive	Difficult to control, not compatible with IC	Gas sensors, solar cells, ITO, large area		Atmospheric 100–180°C

Source: Hess, D. W., and K. F. Jensen, eds. 1989. *Microelectronics processing: chemical engineering aspects.* Washington, DC: American Chemical Society; and Sze, S. M., ed. 1988. *VLSI technology.* New York: McGraw-Hill. With permission.[31]

An intermediate species, $SiCl_2$, is necessary for silicon formation, and at less than 1000°C no $SiCl_2$ forms when starting from $SiCl_4$. Lower-temperature epitaxy can be performed by starting with SiH_2Cl_2 or $SiHCl_3$, which both decompose more readily to $SiCl_2$. When discussing Figure 7.31Bb, we pointed out that using pure silane (in the lowest temperature range) results in gas-phase decomposition of silane, resulting in particles (homogeneous reaction), and silane also must be avoided. Dichlorosilane (SiH_2Cl_2) is the most common and allows growth at ~800–1,050°C.

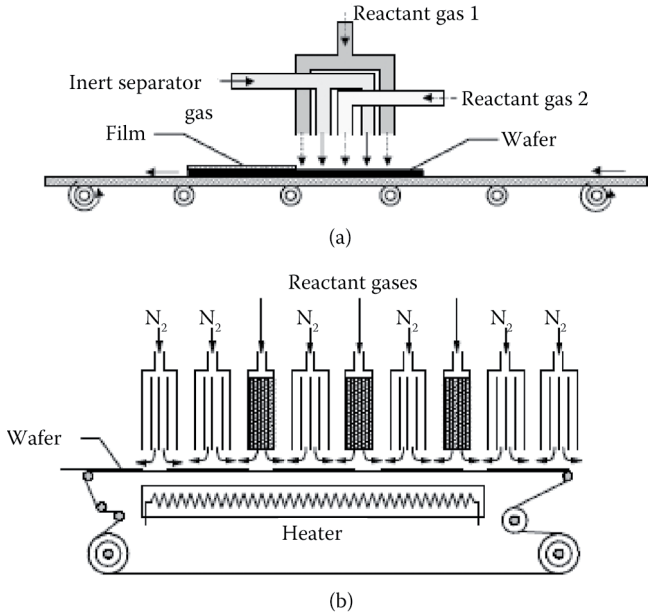


FIGURE 7.37 Simple continuous-feed atmospheric pressure reactor (APCVD): (a) Gas-injection type; (b) plenum type.

In Figure 7.38, we show the growth rate versus temperature for that reaction; the plot delineates a CVD polysilicon deposition region and an epitaxial monocrystalline silicon deposition region.[27] Nominal growth rates for single-crystal Si varies from 0.2–1.5 μm/min, depending on the source gas and the growth temperature. Impurity atoms can be introduced in the gas stream to grow doped epitaxial layers (Figure 7.39) (e.g., arsine-doped epitaxial silicon).

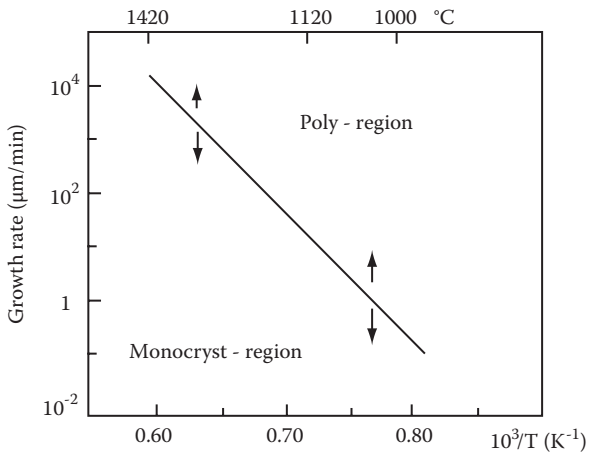


FIGURE 7.38 Growth rate versus temperature. Epitaxy to CVD transition in CVD polysilicon deposition and epitaxial monocrystalline Si deposition. (From Wolf, S., and R. N. Tauber. 1987. *Silicon processing for the VLSI era*. Sunset Beach, CA: Lattice Press. With permission.[27])

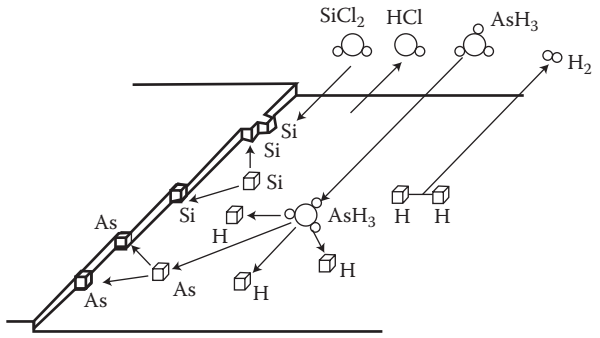


FIGURE 7.39 Arsine-doped epitaxial silicon.

Example 7.3: Oxide APCVD with Tetraethylorthosilicate

In low-temperature oxidation (LTO), silicon dioxide is deposited on the wafer out of the vapor phase from the reaction of silane (SiH_4) with oxygen at relative low temperatures (400–450°C) and low pressure (200–400 mTorr). Silane spontaneously combusts in the presence of oxygen (in other words, silane is pyrophoric), and to avoid a large depletion of the gases at the entrance of the furnace tube, the gases are introduced in the furnace through injectors along the length of the tube. The silane/oxygen gas mixture has been applied in atmospheric pressure CVD (APCVD) and plasma-enhanced CVD (PECVD) conditions. The main advantage of silane/oxygen is the low deposition temperature; the main disadvantage is the nonconformal step coverage.

Because silane is not safe, now tetraethylorthosilicate (TEOS) with oxygen or ozone is commonly used. TEOS is a liquid at room temperature, with a vapor pressure of about 1.5 Torr. TEOS slowly hydrolyzes into silicon dioxide and ethanol when in contact with ambient moisture, but its flammability and toxicity are similar to that of an alcohol. Ozone is more reactive than oxygen; hence the process can be carried out at low temperature without plasma. In general, TEOS brings about a better starting chemistry than the traditional silane-based CVD technologies. Use of TEOS oxide, to replace SiH_4-based oxide in spin-on-glass and photoresist planarization schemes, now has become commonplace for devices with small features. TEOS is especially suited for filling high-aspect-ratio gaps. Trench fill by APCVD of TEOS is illustrated in Figure 7.40.

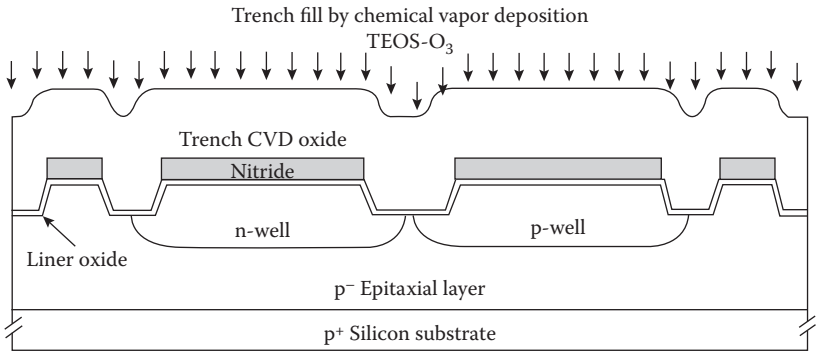


FIGURE 7.40 Trench fill by atmospheric pressure chemical vapor deposition of TEOS.

In APCVD both undoped glass and borophosphosilicate glass have been deposited from TEOS. The process involves the reaction of ozone with TEOS at a pressure of 600 Torr and at a temperature less than 400°C:

$$Si(C_2H_5O)_4 + 8O_3 \rightarrow SiO_2 + 10H_2O + 8CO_2$$

Reaction 7.3

While offering the same good step coverage over submicrometer gaps, the films from the thermal reaction of ozone and TEOS exhibit relatively neutral stress and have a higher film density compared with low-pressure processes. This increased density gives the oxide greater moisture resistance, lower wet etch rate, and smaller thermal shrinkage. Compared with a lower-pressure process, the film density is increased from 2.09–2.15 g/cm^3; the wet etch rate decreased by more than 40%; and the thickness shrinkage changed from 12 to 4% after a 30-min anneal in dry N_2 at 1000°C.[33]

Example 7.4: Carbon Nanotubes Grown by APCVD

Carbon nanotubes are produced in atmospheric pressure CVD (APCVD) by catalytic disproportionation of CO or CH_4 on catalyst particles (e.g., Fe in the form of iron pentacarbonyl). A typical CVD setup consists of a target substrate held in a quartz tube placed inside of a furnace. Typical parameters for carbon nanotube deposition include a pressure of 1 atm and a temperature of 700–900°C. Possible substrates include silicon, mica, quartz, or alumina, and the carbon supply is either CH_4 or CO gas. Common catalysts are nickel, iron, or cobalt. In a typical procedure, catalyst is sputtered, layered, or specifically placed onto the substrate. A carbon-containing gas is then passed over the substrate inside the furnace, and growth is initiated on the catalyst particles. Adding 25% hydrogen increases the SWNT yield. Advantages of the APCVD method for making carbon nanotubes include increased length and purity of the carbon nanotubes, large-scale production capability, increased control, and lower temperatures.

In an intriguing development, fast-heating CVD now offers the ability to grow millimeter-long carbon nanotubes while maintaining control over orientation of the carbon nanotubes. The fast heating minimizes the growing tubes' interaction with the substrate; this is called the *kite mechanism*. In this process, a catalyst is applied by photolithography to a silicon wafer. Using this process, Huang et al.[34] grew a square 2D network of SWNTs guided by flow of carrier gas (Fe/Mo catalyst and CO_2 gas) (Figure 7.41).

Low-Pressure CVD

Low-pressure CVD (LPCVD) at less than 10 Pa allows large numbers of wafers to be coated simultaneously without detrimental effects to film uniformity. This is the result of the large diffusion coefficient at low pressures leading to a growth limited by the rate of surface reactions rather than by the rate of mass transfer to the substrate. The surface reaction rate is very sensitive to temperature, but temperature is relatively easy to control. Typically, reactants can be used without dilution; therefore, growth rates are only an order of magnitude less than operation at atmospheric pressure allows. LPCVD in some cases can overcome the uniformity, step coverage, and particulate contamination limitations of early APCVD systems. One of the difficulties in the cold

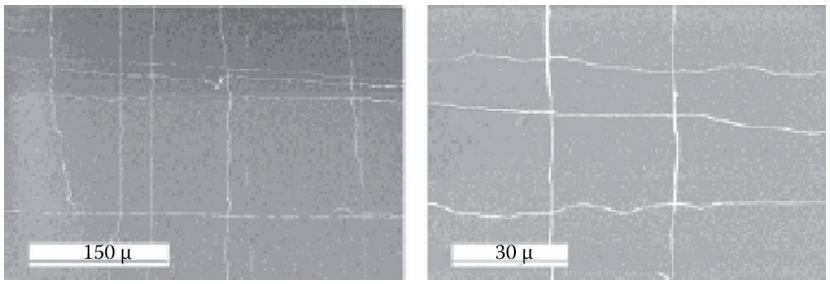


FIGURE 7.41 Square 2D network of SWNTs guided by flow of carrier gas (Fe/Mo catalyst and CO_2 gas). (Huang, S., X. Cai, and J. Liu. 2003. Growth of millimeter-long and horizontally aligned single-walled carbon nanotubes on flat substrates. *J Am Chem Soc* 125:5636–37.[34])

wall systems (where only the susceptor holding the wafer is heated) described earlier is maintaining a uniform temperature across the wafer surface. This can be overcome by putting the entire reaction chamber in a furnace maintained at a uniform temperature, as in the case of standard resistance-heated hot wall tubular furnaces. A large number of wafers can be stacked closely in these furnaces, and good uniformity across the wafer and wafer-to-wafer film thickness can be achieved if they are operated at very low pressures and in the surface reaction-limited regime. Figure 7.42 gives the schematics of a hot-wall, low-pressure CVD reactor used to deposit thin films of materials such as polycrystalline silicon, silicon dioxide, and silicon nitride. The reactor consists of a quartz tube heated by a resistance-heated furnace to maintain a uniform temperature along the reactor. Gases are introduced in one end and pumped out from the other end of the reactor. The operating pressures range from 0.25–2 Torr; temperatures range from 300–900°C; and gas flow rates range from 100–1000 sccm. A large number of wafers (~100) are stacked vertically, perpendicular to the gas flow, in a quartz holder. The inlet gases may undergo homogeneous gas-phase reactions to produce the deposition precursors, which are transported to the wafer surface by gas-phase diffusion. Excellent film uniformities (better than 5%) are obtained in these reactors. Although these reactors have lower deposition rates, it is found that this is more than compensated for by the high wafer capacity. In a hot-wall LPCVD reactor, not only the wafers get coated but also the reaction chamber walls get coated. Such systems require frequent cleaning to avoid serious particulate contamination (in situ cleaning using NF_3 or ClF_3). They find wide application because of their economy, throughput, and uniformity.

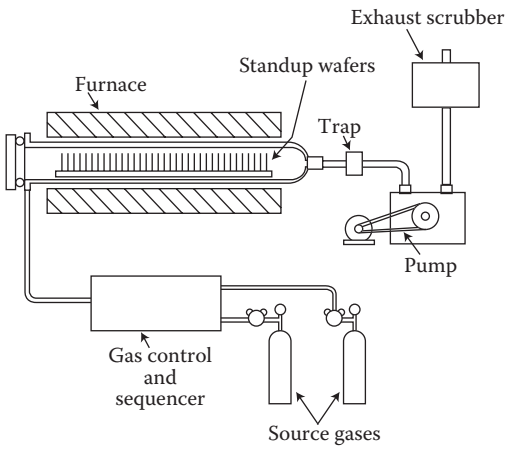


FIGURE 7.42 Low-pressure chemical vapor deposition hot wall reactor.

LPCVD polysilicon is used for structural layers in surface micromachines, and LPCVD SiO_2 and phosphosilicate glass (PSG) are used as sacrificial layers (see below). Two disadvantages of LPCVD are the low deposition rate and the relatively high operating temperatures.

Very low-pressure processes (about 1 Pa) have been used for the growth of single-crystalline Si at relatively low temperatures. Very low-pressure operation is also advantageous for the growth of III–V compound superlattices by reducing flow recirculations and improving interface abruptness.[23]

Example 7.5: Polysilicon Surface Micromachining Using LPCVD

Surface micromachining, we will see below, is based mostly on CMOS manufacturing. Alternating structural and sacrificial layers are deposited, patterned, and etched. Sacrificial

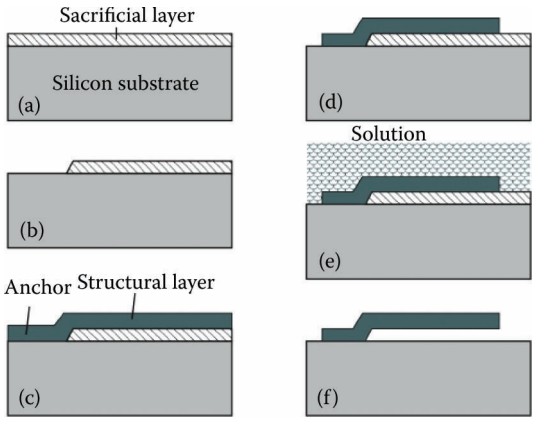


FIGURE 7.43 Basic surface micromachining process sequence. (a) Spacer layer deposition (the thin dielectric insulator layer is not shown). (b) Base patterning with mask 1. (c) Microstructure layer deposition. (d) Pattern microstructure with mask 2. (e–f) Selective etching of spacer layer.

layers are dissolved away at the end to free the structural layers so that they can be free to move as demonstrated in Figure 7.43. Materials are more or less restricted to CMOS-type materials (polycrystalline silicon, silicon oxide, silicon nitride, boron phosphosilicate glass [BPSG], phosphosilicate glass [PSG]). Structures have low aspect ratios and are sometimes referred to as 2.5D (quite planar). Polycrystalline silicon (called *polysilicon* to distinguish it from silicon having a single-crystal structure) has a number of uses in IC fabrication, including conducting materials for leads, gate electrodes in metal oxide semiconductor devices, and contact material in shallow junction devices. It is also the main structural material in surface micromachines. CVD to coat polysilicon onto a wafer involves reduction of silane at temperatures around 600°C, as expressed by the following reaction:

$$SiH_4 \rightarrow Si + 2H_2 \qquad \text{Reaction 7.4}$$

The surface micromachining process sequence depicted in Figure 7.43 is for the creation of a simple freestanding polysilicon anchor. A sacrificial layer, also called a *spacer layer* or *base*, is deposited on a silicon substrate (Figure 7.43a). CVD can be used to deposit layers of doped SiO_2, such as phosphorus-doped silicon dioxide (called *P-glass* or *PSG*). Phosphosilicate glass (PSG) deposited by LPCVD stands out as the best material for the sacrificial layer because it etches even more rapidly in HF than SiO_2. To obtain a uniform etch rate, the PSG film must be densified by heating the wafer to 950–1100°C in a furnace or a rapid thermal annealer (RTA). With a first mask, the base is patterned as shown in Figure 7.43b. A window is opened up in the sacrificial layer, and a microstructural thin film, consisting of polysilicon, is conformably deposited over the patterned sacrificial layer (Figure 7.43c). Furnace annealing of the polysilicon film at 1050°C in nitrogen for 1 h reduces stress stemming from thermal expansion coefficient mismatch and nucleation and growth of the film. Rapid thermal annealing has been found effective for reducing stress in polysilicon as well. With a second mask, the microstructure layer is patterned, usually by dry etching in $CF_4 + O_2$ or $CF_3Cl + Cl_2$ plasma (Figure 7.43d). Finally, selective wet etching of the sacrificial layer, e.g., in 49% HF, leaves a freestanding micromechanical structure (Figure 7.43e and f). The surface micromachining technique is applicable to combinations of thin films and lateral dimensions where the sacrificial layer can be etched without significant etching or attack of the microstructure, the dielectric, or the substrate. Typically, a surface micromachining stack may contain a total of four to five structural and sacrificial layers, but more are possible; the polysilicon surface machining process at Sandia's SUMMiT, for example, stacks up to five polysilicon and five oxide layers (see page 462).

Plasma-Enhanced CVD

One of the most important limitations of LPCVD is that it requires high deposition temperatures, which may be incompatible with other process steps, e.g., the deposition on an aluminum film. In plasma-enhanced CVD (PECVD), homogenous gas-phase reactions are initiated by collisions of neutrals with nonequilibrium energetic electrons produced by RF plasma instead of by thermal energy. Thus, the deposition precursors can be produced at much lower temperatures by these homogenous reactions. Furthermore, ion bombardment makes the surface more reactive by creating more dangling bonds (active sites) and/or supplying instantaneous energy at the surface for the heterogeneous reaction to take place. In general, PECVD can be carried out at a much lower temperature than LPCVD.

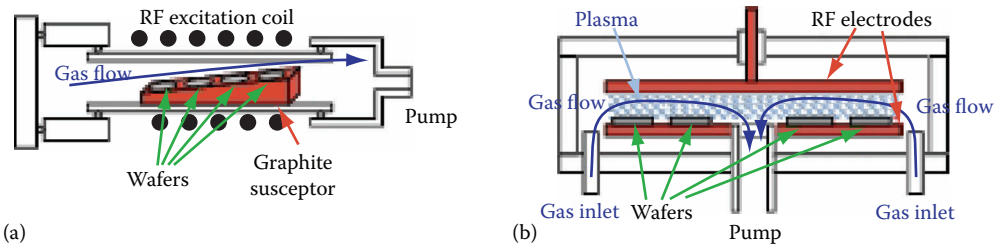


FIGURE 7.44 (a) Horizontal tube reactor. (b) Schematic of a cold wall, parallel-plate plasma deposition reactor, which is very similar to an etching reactor.

Moreover, the deposition rate is found to be higher in PECVD than in LPCVD. There are a number of different reactor configurations used for PECVD, but the physical principle is the same in all cases. Two basic PECVD setups are illustrated in Figure 7.44. In these setups, RF-induced plasma transfers energy into the reactant gases, allowing the substrate to remain at lower temperatures than in APCVD and LPCVD processes. We will only detail the cold-wall, parallel-plate plasma deposition reactor, which is very similar to an etching reactor.

With a simple parallel-plate reactor (Figure 7.44b), substrates can be placed horizontally or, if the pressure is low enough, vertically. The vertical position is used to increase throughput. No loss of film-thickness uniformity occurs in the latter case because the PECVD method is surface reaction limited. Adequate substrate temperature control ensures uniformity. The chamber consists of two electrodes, one of which is grounded, and the other is powered by an RF source. Wafers are placed on the grounded electrode and are subjected to a less-energetic bombardment than wafers placed on the powered electrode. The wafers are heated from below to the desired temperature by resistive heaters or high-intensity lamps. The reactant gases are introduced at the outer periphery, and they flow radially toward the center from where they are pumped out. Plasma is generated between the electrodes by the RF voltage applied across the electrodes. The electrode spacing is about 5–10 cm, and the operating temperature and pressure are ~400°C and 0.1–5 Torr, respectively. Homogenous reaction between the reactant gases is initiated by the electrons produced in the plasma. The ions are accelerated toward the electrodes as a result of the self-bias voltage between the plasma and the electrode. These energetic ions bombard the surface, making it more reactive, leading to higher deposition rates.

In most PECVD systems, the reactor configuration is changed so that the potential of both the powered and the grounded electrode, relative to the plasma, become equal.[27] Compared with sputter deposition, PECVD offers several advantages. The lower power densities, higher pressures, and somewhat higher substrate temperatures (e.g., >200°C) all lead to less-severe radiation damage than in sputter deposition. Moreover, for radiation-sensitive substrates such as compound semiconductors, afterglow or downstream deposition systems can be used where the radicals are formed in the glow discharge and then transported out of the region to a downstream deposition system. Thus, selective activation of reactants becomes possible without damaging the surface of the substrate.[35] PECVD films are not stoichiometric because the deposition reactions vary widely, and particle bombardment during growth of a multicomponent system changes the composition according to the ratios of sputtering yields of the component materials (nonthermal equilibrium). Because of the rapid deposition without the temperatures required for surface migration, film tends to be porous (possess holes like a sponge), leading to lower density than thermally grown films. Film stoichiometry and density can be monitored by a combination of etch rate comparisons to thermally grown films and ellipsometry to determine the index of refraction of the films.

Despite this negative consequence of particle bombardment, in general, the more ion bombardment, the better the film quality. Microstructure, stress, density, and other film properties show marked response, mostly for the better, to ion bombardment during deposition. Good adhesion, low

pinhole density, good step coverage, adequate electrical properties, and compatibility with fine linewidth pattern transfer processes have led to wide use of PECVD in very large-scale integration (VLSI). PECVD enables dielectric films such as oxides, nitrides, and oxynitrides to be deposited on wafers with small feature sizes and line widths at low temperatures and on devices unable to withstand the high temperatures of a thermally activated reaction. Planarization represents only one of the many applications of this versatile technology. Another application is deposition of amorphous silicon thin films, as used in flat-panel displays, eyeglasses, and photovoltaic panels. The most significant application is probably the deposition of SiO_2 or Si_3N_4 over metal lines.[35]

Summarizing, in PECVD, low RF excitation frequency, low reactor pressures, and low deposition rate at high temperature contribute to:

1. Lower processing temperature (250–450°C)
2. Excellent gap-fill for high-aspect-ratio gaps (with high-density plasma)
3. Good film adhesion to the wafer
4. High deposition rates
5. High film density as a result of low pinholes and voids
6. Low film stress as a result of lower processing temperature

Example 7.6: Si_3N_4 by PECVD

Silicon nitride is used as a passivation layer (protecting against sodium diffusion and moisture). Si_3N_4 has a low oxidation rate compared with silicon; therefore, a nitride mask can be used to prevent oxidation in coated areas on the silicon surface. Plasma nitride films are almost universally used for final passivation layers in IC fabrication, often in conjunction with a deposited oxide or PSG layer. A conventional CVD process for coating Si_3N_4 onto a silicon wafer involves reaction of silane and ammonia (NH_3) at around 800°C as follows:

$$3SiH_4 + 4NH_3 \rightarrow Si_3N_4 + 12H_2 \quad \text{Reaction 7.5}$$

Plasma-enhanced CVD is also used for basically the same coating reaction, the advantage being that it can be performed at a much lower temperature, around 300°C, compatible with aluminum metallization. In the latter application, stress control is important to avoid film cracking and degradation of metal reliability. Transparency in the UV must also be ensured in UV-erasable EPROM (erasable programmable read only memory) applications. PECVD silicon nitride is also often used for passivation of GaAs MESFETs [MESFET stands for metal semiconductor field effect transistor. Instead of using a p-n junction for a gate, a Schottky (metal-semiconductor) junction is used] to avoid oxidation of the exposed GaAs surface.

Properties of silicon nitride made by LPCVD versus PECVD are compared in Table 7.15.

Table 7.15 Comparison of LPCVD and PECVD Silicon Nitride

Property	LPCVD	PECVD
Deposition temperature (°C)	700–800	300–400
Composition	Si_3N_4	$Si_xN_yH_z$
Step coverage	Conformal	Fair
Stress at 23°C on silicon (dyn/cm^{-2})	$1.2–1.8 \times 10^{10}$ (tensile)	$1–8 \times 10^9$ (tensile or compressive)

Example 7.7: Polysilicon by PECVD

We remarked earlier on the quest for low-temperature polysilicon deposition processes, driven by the need for compatibility with prefabricated aluminum-metallized CMOS circuitry. This makes the 320°C PECVD polysilicon interesting (as is the ~350°C sputter deposition method).

PECVD polysilicon films, deposited in a 50-kHz parallel-plate diode reactor, can be doped in situ and crystallized by RTA (1100°C, 100 s). It was shown that small-grained PECVD films annealed by RTA have good electrical properties and gauge factors between 20 and 30, similar to those reported for other alternative types of polycrystalline silicon.[35]

A low-temperature polysilicon PECVD process is not only compatible with conventional oxide sacrificial layers (as used in surface micromachining; see below) but also with certain organic sacrificial layers such as polyimides.

More information on CVD thin films and how they are influenced by the above

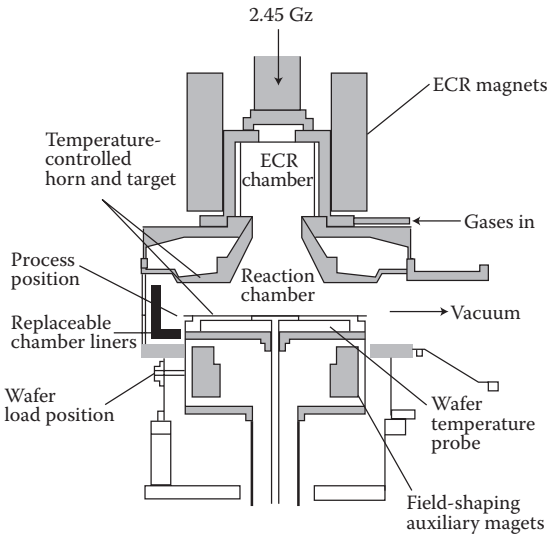


FIGURE 7.45 In ECRCVD, the plasma can deposit a film while simultaneously sputtering.

deposition parameters, as well as other parameters (e.g., moisture, flow rate, and gas composition), will be given in Volume III, Chapter 5 on selected MEMS materials.

High-Density Plasma CVD

High-density plasma sources such as inductively coupled plasma (ICP) and electron cyclotron resonance (ECR) stations (see Figure 3.36)—machines discussed in Chapter 3—can be used in high-density plasma CVD (HDPCVD) for both enhanced deposition and enhanced etching. Both systems work at low pressures (a few mTorr) and can independently control ion flux and ion energy. In ECRCVD, one uses ECR to generate a high-density plasma (HDP) that can deposit materials at high rates while pressures and temperatures remain low (Figure 7.45). ECRCVD technology provides a way to ensure durable ultrathin films. Alternating deposition and etch back is used for very effective planarization of topography. In submicrometer device fabrication, interlevel dielectric gap filling of high-aspect-ratio topography generally requires the use of multistep deposition/etch or spin-on dielectric processing to produce void-free, filled structures. With HDPCVD processing, high-aspect-ratio (up to 4:1) sub-half-micrometer structures can be filled and locally planarized in a *single* processing step. For gap filling, HDPCVD processing consists of a simultaneous deposition/etching process in which films over topographical features are sputtered off by reactive ions and radicals during deposition while deposition refills the newly created voids (see Example 7.8). The deposition/sputtering rate ratio (D/S) is an important measure of the gap-filling capability of the processes. For further reading, see Nguyen and references therein.[36] Alternative planarization techniques are reviewed in Chapter 1.

A typical ECRCVD machine is the Tek-Vac DRIE-1100-LL-ECR (http://www.tekvac.com/prd-ecr-1.htm) designed for both high-density plasma CVD and reactive ion etching of submicrometer integrated circuits (ICs), optical devices, and RF microwave electron devices. This type of equipment is used for CVD of diamond, diamond-like carbon, SiC, Si_3N_4, BN, and other refractory or high dielectric materials. The equipment provides high-density plasma with low damage and high etching rates and is well suited for etching devices sensitive to high energy electron damage, such as low-noise amplifiers. Various materials include aluminum, copper, gallium nitride, gallium arsenide, and HeCdTe, and semiconductor dielectrics have been etched this way. The method provides high uniformity of deposition and etching across large area substrates.

Example 7.8: Gap Filling by ECRCVD

ECRCVD simultaneously deposits and etches films to prevent bread-loaf and key-hole effects typical for PECVD processes, as shown in Figure 7.46. As shown in Figure 7.46b, ion-induced deposition is filling the gap between two metal lines. Argon ions then sputter-etch the excess film at the gap entrance, resulting in a beveled appearance of the film (Figure 7.46c). The etched material is redeposited, and the process is repeated, resulting in an equal "bottom-up" profile (Figure 7.46d).

Metalorganic CVD

Many materials that one wishes to deposit have too low a vapor pressure and thus are difficult to transport via gases. One solution is to chemically attach the material (e.g., Ga, Al, Cu) to an organic compound that has a high vapor pressure. Organic

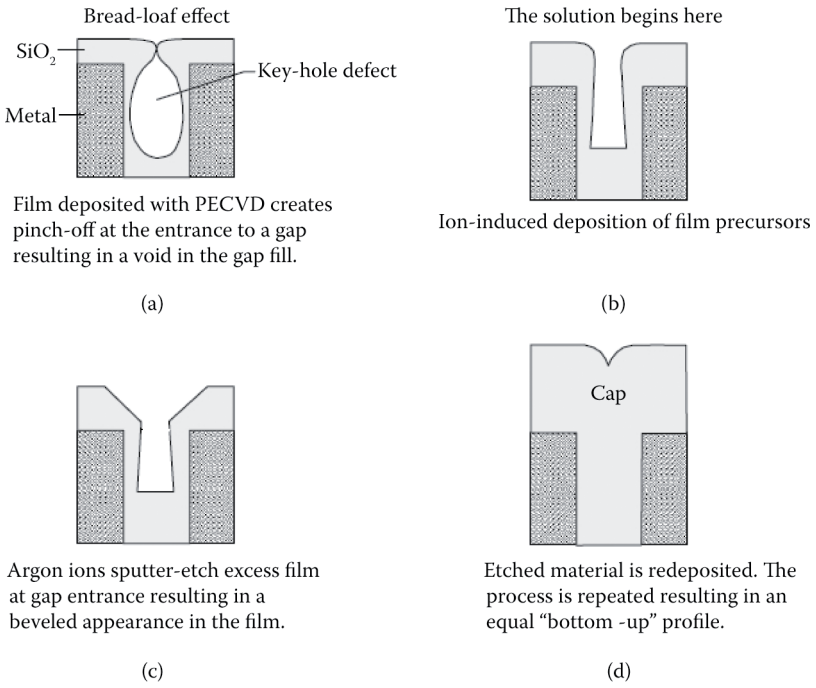


FIGURE 7.46 ECRCVD planarization.

compounds often have very high vapor pressure. The organic-metal bond is very weak and can be broken via thermal means on a heated wafer, depositing the metal with the high vapor pressure organic being pumped away. Care must be taken to ensure little of the organic byproducts are incorporated. Carbon contamination and unintentional hydrogen incorporation are sometimes a problem. Crystal growth by metalorganic CVD (MOCVD) began in the late 1960s. Metalorganic CVD, sometimes called *organometallic vapor phase epitaxy*, relies on the flow of gases (hydrides such as arsine and phosphine and organometallics such as trimethyl gallium and trimethyl aluminum) past heated samples placed in the gas stream. Typically, the reactive metal alkyls [e.g., $Ga(CH_3)_3$] react with a hydride of the nonmetal component (e.g., AsH_3) to produce the pure films (e.g., GaAs) and some byproducts (e.g., CH_4). For example:

$$(CH_3)_3Ga\ (g) + AsH_3\ (g) \rightarrow GaAs(s) + 3CH_4\ (g)$$

Reaction 7.6

The biggest advantage of MOCVD is the low growth temperature (cold wall reactor); the biggest disadvantage is the high cost of the very pure chemicals. MOCVD provides thickness control within one atomic layer and has become the preferred epitaxial process, a cost-effective manufacturing process for a variety of compound semiconductor devices. Foremost, MOCVD plays a key role in the manufacture of many optoelectronic devices with III–V compounds for solar cells, lasers, photocathodes, light-emitting diodes (LEDs), high electron mobility transistors (HEMTs), and quantum wells.[37] By using MOCVD, it is easier to scale it up to larger growth areas than it is to scale up liquid-phase epitaxy (LPE) (see below), and the uniformity and surface morphology of MOCVD grown lasers are superior to those attainable with LPE.

Atomic Layer Deposition

Atomic layer deposition (ALD) is a method of applying thin films to various substrates with atomic scale precision and constitutes one of the most promising "new" nanotechnology fabrication techniques. Dr. Tuomo Suntola and coworkers in Finland introduced ALD in 1974 to improve the quality of ZnS films used in electroluminescent displays.* It turns

* Picosun in Espoo, Finland, develops and manufactures ALD reactors for microelectronics and nanotechnology. With the help of Dr. Tuomo Suntola, Picosun has succeeded in developing inexpensive yet highly efficient ALD tools for research and development.

Table 7.16 List of ALD Applications

- High-k dielectrics (Al_2O_3, HfO_2, ZrO_2, Ta_2O_5, La_2O_3, etc.) for transistor gates and DRAM capacitors in Si, heterostructures, compound semiconductors, MESFETs, III–V semiconductor materials, organic transistors, graphene, graphite, nanotubes, nanowires, and molecular electronics
- Conductive gate electrodes (Ir, Pt, Ru, TiN, etc.)
- Metal interconnects and liners (Cu, WN, TaN, WNC, Ru, Ir, etc.); metallic diffusion barrier layers for copper interconnects and semiconductor vias for transistor gate and memory cell applications, DRAM capacitors, and passivation layers
- Catalytic materials (Pt, Ir, Co, TiO_2, V_2O_5); coatings inside filters, membranes, catalysts (thin economical Pt for automobile catalytic converters), fuel cells, ion exchange coatings
- Nanostructures (all materials); conformal deposition around and inside nanostructures and MEMS
- Biomedical coatings: (TiN, ZrN, CrN, TiAlN, AlTiN, etc.); biocompatible materials for in vivo medical devices and instruments
- Piezoelectric layers (ZnO, AlN, ZnS, etc.)
- Transparent electrical conductors (ZnO:Al, ITO)
- UV blocking layers (ZnO, TiO_2)
- OLED passivation (Al_2O_3)
- Solid lubricant layers (WS2)
- Photonic crystals (ZnO, ZnS:Mn, TiO_2, Ta_2N_5); coatings inside porous alumina, inverted opals
- Antireflection and optical filters (Al_2O_3, ZnS, SnO_2, Ta_2O_5); Fabry-Perot, flip-flop optical filters
- Electroluminescent devices (SrS:Cu, ZnS:Mn, ZnS:Tb, SrS:Ce)
- Processing layers (Al_2O_3, ZrO_2); etch barriers, ion diffusion barriers, fill layers for magnetic read heads
- Optical applications (AlTiO, SnO_2, ZnO); photonics, nanophotonics, solar cells, integrated optics, optical coatings, lasers, variable dielectric constant nanolaminates
- Sensors (SnO_2, Ta_2O_5, etc.); gas sensors, pH sensors
- Wear and corrosion inhibiting layers (Al_2O_3, ZrO_2)

out that ALD also produces outstanding high-k* dielectrics, metal electrodes, barriers, and spacers (see list of applications in Table 7.16). Objects of almost any size and shape can be coated, even atomic force microscope (AFM) tips. The method is similar in chemistry to CVD, except that the ALD reaction breaks up the CVD reaction into two half-reactions, keeping the precursor reagents separate during their reaction at the surface. As opposed to conventional CVD where one has continuous deposition and concurrent flow of precursors, ALD is based on the sequential deposition of individual monolayers at the surface in a well-controlled manner: the growth surface is alternately exposed to only one of two complementary chemical environments, i.e., individual precursors are supplied to the reactor one at a time. The precursors need to be thermally stable and should chemisorb fast and react aggressively with each other. Between exposure steps, an inert gas purge or a pump-down step is used to remove any residual chemically active source gas or byproducts before another precursor is introduced into the reactor. The overall ALD process then consists of a repetition of a number of growth cycles. Each cycle is made up of a typical sequence: 1) flow of precursor 1, 2) purge, 3) flow of precursor 2 and 4) purge. During a precursor exposure step, precursor molecules react with the surface until all available surface sites are saturated. Precursor chemistries and process conditions are chosen such that no further reaction takes place once the surface is completely saturated. This surface saturation guarantees the self-limiting nature of ALD. Precursors are preferably overdosed so that process results become independent of potential slight variations in the amount of precursor supplied to the surface. Thus, surface chemistry governs film growth rather than a precise control of tool-specific process parameters such as precursor flow and partial pressure. A known and constant thickness is deposited per growth cycle. Typically, deposition rates on the order of 0.1–1.0 Å/cycle are obtained, with cycle times ranging from 1–10 s. Cycle time critically depends on saturation behavior, chamber volume, and reactor design. The film is grown in a layer-by-layer mode, and the total film thickness is given by the number of cycles. Exceptional across-wafer thickness uniformity and excellent step coverage on high-aspect-ratio features can be achieved with ALD, making this technique one of the most exciting nanotechnology methods. In Figure 7.47 we illustrate a typical ALD setup, and in Figure 7.48 the ALD cycle for Al_2O_3 deposition is demonstrated.

* Many materials are currently under consideration as potential candidates for gate dielectrics for sub-0.1-μm CMOS technology. Silicon dioxide with a dielectric constant of 3.9 will be replaced by so-called *high-k materials*. Candidates for gate dielectrics include the oxides of yttrium, lanthanum, zirconium, and hafnium, which have much higher dielectric constants than SiO_2.

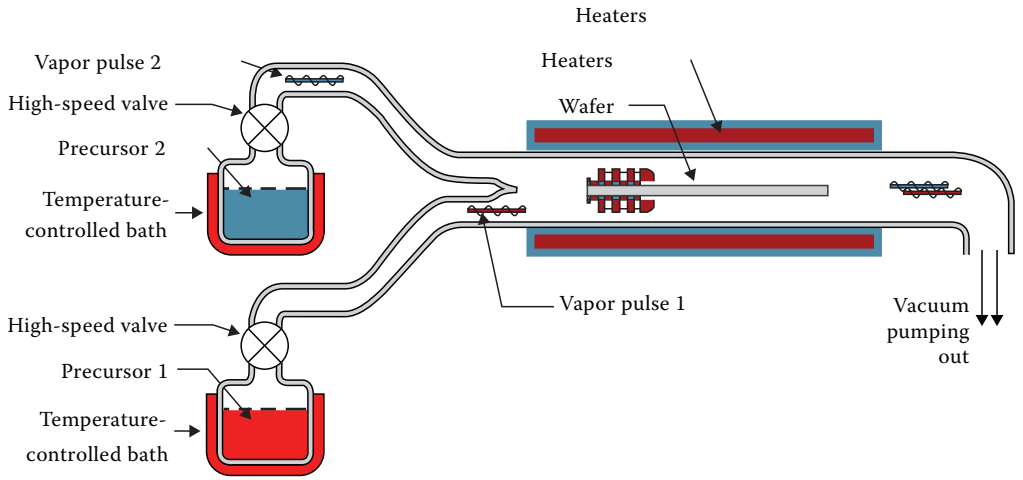
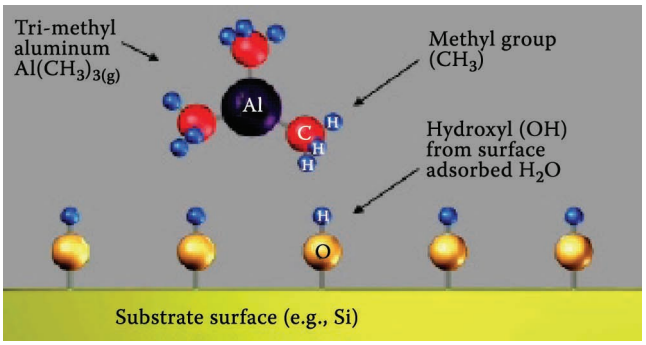
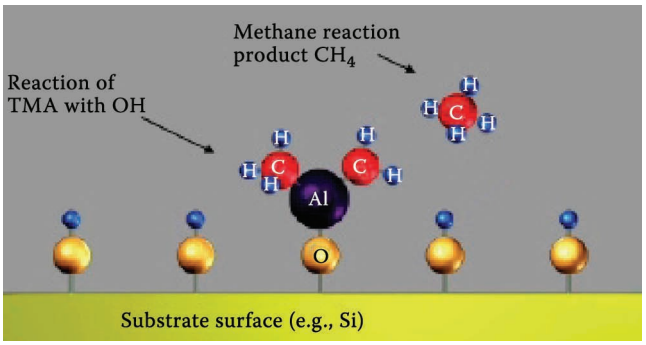


FIGURE 7.47 Typical atomic layer deposition (ALD) setup. (From *Atomic layer deposition.* Cambridge NanoTech Inc., April 24, 2006. http://www.cambridgenanotech.com.[231])

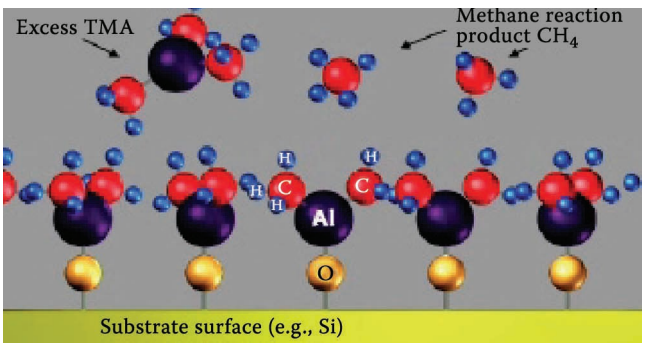


1. Flow of precursor 1: In air, water vapor is adsorbed on most surfaces, forming a hydroxyl group. In the case of Si this forms Si-OH. After placing the substrate in the reactor, trimethyl aluminum (TMA) is pulsed into the reaction chamber.



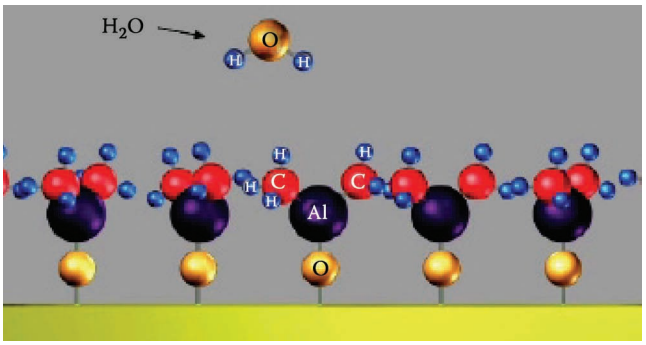
TMA reacts with the adsorbed hydroxyl groups, producing methane as a reaction product.

$$Al(CH_3)_3(g) + :Si-O-H(s) \rightarrow :Si-O-Al(CH_3)_2(s) + CH_4(g) \qquad \text{Reaction 7.7}$$

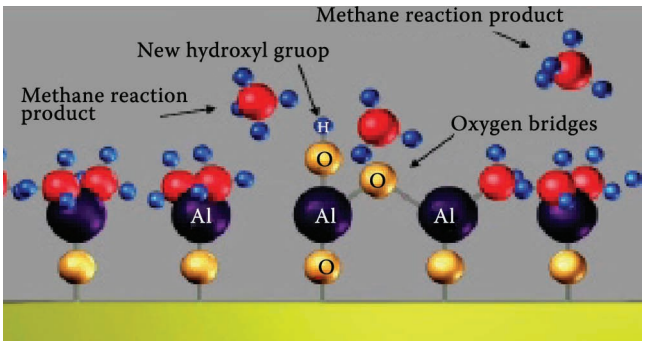


2. Purge cycle: TMA reacts with the adsorbed hydroxyl groups until the surface is passivated. TMA does not react with itself, limiting the reaction to one monolayer. This leads to the perfect uniformity and conformality of the ADL process. In the purge pulse, excess TMA and the methane reaction product are pumped away.

FIGURE 7.48 The atomic layer deposition (ALD) process illustrated for the Al_2O_3 process. (From *Atomic layer deposition.* Cambridge NanoTech Inc., April 24, 2006. http://www.cambridgenanotech.com.[231])

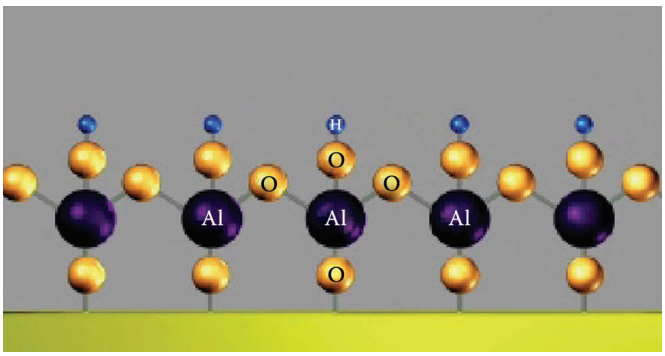


3. Flow of precursor 2: After all the TMA and methane have been pumped away, water is pulsed into the reactor chamber.

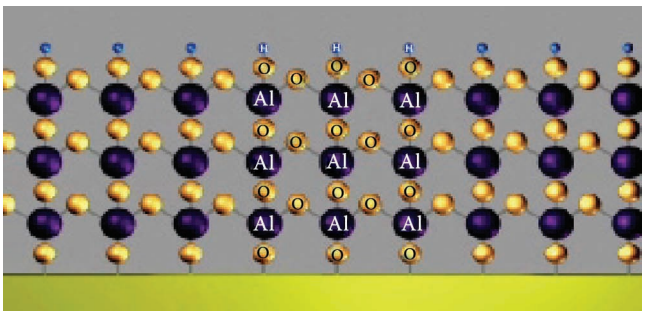


Water reacts with the dangling methyl groups on the new surface, forming aluminum-oxygen (Al-O) bridges and hydroxyl surface groups waiting for a new TMA pulse.

$$2H_2O(g) + :Si{-}O{-}Al(CH_3)_2(s) \rightarrow :Si{-}O{-}Al(OH)_2(s) + 2CH_4(g) \qquad \text{Reaction 7.8}$$



4. New purge cycle: The reaction product methane is pumped away. Excess water does not react with the hydroxyl surface groups, again causing perfect passivation limited to one monolayer.



One TMA pulse, one purge, and one water pulse form one cycle. Here three cycles are shown, with approximately 1 Å/cycle. Each cycle, including pulsing and purging, takes about 3 s.

FIGURE 7.48 (*Continued*)

Table 7.17 Comparison of ALD with CVD

ALD	CVD
Highly reactive precursors	Less reactive precursors
Slower deposition	Faster deposition
Precursors react separately on the substrate	Precursors react at the same time on the substrate
Low-temperature deposition possible	Higher deposition temperature
Precursors must not decompose at process temperature	Precursors can decompose at process temperature
Uniformity ensured by the saturation mechanism	Uniformity requires uniform flux of reactant and temperature
Stoichiometric films with large area uniformity and 3D conformality	
Gentle deposition process for sensitive substrates	
Thickness control by counting the number of reaction cycles	Thickness control by precise process control and monitoring
Surplus precursor dosing acceptable	Precursor dosing important

The list of applications for ALD keeps expanding, as evident from Table 7.16. In Table 7.17 we compare ALD with CVD.

Epitaxy

Introduction

Epitaxial techniques arrange atoms in single-crystal fashion on a crystalline substrate acting as a seed crystal so that the lattice of the newly grown film duplicates that of the substrate. The lattice constant of the epitaxially grown layer needs to be close to the lattice constant of the substrate wafer. Otherwise, the bonds cannot stretch far enough, and dislocations will result.

Si epitaxy is used for high-purity layer growth and can form very thick doped structures (30–100 μm) not possible with implantation or diffusion. Such thick, pure layers are often used in power devices with thinner, 1–5-μm structures, commonly used for CMOS and bipolar technology. In surface micromachining, where polycrystalline silicon functional layers are built up on a substrate rather than etched in the bulk of a single-crystal substrate as in bulk micromachining, epi-grown SOI wafers are a very attractive alternative. Surface micromachines built from layers of epi silicon and isolated from the substrate by a SiO_2 layer combine the most attractive features of surface micromachining (i.e., CMOS compatibility and freedom in types of structural shapes) with the superior single-crystal properties of the epilayer. For ICs, SOI is compatible with existing wafer processes, provides 50% faster circuit speed than bulk silicon, allows easier scaling to finer line widths, and can reduce the number of required mask levels for a given design by ~30%. SOI may very well represent the wave of the future in the IC industry rather than GaAs.[38] Silicon micromachining has typically followed in the footsteps of the IC industry; therefore, the prevalence of SOI in IC is a good indicator of coming trends in micromachining. SOI simplifies building micromachines by reducing the number of masks, reducing packaging concerns, and making integration of electronics easier. In the IC industry, many device parameters such as transistor breakdown voltage, junction capacitance, transistor gain, and AC performance depend on the epilayer thickness, thus necessitating precise control of it. This precise control also enables more reproducible and predictable mechanical micromachines.

Liquid- and Solid-Phase Epitaxy

In the section on PVD techniques we discussed molecular beam epitaxy (MBE) as the most advanced. In the CVD section we encountered two types of vapor phase epitaxy (VPE) techniques, APCVD and MOCVD. For silicon processing, VPE has met with the widest acceptance because excellent control of impurity concentration and crystalline perfection can be achieved. In growing an epilayer with VPE, the processing is accomplished under closely controlled conditions at higher temperatures than conventional CVD of Si, using diluted reacting gases to slow the process so that an epitaxial layer can be successfully formed. Here, we briefly discuss liquid- and solid-phase epitaxy. For depositing multilayer structures

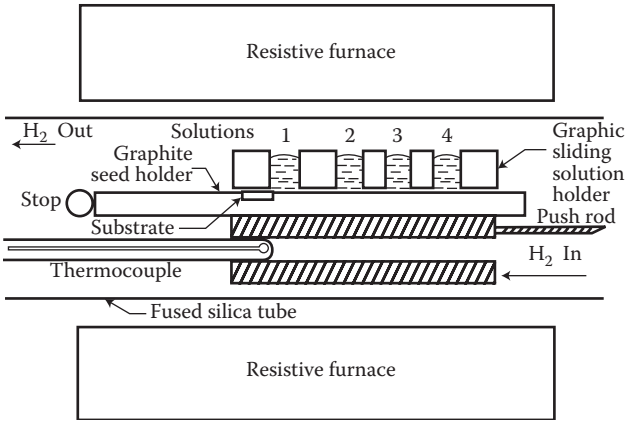


FIGURE 7.49 Liquid-phase epitaxy setup. (From Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press. With permission.[1])

of different materials on the same substrate, liquid phase epitaxy (LPE) is used. LPE films grow from a liquid solution very near the equilibrium state, making the technique reproducible and resulting in films with low concentrations of growth-induced defects. A schematic for a typical LPE setup is presented in Figure 7.49.[1] In operation, a graphite slider plate moves relative to a multiple-well assembly to bring the substrate in a recess in the slider plate in contact with the different solutions. LPE has found its widest application in producing epitaxial layers of III–V compounds (e.g., InP, GaAs) for lasers and photodetectors. Anderson reviews the technique well.[39]

Solid-phase epitaxy describes the crystalline regrowth of amorphous layers that extend continuously to the underlying single-crystal substrate. At temperatures between 500 and 600°C, in the case of silicon, a recrystallization process occurs on the underlying crystalline substrate, and regrowth proceeds toward the surface. Regrowth is faster on (100) than on (111) Si, and impurities such as B, P, and As enhance the regrowth, whereas O, C, N, and Ar retard it.[27]

Selective Epitaxy

The incorporation of selective epi in the micromachining arsenal is making more versatile microstructures possible. Under the correct growth conditions and/or surface treatment, it is possible to initiate Si growth in selected areas.[27,40] Selective epitaxial growth allows the formation of closely spaced silicon features isolated by SiO_2 (see Figure 7.50). Besides increased density over other insulation techniques (important mainly for the IC industry), structures as shown in Figure 7.50a and b, might enable a host of interesting mechanical microstructures. For example, imagine that the SiO_2 in Figure 7.50b is selectively etched away, resulting in an epi anchor with suspended polybeams. In selective epitaxy of the type shown in Figure 7.50a, silicon atoms possessing high surface mobility migrate to sites on the single crystal where nucleation is favored. In the ideal case, all the epi grows exclusively in the oxide openings. Silicon mobility improves in the presence of halides. The higher the number of chlorine atoms in the silicon source, the better the degree of selectivity (e.g., $SiHCl_3$ is a better source than SiH_2Cl_2). The selective deposition shown in Figure 7.50b is accomplished by simultaneous deposition of epitaxial silicon in the oxide openings and polysilicon on the oxide surfaces. The angle between the epi and the poly in Figure 7.50b depends on the crystallographic orientation of the

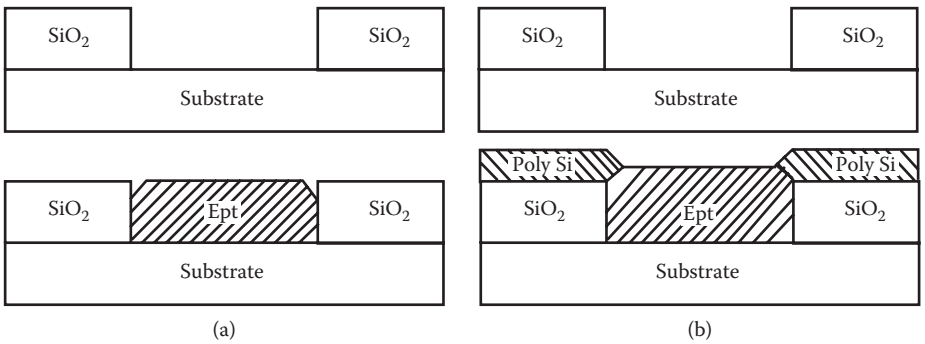


FIGURE 7.50 Selective deposition of epitaxial silicon. (a) Selective deposition of epi Si on Si in SiO_2 windows. (b) Simultaneous deposition of epi Si on Si, and polysilicon on SiO_2. (From Wolf, S., and R. N. Tauber. 1987. *Silicon processing for the VLSI era*. Sunset Beach, CA: Lattice Press.[27])

substrate. The angle is 90° for <110> orientations, 72° toward the polysilicon for <100> orientations, or tapered toward the single-crystal silicon at 70° for the <111> orientation.[27] These figures represent interesting angles from which to construct micromechanical structures. This selective deposition process also lends itself to CVD, for example, for selective deposition of W.

Epilayer Thickness

Epitaxial layer thickness is a critical parameter both in IC applications and micromachines and therefore must be accurately measured and controlled. Many devices, from discrete transistors and 16-megabyte dynamic random access memory (DRAM) to membrane-based micromechanical structures, use silicon epitaxial layer wafers as their starting material. The thickness of an epitaxial layer forms an integral part in the design of many micromachined devices. For example, in a piezoresistive pressure sensor, epilayer thickness control ultimately determines the pressure sensitivity control.

Epilayer thickness can be measured from infrared reflectance, angle lap and stain, tapered groove, weighing, capacitance-voltage measurements, and profilometry. The most widely used, nondestructive method of measuring epi thickness is with infrared instruments. Fourier transform infrared offers automated epilayer thickness measurements.[41]

CVD Equipment

For a detailed historical timeline of vacuum coating and vacuum/plasma technology, visit http://www.svc.org/H/H_HistoricalTime.html, posted on the Internet by the Society of Vacuum Coaters.

Typical sources for CVD equipment are CVD Equipment Corporation (http://www.cvdequipment.com) and IonBond Inc. (http://www.ionbond.com). Commercial epitaxial services offer layers ranging from 0.5–150 μm and N, P, N^+, and P^+ type with uniformity better than ±5%.

Planar (http://www.planar.com) and Cambridge NanoTech (http://www.cambridgenanotech.com) manufacture atomic layer deposition (ALD) systems for many coating applications. Vacuum equipment to clean, deposit, and etch is combined more and more in so-called *cluster tools* (Figure 7.51).[42]

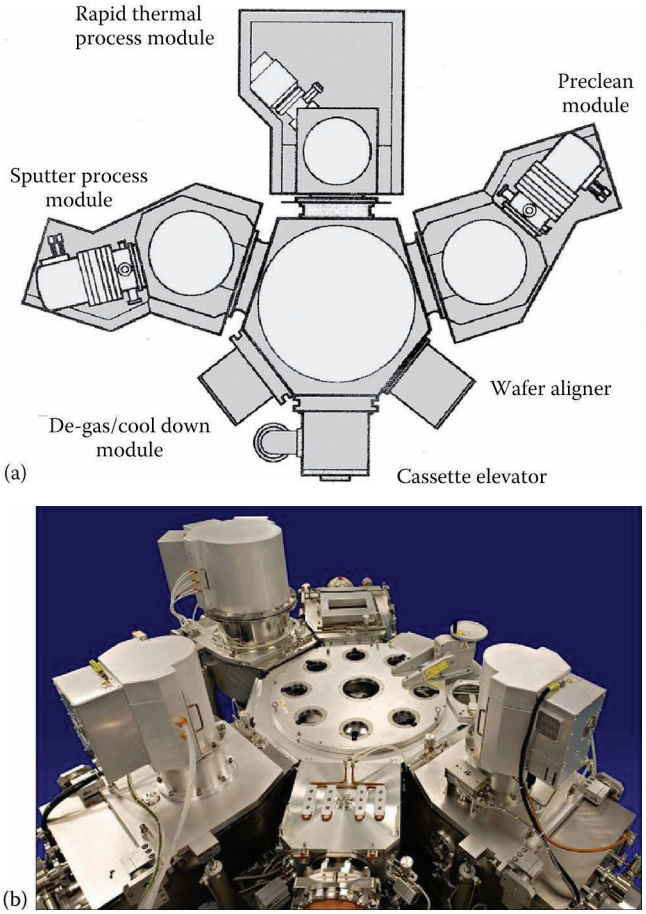


FIGURE 7.51 Cluster tools. (a) Schematic. (b) Example from Ovonyx (http://www.ovonyx.com/corporate).

Selection criteria for the additive processes reviewed in this chapter depend on a variety of considerations, such as:

1. Limitations imposed by the substrate or the mask material: T_{max} (maximum temperature), surface morphology, substrate structure and geometry, and so on
2. Apparatus requirement and availability
3. Limitations imposed by the material to be deposited: chemistry, purity, thickness, T_{max}, morphology, crystal structure, and so on
4. Rate of deposition to obtain the desired film quality
5. Adhesion of deposit to the substrate; necessity of adhesion layer or buffer layer
6. Total running time, including setup time and postcoating processes
7. Cost
8. Ease of automation
9. Safety and ecological considerations

Mechanical Properties of Thin Films

Introduction

The study of the mechanical behavior of thin films started in the 1950s. This early work was initiated to better understand the stresses and failures in integrated circuit (IC) structures. The MEMS field, especially surface micromachining, provided an additional impetus in the late 1980s, and the investigations were extended to more general applications of thin films. In this section, advances in the understanding of mechanical properties of thin films are reviewed with special focus on the mechanical behavior of polycrystalline thin films for MEMS and NEMS applications.

Thin films in surface micromachines must satisfy a large set of rigorous chemical, structural, mechanical, and electrical requirements. Excellent adhesion, low residual stress, low pinhole density, good mechanical strength, and chemical resistance all may be required simultaneously. For many microelectronic thin films, the material properties depend strongly on the details of the deposition process and the growth conditions, and some properties may depend on postdeposition thermal processing, referred to as *annealing*. Furthermore, the details of thin-film nucleation and/or growth may depend on the specific substrate or on the specific surface orientation of the substrate. Although the properties of a bulk material might be well characterized, its thin-film form may have properties substantially different from those of the bulk. For example, thin films generally display smaller grain size than bulk materials. An overwhelming reason for the many differences stem from the properties of thin films, which exhibit a higher surface-to-volume ratio than large chunks of material and are strongly influenced by the surface properties.

Examples where thin films are directly "mechanically challenged" include the myriad of structural and moving parts in MEMS devices, such as the membranes in pressure sensors, hinges in micromirror arrays, wear-resistance coatings on cutting tools, and protective coatings for magnetic disks. In many other applications, thin films are selected primarily because of their unique electronic, magnetic, optical, or thermal properties; however, the mechanical characteristics of the materials of choice in such applications are not unimportant because thin films are often subjected to large mechanical stresses during both the manufacturing process and normal operation of the end-use devices.

Some terminology characterizing thin films and their deposition is introduced in Table 7.18.

Because thin films were originally not intended for load-bearing applications, their mechanical properties were ignored for a long time. As we indicated earlier, it was the IC and MEMS world that

Table 7.18 Thin-Film Terms Used in Characterizing Deposition

Term	Remark
Film	Bond energy <10 kcal/mol
Chemisorbed film	Bond energy >20 kcal/mol
Nucleation	Adatoms forming stable clusters
Condensation	Initial formation of nuclei
Island formation	Nuclei grow in three dimensions, especially along the substrate surface
Coalescence	Nuclei contact each other and larger, rounded shapes form
Secondary nucleation	Areas between islands are filled in by secondary nucleation, resulting in a continuous film
Grain size of thin film	Generally smaller than for bulk materials and function of deposition and annealing conditions (higher T, larger grains)
Surface roughness	Lower at high temperatures except when crystallization starts; at low temperature the roughness is higher for thicker films, also oblique deposition and contamination increases roughness
Epitaxial and amorphous films	Very low surface roughness
Density	More porous deposits are less dense; density reveals a lot about the film structure
Crystallographic structure	Adatom mobility: amorphous, polycrystalline, single crystal or fiber texture, or preferred orientation

precipitated a strong appreciation for the need of an understanding of the mechanical properties as essential for improving the reliability and lifetime of thin films, even in nonstructural applications.[43] Generally, mechanical stresses in thin films can be divided into intrinsic or growth stresses that develop during the deposition process and extrinsic stresses that are induced by external physical effects. The extrinsic stresses are induced by various physical effects after the film is grown.

Thin-film materials are fabricated using very different processing methods than those for bulk materials, such as the various CVD and PVD deposition techniques studied in the current chapter. These processing methods often result in development of intrinsic mechanical stresses in the films. We saw earlier that metal films sputtered at room temperature typically consist of very fine grains, whereas evaporated films are often highly textured with larger grain sizes than sputtered films. It was also pointed out that mobility of the atoms of the deposited material significantly affects the microstructure formation in both deposition and postdeposition processes. The microstructure evolution during the growth process typically involves 1) nucleation of crystallite islands from the condensed material atoms at many sites on the substrate surface, 2) growth of individual crystallite islands until they coalescence with other islands, and 3) coarsening of grains during thickening of the film. For bulk materials, the crystal grain growth occurs at the expense of small grains through the motion of grain boundaries, driven by the reduction in the total grain boundary energy. In thin films the situation is a bit more complex because free surfaces and interfaces also play an important role in the microstructure development. Mismatch strains often develop between the film and the substrate, and this leads to higher strain energy in the film. Grain growth and crystallographic texture development in thin films are driven by the reduction of the total energy, including the grain boundary energy, the surface and interface energy, and the strain energy. This is a thermally activated kinetic process, and for films grown at room temperature, grain growth and texture development are often slow because of lack of thermal activation, resulting in a metastable structure with very fine grains (see section on sputtered films). The microstructure of thin films can be further modified through a postdeposition process, such as annealing. During annealing, grains grow and crystallographic textures develop to minimize the total energy at a rate determined by the annealing temperature. When a grain grows to a size on the order of the film's thickness, the grain boundaries start intersecting the film surface, forming grooves at the surface. These grooves suppress further growth of the grain. As a result, a columnar grain structure with grain boundaries traversing the film thickness results. Thus, the grain size in the plane of the film is on the order of the film thickness.

There have been many mechanisms proposed for the development of growth stresses, which depend sensitively on the material systems, deposition technique, and process parameters. Mechanics at small length scales in micro- and nanotechnology is a challenge for those who recognize its existence and want to optimize the mechanical performance of micro- and nanosystems and a threat for those who ignore it. We will see in Volume III, Chapter 6 on metrology and analytical techniques that surface micromachined devices contribute strongly to the understanding of mechanical properties of thin films.

Strength of Thin Films

Most thin film properties can be simply explained through extrapolation of bulk behavior, but other properties depend on mechanisms that remain unexplained. The strength of thin films is such an example. A substantial body of work has shown that thin metal films can support much higher stresses than the same material in bulk form. This strengthening has generally been attributed to dimensional and microstructural constraints on dislocation activity in these films. Dimensional constraints are imposed by the interfaces and the small dimensions typically encountered in thin films, whereas microstructural constraints arise from the very fine grains that often typify thin films. In bulk materials, microstructural constraints dominate the plastic behavior of the material. However, when material dimensions are comparable to microstructural length scales—as is typically the case for thin films—free surfaces and interfaces play an important role as well. For example, dislocations can exit the material through free

surfaces, whereas interfaces may prevent them from doing so. Consequently, strong interfaces lead to a higher cumulative dislocation density in the film and result in a higher flow stress and a greater strain-hardening rate. In addition to dimensional constraints, the microstructure of thin films also affects their mechanical properties. Because thin films are fabricated using different techniques than those for bulk materials, the microstructure of thin films is often very different from that of bulk materials.

Let us remember what controls the strength of a material. Based on their bond strength most materials should be much stronger than they are in practice. For example, the strength for an ionic bond should be about 10^6 psi, but a more typical strength for an ionic material is 40 10^3 psi. The reason, as we saw in Volume I, Chapter 2, is that materials do not usually fail by breaking bonds. Real materials have lots of dislocations; therefore, the strength of the material depends on the force required to make the dislocation move, not the bonding energy. When dislocations move, slip occurs and the direction of the movement is the same as the Burgers vector (= the deformation direction, e.g., for an edge dislocation it is perpendicular to the dislocation line, and for a screw dislocation it is parallel to the line of the dislocation). When dislocations run into each other a traffic jam effect results; thus, more dislocations increase the strength of a material. Dislocations may be pinned by solutes, interstitials, and precipitates. An example of a solute is Zn atoms in a Cu matrix: the bigger Zn atoms make the slip plane "rougher," and thus increase the resistance to dislocation motion. Precipitates (small particles) also can promote strengthening by impeding dislocation motion. Applying a force to the material increases the number of dislocations, and this is called *strain hardening* or *cold work*. Work hardening accounts for the higher strength of cold rolled steel. Grain boundaries act to impede dislocation motion because the slip systems in adjoining grains will usually not line up; increases in yield strength arising from this mechanism are called *boundary strengthening*. In a small grain, a dislocation gets to the boundary and stops. In a large grain, the dislocation can travel farther. Therefore, a small grain size equates to more strength. Hall and Petch, working independently around 1950, found that the yield strength of a polycrystalline material follows the following dependence:

$$\sigma_Y = \sigma_0 + \frac{K}{\sqrt{d}}$$

(Volume I, Equation 2.70)

where σ_0 is the lattice friction stress needed to move dislocations and K is a constant (e.g., the elastic limit of copper doubles when the grain size decreases from 100 μm to 25 μm). This relation is essentially empirical, but it can be rationalized by viewing the second term as related to the stress needed to activate a new mobile dislocation in an unfavorably oriented grain. The Hall-Petch relation is illustrated in Figure 7.52 for steel at room temperature. The grain size is often given in ASTM grain size number (n)—a measure of the size of the grains in a crystalline material obtained by counting the number of grains per square inch at a magnification × 100. Instead of yield strength one might also plot the hardness (H) of the material in Figure 7.52. Indentation (hardness) testing is very common for bulk materials where the direct relationship between bulk hardness and yield strength is well known. Pressing a hard, specially shaped point into the surface and observing the indentation can measure hardness. This type of measurement is of little use for measuring thin films less than 5 × 10^4 Å. Consequently, very little is known

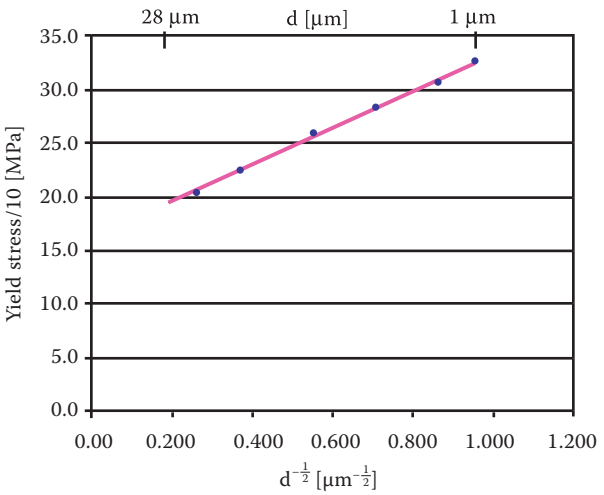


FIGURE 7.52 The Hall-Petch equation—the relationship between yield strength and grain size—in steel at room temperature.

about the hardness of thin films. Recently, specialized instruments have been constructed (e.g., the nanoindenter) in which load and displacement data are collected while the indentation is introduced in a thin film. This eliminates the errors associated with later measurement of indentation size and provides continuous monitoring of load/displacement data similar to a standard tensile test. Load resolution may be 0.25 μN, displacement resolution 0.2–0.4 nm, and *x*-*y* sample position accuracy 0.5 μm. Empirical relations have correlated hardness with Young's modulus and with uniaxial strength of thin films. Hardness calculations must include both plastic and long-distance elastic deformation. If the indentation is deeper than 10% of the film, corrections for elastic hardness contribution of the substrate must also be included.[43] Mechanical properties such as hardness and modulus of elasticity can also be determined on the micro- to picoscales using an atomic force microscope (AFM).[44] Bushan[45] provides an excellent introduction to this field in the *Handbook of Micro/Nanotribology*.

The Hall-Petch relationship has been well established experimentally for grain sizes in the millimeter through submicrometer regimes. Consequently, it was thought that nanosized grains would produce materials with even greater mechanical integrity. The above relation indicates that simply by decreasing the grain size one can make a material arbitrarily strong! Of course, this is not sensible because a polycrystalline material with grain size of a single atom would just be amorphous or a single crystal. In other words, there must be a limit where Hall-Petch breaks down! It is easily recognized that the Hall-Petch behavior is expected to break down, for example, if the smallest dislocation loop no longer fits in a grain (Figure 7.53). For very small grain

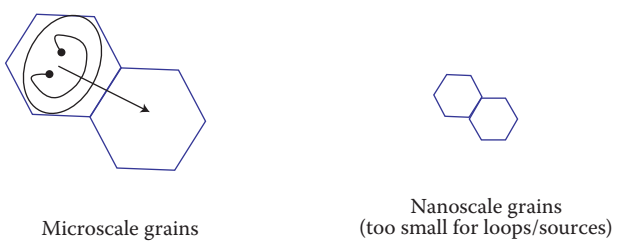


FIGURE 7.53 Hall-Petch behavior breaks down if the smallest dislocation loop no longer fits in a grain.

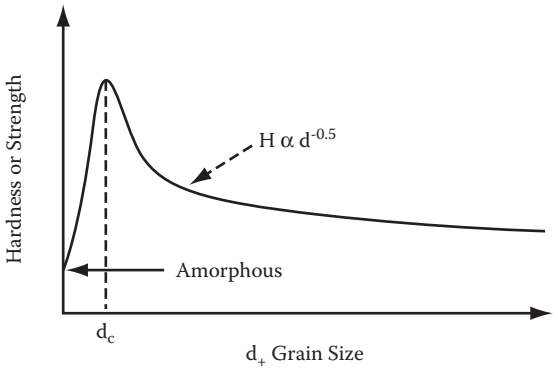


FIGURE 7.54 Hall-Petch and reverse Hall-Petch.

sizes one can perform atomistic simulations of the deformation, and these simulations suggest that for very small grain sizes there is a reverse Hall-Petch effect, i.e., the strength decreases with increasing grain size; in other words, an optimal grain size (d_c) exists as suggested by the plot in Figure 7.54. The classic Hall-Petch relationship is based on the idea that grain boundaries act as obstacles to dislocation movement and that dislocations require greater amounts of energy to overcome these barriers. Because dislocations are carriers of plastic deformation, this manifests itself macroscopically as an increase in material strength. The deformation mechanism for materials with very small grains (<20 nm) is different, and it has been suggested that plastic deformation in this case is no longer dominated by dislocation motion but by atomic sliding of grain boundaries.[46] For very small grain sizes, this sliding effect might conceivably constitute the dominant effect because of the larger ratio of grain boundary to crystal lattice, and this mode of deformation could then lead to the observed softening of a nanomaterial. Other explanations to rationalize the apparent softening of metals with nanosized grains include poor sample quality and the suppression of dislocation pileups.[47]

A reverse Hall-Petch behavior has been observed experimentally in nanocrystalline materials with sufficiently small grain sizes.[48] There are also observations of "kinked" Hall-Petch graphs, i.e., cases where the slope is reduced (but still positive) below a certain grain size. However, today there is still a lot of uncertainty about the exact interpretation of a reverse Hall-Petch effect. It is possible that many of the early measurements of a reverse Hall-Petch

effect were likely the result of unrecognized pores in the samples. The presence of voids in nanocrystalline metals would undoubtedly lead to their having weaker mechanical properties. It is not entirely clear what exactly the dependency of yield stress should be on grain sizes less than a diameter d of 20 nm. Nanocrystalline materials have been the subject of widespread research in recent decades; since the review by Gleiter[49] in 1989, thousands of papers have been published on the topic.

The yield strength of metallic films is affected by several microstructural dimensions: grain size, film thickness, free-standing or deposited on a substrate, and obstacle spacing (e.g., precipitate concentration). A common approach to the quantification of the role of different parts of the microstructure to the yield stress of a deposited film is to assume that their contributions have a power-law form and can be superimposed:

$$\sigma_Y = \sigma_0 + Kd^{-n} + K't^{-m} \qquad (7.22)$$

where the first term represents the bulk yield stress for a large-grained polycrystal, the second represents the contribution from the grain boundaries (with d the grain size), and the third represents the contribution from the film surface or interface with t, the film thickness. The first two terms together form the well-known Hall-Petch relation, in which, commonly, n = 0.5.[50] For thin films with very fine grains ($d \ll t$) or in multilayers with very thin constituent layers, the deposition of individual dislocation segments becomes more important than the formation of pileups, which makes n approach 1. In several studies of thin films deposited on a substrate, an increase of the yield strength with decreasing film thickness has been observed. For example, for thin films of copper, with thickness between 0.1 and 3 μm, vapor deposited on 12.7- or 7.6-μm-thick polyimide Kapton substrates, Yu et al.[50] find that the yield stress depends strongly on the film thickness and is fit by $\sigma_Y = 116 + 355t^{-0.473}$ where t is the thickness in meters and σ_Y is in MPa. The origin of the hardening induced by the finite size of the film is the deposition of dislocation segments at the film surface if oxidized or otherwise passivated and at the film substrate interface. The blocking of the dislocations at the film substrate interface leads to an inhomogeneous strain distribution, in which the plastic strain approaches zero near the interface.

For free-standing films, the effect of film thickness is reversed, with thinner films being weaker. It has been found that grain-size effects are diminished when Cu test sample size decreases to less than about 1 mm, an observation that has important consequences for micromanufacturing. When the grain size d varied between 18 μm and 200 μm, it was found that strength is reduced with an increase in d/t, not the Hall-Petch grain size d.[51] The tensile strength may apparently be described by a rule of mixtures between surface grains with a grain-size independent strength and interior grains that obey the Hall-Petch relation. As shown in Figure 7.55, the tensile strength of Al films decreases with decreasing part thickness in the range of 0.1–0.5 mm.[51]

From the above, size effects may strengthen or weaken the material, and interface strengthening or weakening will play a dominant role. Let us consider Si yield strength as an example. Because of the high activation energy for dislocation motion in silicon (2.2 eV), hardly any plastic flow occurs in single crystalline silicon for temperatures lower than 673°C. Grain boundaries in polysilicon further enhance this blocking of dislocation motion; hence polysilicon films can be treated as an ideal brittle material at room temperature.[52] From the above, we now understand why high yield strengths often are obtained in thin deposited films with values up to 200 times as large as those found in the corresponding bulk material.[53] In this light, the fact that fracture stress of polysilicon is between 2 and 10 times smaller than that of bulk single crystal is surprising. Greek et al.[54] explain this deviation by pointing at the high surface roughness of polysilicon films compared with single-crystal Si. They believe that a reduction in surface roughness would improve the tensile fracture strength considerably.

Adhesion

The importance of adhesion of various films to one another and to the substrate in overall IC performance and reliability cannot be stressed enough. Because

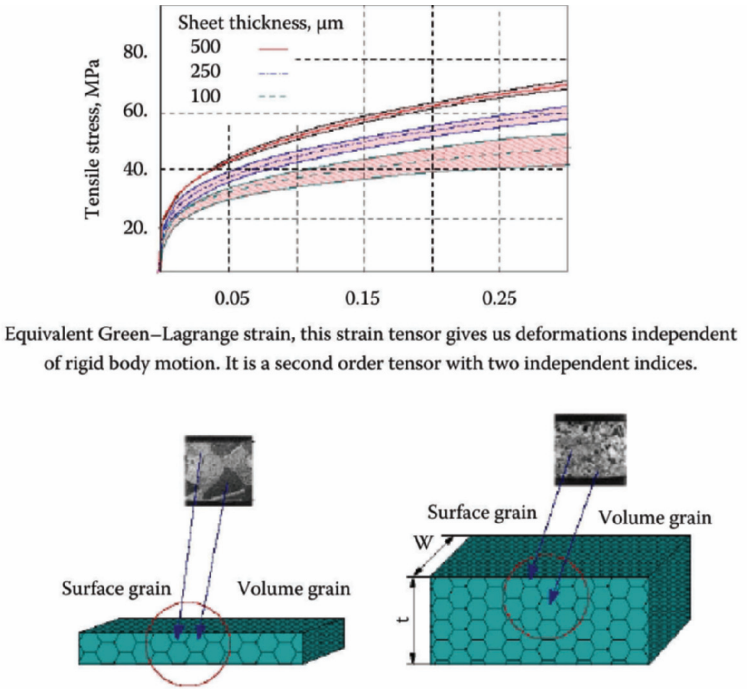


FIGURE 7.55 Thinner Al films mean weaker.[51]

mechanical pulling forces might be involved, adhesion is even more crucial in micromachining. If films lift from the substrate under a repetitive, applied mechanical force, the device will fail. Classical adhesion tests include the scotch tape test, abrasion method, scratching, deceleration (ultrasonic and ultracentrifuge techniques), bending, pulling, and so on (see also Volume III, Chapter 6 on metrology and analytical techniques).[55] Micromachined structures, because of their sensitivity to thin-film properties, enable some innovative new ways of in situ adhesion measurement. Figure 7.56 illustrates how a suspended membrane may be used for adhesion measurements. In Figure 7.56a, the membrane is suspended but still adherent to the substrate. Figure 7.56b shows the membrane after it has been peeled partially from its substrate by an applied load (gas pressure). Figure 7.56c illustrates the accompanying pressure-volume (*P-V*) cycle, in which the membrane is inflated, peeled, and then deflated. The shaded portion of Figure 7.56c illustrates the *P-V* work creating the new surface, which equals the average work of adhesion for the film-substrate interface times the area peeled during the test.[55]

Cleanliness of a substrate is a *conditio sine qua non* for good film adhesion. Roughness, providing more bonding surface area and mechanical interlocking, further improves it. Adhesion also improves with increasing adsorption energy of the deposit and/or increasing number of nucleation sites in the early growth stage of the film. Sticking energies between film and substrate range from less than 10 kcal/mol in physisorption to more than 20 kcal/mol for chemisorption. The weakest form of

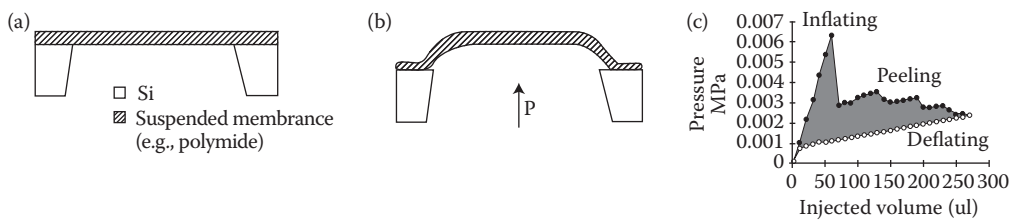


FIGURE 7.56 Micromachined structure to evaluate adhesion. (a) Suspended membrane. (b) Partially detached membrane-outward peel. (c) Pressure-volume curve during inflate-deflate cycle. (From Senturia, S. 1987. *Proceedings: IEEE micro robots and teleoperators workshop.* Hyannis, MA: IEEE. With permission.[55])

adhesion involves Van der Waals forces only (see also Table 7.17).

It is highly advantageous to include a layer of oxide-forming elements between a metal and a substrate. These adhesion layers, such as Cr, Ti, and Al, provide good anchors for subsequent metallization. Intermediate film formation allows for a continuous transition from one lattice to the other and results in the best adhesion. Adhesion also improves when formation of intermetallic metal alloys takes place.

Stress in Thin Films

Qualitative Description

Film cracking, delamination, and void formation may all be linked to film stress. Nearly all films foster a state of residual stress as a result of mismatch in the thermal expansion coefficient, nonuniform plastic deformation, lattice mismatch, substitutional or interstitial impurities, and growth processes. Figure 7.57 illustrates stress-causing factors categorized as either intrinsic or extrinsic.[55] The intrinsic stresses (also growth stresses) develop during the film nucleation. Extrinsic stresses are imposed by unintended external factors such as temperature gradients or sensor package-induced stresses. Thermal stresses, the most common type of extrinsic stresses, are well understood and often rather easy to calculate (see below on page 448). They arise either in a structure with inhomogeneous thermal expansion coefficients subjected to a uniform temperature change or in a homogeneous material exposed to a thermal gradient.[56] Intrinsic stresses in thin films often are larger than thermal stresses. They usually are a consequence of the nonequilibrium nature of the thin-film deposition process. For example, in chemical vapor deposition (CVD), depositing atoms (adatoms) may at first occupy positions other than the lowest energy configuration. With too high a deposition rate and/or too low adatom surface mobility, these first adatoms may become pinned by newly arriving adatoms, resulting in the development of intrinsic stress. Other types of intrinsic stresses illustrated in Figure 7.57 include transformation stresses occurring when part of a material undergoes a volume change during a phase transformation; misfit stresses arising in epitaxial

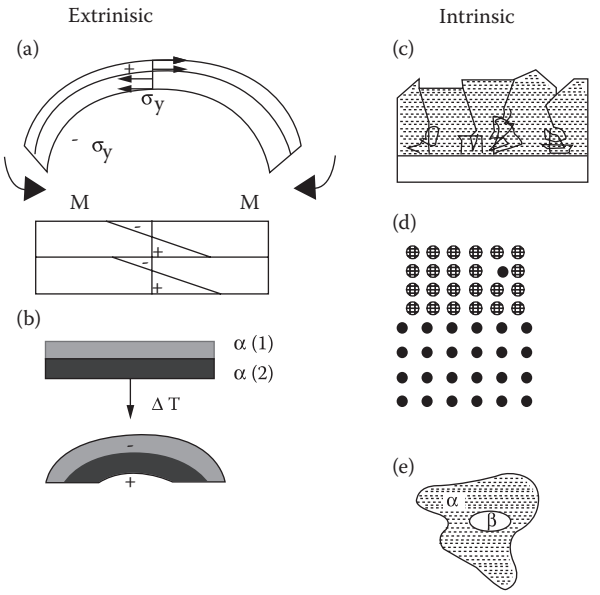


FIGURE 7.57 Examples of intrinsic and extrinsic residual stresses. (a) Nonuniform plastic deformation results in residual stresses on unloading. M = bonding moment. (b) Thermal expansion mismatch between two materials bonded together. α (1) and α (2) are thermal expansion coefficients. (c) Growth stresses evolve during film deposition. (d) Misfit stresses resulting from mismatches in lattice parameters in an epitaxial film and stresses from substitutional or interstitial impurities. (e) Volume changes accompanying phase transformations cause residual stresses. (From Krulevitch, P. A. 1994. *Micromechanical investigations of silicon and Ni–Ti–Cu thin films.* PhD thesis, University of California, Berkeley.[56])

films as a result of lattice mismatch between film and substrate; and impurities, either interstitial or substitutional, which cause intrinsic residual stresses as a result of the local expansion or contraction associated with point defects. Intrinsic stress in a thin film does usually not suffice to result in delamination unless the film is thick. For example, to overcome a low adsorption energy of 0.2 eV, a relatively high stress of about 5×10^9 dyn/cm^2 (10^7 dyn/cm^2 = 1 MPa) is required.[53] High stress can result in buckling or cracking of films.

The stress developing in a film during the initial phases of a deposition may be compressive (i.e., the film tends to expand parallel to the surface), causing buckling and blistering or delamination in extreme cases (especially with thick films). Alternatively, thin films may be in tensile stress (i.e., the film tends to contract), which may lead to cracking if forces high enough to exceed the fracture limit of the film

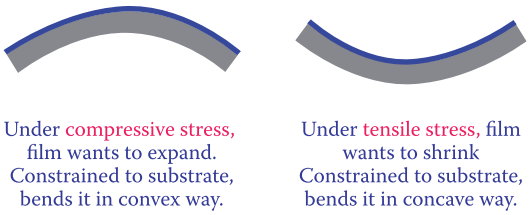


FIGURE 7.58 Tensile and compressive stress in thin films.

material are present (Figure 7.58). Subsequent rearrangement of the atoms, either during the remainder of the deposition or with additional thermal processing, can lead to further densification or expansion, decreasing remaining tensile or compressive stresses, respectively.

The mechanical response of thin-film structures is affected by the residual stress, even if the structures do not fail. For example, if the residual stress varies in the direction of film growth—in other words, if there is a stress gradient—the resulting built-in bending moment will warp released structures, such as cantilever beams (Figure 7.59). The presence of residual stress also alters the resonant frequency of thin-film, resonant microstructures (see Equation 7.29 below).[57] In addition, residual stress can lead to degradation of electrical characteristics and yield loss through defect generation.[56] For example, it was found that the resistivity of stressed metallic films is higher than that of their annealed counterparts. Residual stress has been used advantageously in a few cases, such as in self-adjusting microstructures,[58] and for altering the shape-set configuration in shape memory alloy films.[56]

In general, the stresses in thin films, by whatever means produced, are in the range of 10^8 to 5×10^{10} dyn/cm^2 and can be either tensile or compressive. For normal deposition temperatures (50 to a few hundred °C), the stress in metal films typically

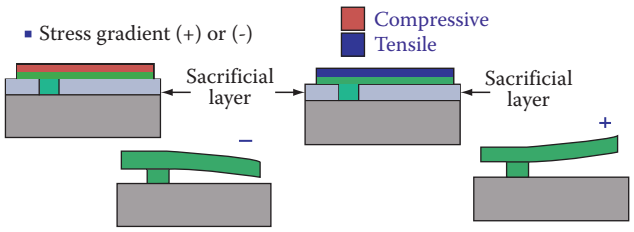


FIGURE 7.59 Stress gradients cause a bending moment in a cantilever upon release, i.e., removal of the sacrificial material.

ranges from 10^8–10^{10} dyn/cm^2 and is tensile, with the refractory metals (Mo, Ta, Nb, W, and Ta) at the upper end and the soft ones (Cu, Ag, Au, and Al) at the lower end. At low substrate temperatures, metal films tend to exhibit tensile stress. This often decreases in a linear fashion with increasing substrate temperature, finally going through zero or even becoming compressive. The changeover to compressive stress occurs at lower temperatures for lower melting point metals. The mobility of the adatoms is key to understanding the ranking for refractory and soft metals. A metal such as aluminum has a low melting temperature and a corresponding high diffusion rate even at room temperature; thus, usually it is fairly stress free. By comparison, tungsten has a relatively high melting point and a low diffusion rate, and tends to accumulate more stress when sputter deposited. With dielectric films, stresses often are compressive and have slightly lower values than commonly noted in metals.

Tensile films result, for example, when a process byproduct is present during deposition and later driven off as a gas. If the deposited atoms are not sufficiently mobile to fill in the holes left by these departing byproducts, the film will contract and go in tension. Nitrides deposited by plasma CVD are usually compressive because of the presence of hydrogen atoms in the lattice. By annealing, which drives out the hydrogen, the films can become highly tensile. Annealing also has a dramatic effect on most oxides. Oxides often are porous enough to absorb or give off a large amount of water. Full of water, they are compressive; devoid of water, they are tensile. Thermal SiO_2 is compressive even when dry. If atoms are jammed in place (such as with sputtering), the film tends to act compressively. The stress in a thin film also varies with depth.

When reviewing dry etching in Chapter 3 we discussed plasma chemistry and physics. The following listing further completes that information. The RF power of a plasma-enhanced CVD (PECVD) deposition influences stress, e.g., a thin film may start out tensile, decrease as the power increases, and finally become compressive with further RF power increase. CVD equipment manufacturers concentrate on building stress-control capabilities into new equipment by controlling plasma frequency.

Therefore, in what follows we pay specific attention to PECVD parameter settings influencing thin-film properties, important in building surface micromachined microstructures, i.e., film stress and density. Surface micromachining uses deposited thin films, such as polysilicon, silicon nitride, and silicon dioxide. One of the challenges of any surface-micromachined process is to control the intrinsic stresses in those deposited films. Plasma settings act as critical controlling parameters of intrinsic stress in CVD films. Recall that the plasmas of interest are low-pressure glow discharges consisting of ions, electrons, and neutral species. The neutral species, molecules, and radicals greatly outnumber the electrons and ions and are not influenced by external electrical fields. Just as etching occurs mainly through neutral species, deposition also almost exclusively involves neutrals. In etching, the ions impart anisotropy to the etching process, whereas in deposition they alter the properties of the deposited films. The different reactor parameter settings have the following influence on the CVD deposited films:

1. *Total reactor pressure*: Because gas density varies with pressure, the mean free path is longer at lower pressures. Consequently, ions accelerated toward the cathode at lower pressure can gain more energy before a collision takes place. Therefore, the effect of ion bombardment is more pronounced at lower pressures, and better quality CVD films ensue, characterized as films with a low wet etch rate (high film density) and low compressive stress. Experimental results show that as the reactor pressure is lowered, film stress goes from tensile to compressive and wet etch rates decrease. Often, stress dynamically changes when the film is exposed to the atmosphere and subsequent heating. In the case of oxide films, for example, tensile films take up water, swell, and become more compressive (the substrate bends down). Compressive oxide films have a greater moisture resistance. A nitride film, on the other hand, absorbs no water and shows no tendency toward compressive stress with time. In the latter case, the amount of Si-H bonds, controlled by annealing at 490°, seems to dominate the stress behavior.[59] Too high pressures promote gas-phase polymerization, increasing defect density in the deposited material. In the other extreme, too low pressures (alternatively, the reactant gas is too diluted) change the process from CVD-like to PVD-like, giving way to a columnar film morphology with more defects.
2. *Frequency of the RF excitation*: At low frequencies, ions experience the full amplitude of the RF voltage, whereas above the ion-transition frequency (>3 MHz), where ions cannot follow anymore, the ion energy is determined by the time average of the RF amplitude. Consequently, lower frequency shifts the ion energy upward. At the lower frequencies, wet etch rates decrease, and compressive film stress results. Again, higher-energy bombardment yields better films. Higher ion bombardment also improves the film quality on sidewalls. Multiple-frequency plasma equipment is emerging rapidly because it allows the user precise control over film properties (particularly stress) for a wide range of process conditions and can also improve step coverage.
3. *RF power effects*: An increase in RF power leads to more intense ion bombardment due to the increase in ion current. With a higher ion current, the film deposition rate goes up. To separate the effect of RF power on growth rate and film quality, the ratio of power density to deposition rate must be evaluated. This ratio represents a rough measure of ion bombardment per deposited molecule. In a plot of wet etch rate versus power density divided by the deposition rate, the maximum ion bombardment, corresponding to highest energy density, leads to the lowest wet etch rate.[60]
4. *Growth temperature*: The growth temperature has a strong influence on the structure of the film. At low temperatures (and high growth rates), the surface diffusion is slow relative to the arrival rate of film precursors. In this situation, the adsorbed precursor molecule is likely to interact with an impinging precursor molecule before it has a chance to diffuse away on the surface, and an amorphous film is formed. At high temperatures (and low growth rates),

the surface diffusion is fast relative to the incoming flux. The adsorbed species can diffuse to step growth sites, forming single crystalline materials.

Summarizing, low RF excitation frequency, low reactor pressures, and low deposition rate at high temperature contribute to improved film quality as evidenced by stress and (wet) etch rate measurements. More information on CVD thin films and how they are influenced by the above deposition parameters, as well as other parameters (e.g., moisture, flow rate, and gas composition) will be given in Volume III, Chapter 5 on selected MEMS and NEMS materials.

Stress in Thin Films on Thick Substrates—Quantitative Analysis

The total stress in a thin film typically is given by:

$$\sigma_{tot} = \sigma_{th} + \sigma_{int} + \sigma_{ext} \tag{7.23}$$

i.e., the sum of any intentional external applied stress (σ_{ext}), the thermal stress (σ_{th}, an unintended external stress), and different intrinsic components (σ_{int}). With constant stress through the film thickness, the stress components retain the form of:

$$\begin{aligned} \sigma_x &= \sigma_x(x,y) \\ \sigma_y &= \sigma_y(x,y) \\ \tau_{x,y} &= \tau_{x,y}(x,y) \\ \tau_{xz} &= \tau_{yz} = \sigma_z = 0 \end{aligned} \tag{7.24}$$

In other words, the three nonvanishing stress components are functions of x and y alone. No stress occurs in the direction normal to the substrate (z). With x,y as principal axes, the shear stress τ_{xy} also vanishes,[61] and Equation 7.24 reduces to the following strain-stress relationships:

$$\begin{aligned} \varepsilon_x &= \frac{\sigma_x}{E} - \frac{\nu\sigma_y}{E} \\ \varepsilon_y &= \frac{\sigma_y}{E} - \frac{\nu\sigma_x}{E} \\ \sigma_z &= 0 \end{aligned} \tag{7.25}$$

In the isotropic case, $\varepsilon = \varepsilon_x = \varepsilon_y$, so that $\sigma_x = \sigma_y = \sigma$, or:

$$\sigma = \left(\frac{E}{1-\nu}\right)\varepsilon \tag{7.26}$$

where the Young's modulus E of the film and the Poisson ratio of the film act independently of orientation. The quantity $E/1 - \nu$ often is called the *biaxial modulus*. Uniaxial testing of thin films is difficult, prompting the use of the biaxial modulus rather than Young's uniaxial modulus (see Volume III, Chapter 6 on metrology and analytical techniques). Plane stress, as described here, is only a good approximation when several thicknesses (e.g., three) away from the edge of the film.

Thermal Stress

The most commonly encountered example of thin film stress is the thermal mismatch stress due to the thermal expansion mismatch between a film and its substrate. The manufacturing process and normal operation of most thin film-based devices often involve large temperature cycles. Therefore, thermal mismatch stresses are inevitable, and they can be very high if the thermal mismatch is large and the elastic modulus of the film is large, such as is the case for Cu films used in integrated circuits (ICs). Thermal stresses develop in thin films when high-temperature deposition or annealing is involved, and they usually are unavoidable because of mismatch of thermal expansion coefficients between film and substrate. The problem of a thin film under residual thermal stress can be modeled by considering a thought experiment involving a stress-free film at high temperature on a thick substrate. Imagine detaching the film from the high-temperature substrate and cooling the system to room temperature. Usually, the substrate dimensions undergo minor shrinkage in the plane, whereas the film's dimensions may reduce significantly. To reapply the film to the substrate with complete coverage, the film needs stretching with a biaxial tensile load to a uniform radial strain, ε, followed by perfect bondage to the rigid substrate and load removal. The film stress is assumed to be the same in the stretched and freestanding film as

in the film bonded to the substrate, i.e., no relaxation occurs in the bonding process. To calculate the thermal residual stress from Equation 7.26, the elastic moduli of the film must be known, as well as the volume change associated with the residual stress, i.e., the thermal strain, ε_{th}, resulting from the difference in the coefficients of thermal expansion between the film and the substrate.

Let us now consider whether, qualitatively, the above assumptions apply to the measurement of thin films on Si wafers. Such films typically measure 1 μm thick and are deposited on 4-in. wafers, nominally 550 μm thick. In this case, the substrate measures nearly three orders of magnitude thicker than the film, and, because the bending stiffness is proportional to the thickness cubed, the substrate essentially is rigid relative to the film. The earlier assumptions clearly apply.

Figure 7.60 portrays a quantitative example where a polyimide film, strain-free at the deposition temperature (T_d) of 400°C, is cooled to room temperature T_r (25°C) on a Si substrate with a different coefficient of thermal expansion. The resulting strain is given by:

$$\varepsilon_{th} = \int [\alpha_f(T) - \alpha_s(T)]dT \qquad (7.27)$$

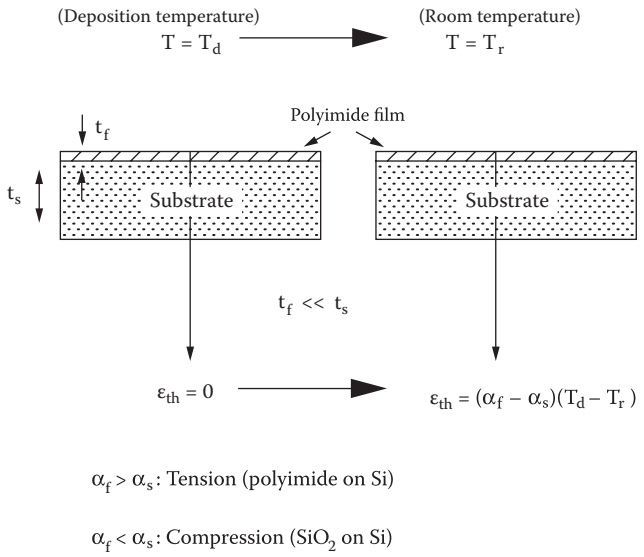


FIGURE 7.60 Thermal stress. Tension and compression are determined by the relative size of thermal expansion coefficients of film and substrate. Suppose a strain-free film at deposition temperature, T_d, is cooled to room temperature, T_r, on a substrate with a different coefficient of thermal expansion.

where α_f and α_s represent the coefficients of thermal expansion for the polyimide film and the Si substrate, respectively. The thermal strain can be of either sign; based on the relative values of α_f and α_s, positive is tensile, and negative is compressive. Polyimide features a thermal expansion ($\alpha_f = 70 \times 10^{-6}°C^{-1}$) larger than the thermal expansion coefficient of Si ($\alpha_s = 2.6 \times 10^{-6}°C^{-1}$); hence a tensile stress is expected. With SiO_2 grown or deposited on silicon at elevated temperatures, a compressive component [$\alpha_f(0.35 \times 10^{-6}°C^{-1}) < a_s(2.6 \times 10^{-6}°C^{-1})$] is expected. Assuming that the coefficients of thermal expansion are temperature independent, Equation 7.27 simplifies to:

$$\varepsilon_{th} = [\alpha_f - \alpha_s](T_d - T_r) \qquad (7.28)$$

The calculated thermal strain for polyimide on Si then measures 25×10^{-3} at room temperature. The biaxial modulus ($E/1 - \nu$), with E = 3 GPa and ν = 0.4, equals 5 GPa, and the residual stress σ, from Equation 7.26, is 125 MPa and tensile.

Intrinsic Stress

The intrinsic stress, σ_i, reflects the internal structure of a material and is less clearly understood than the thermal stress, which it often dominates.[62] Several phenomena may contribute to σ_i, making its analysis very complex. Intrinsic stress depends on thickness, deposition rate (locking in defects), deposition temperature, ambient pressure, method of film preparation, type of substrate used (lattice mismatch), incorporation of impurities during growth, and so on. Some semiquantitative descriptions of various intrinsic stress-causing factors follow:

- Doping ($\sigma_{int} > 0$ or $\sigma_{int} < 0$): When doping Si, the atomic or ionic radius of the dopant and the substitutional site determine the positive or negative intrinsic stress ($\sigma_{int} > 0$ or tensile and $\sigma_{int} <$ 0 or compressive). With boron-doped polysilicon, a small atom compared with Si, the film is expected to be tensile ($\sigma_{int} > 0$); with phosphorous doping, a large atom compared with Si, the film is expected to be compressive ($\sigma_{int} < 0$).
- Atomic peening ($\sigma_{int} < 0$): Ion bombardment by sputtered atoms and working gas densifies thin films, rendering them more compressive.

Magnetron sputtered films at low working pressure (<1 Pa) and low temperature often exhibit compressive stress. This topic was discussed in some detail higher above.

- Microvoids ($\sigma_{int} > 0$): Microvoids may arise when byproducts during deposition escape as gases and the lateral diffusion of atoms evolves too slowly to fill all the gaps, resulting in a tensile film.
- Gas entrapment ($\sigma_{int} < 0$): As an example we can cite the hydrogen trapped in Si_3N_4. Annealing removes the hydrogen, and a nitride film, compressive at first, may become tensile if the hydrogen content is sufficiently low.
- Shrinkage of polymers during cure ($\sigma_{int} > 0$): The shrinkage of polymers during curing may lead to severe tensile stress, as becomes clear in the case of polyimides. Special problems are associated with measuring the mechanical properties of polymers because they exhibit a time-dependent mechanical response (viscoelasticity), a potentially significant factor in the design of mechanical structures where polymers are subjected to sustained loads.[63]
- Grain boundaries (σ_{int} = ?): Based on intuition it is expected that the interatomic spacing in grain boundaries differs depending on the amount of strain, thus contributing to the intrinsic stress. However, the origin of, for example, the compressive stress in polysilicon and how it relates to the grain structure and interatomic spacing are not yet completely clear (see also below on coarse and fine-grated Si).

For further reading on thin-film stress, refer to Hoffman.[64,65] For a short tutorial, visit http://www.uccs.edu/~tchriste/courses/PHYS549/549lectures/mechchar.html.

Case Study: Polysilicon Deposition and Film Microstructure

Introduction Thin-film properties prove not only difficult to measure but also cumbersome to reproduce, given the many influencing parameters. Evaporation, sputtering, and molecular beam techniques can deposit dielectric and polysilicon films. In VLSI and surface micromachining, none of these techniques is as widely used as CVD. The major problems associated with the former methods are defects caused by excessive wafer handling, low throughput, poor step coverage, and nonuniform depositions. From the comparison of CVD techniques for typical IC and MEMS materials in Table 7.19, we can conclude that LPCVD, at medium temperatures, prevails above all others. VLSI devices and integrated surface micromachines require low processing temperatures to prevent movement of shallow junctions, uniform step coverage, few process-induced defects (mainly from particles generated during wafer handling and loading), and high wafer throughput to reduce cost. These requirements are best met by hot-wall, low-pressure depositions.[66] While depositing a material with LPCVD the following process parameters can be varied: deposition temperature, gas pressure, flow rate, and deposition time.

The IC industry applies LPCVD polysilicon in applications ranging from simple resistors, gates

Table 7.19 Comparison of Different Deposition Techniques

	Atmospheric CVD (APCVD)	Low T LPCVD	Medium T LPCVD	Plasma-Assisted CVD (PECVD)
Temp (°C)	300–500	300–500	500–900	100–350
Materials	SiO_2, P-glass	SiO_2, P-glass, BP-glass	Polysilicon, SiO_2, P-glass, BP-glass Si_3N_4, SiON	SiN, SiO_2, SiO_2, SiON
Uses	Passivation, insulation, spacer	Passivation, insulation, spacer	Passivation, gate metal, structural element, spacer	Passivation, insulation, structural elements
Throughput	High	High	High	Low
Step coverage	Poor	Poor	Conformal	Poor
Particles	Many	Few	Few	Many
Film properties	Good	Good	Excellent	Poor

Source: Adapted from Adams, A. C. 1988. *VLSI technology*. Ed. S. M. Sze. New York: McGraw-Hill.[30]

for MOS transistors, thin-film transistors (TFT) based on amorphous hydrogenated silicon (a-Si:H), DRAM cell plates, and trench fills, as well as in emitters in bipolar transistors and conductors for interconnects. For the last application, highly doped polysilicon is especially suited: it is easy to establish ohmic contact; it is light insensitive and corrosion resistant; and its rough surface promotes adhesion of subsequent layers. Doping elements such as arsenic, phosphorous, or boron reduce the resistivity of the polysilicon. Commercial LPCVD equipment is available to deposit polysilicon on wafers up to 8 in. in diameter. A typical batch is 100–200 wafers. The mechanical properties of polycrystalline silicon are very good:

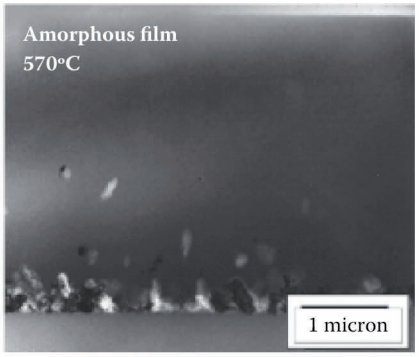
- Stronger than stainless steel: fracture strength of polysilicon ~ 2–3 GPa (steel ~ 0.2 GPa–1 GPa)
- Young's modulus E ~ 140–190 GPa
- Extremely flexible: maximum strain before fracture ~ 0.5%
- Does not fatigue readily
- Compatible with IC fabrication processes: process parameters are well known

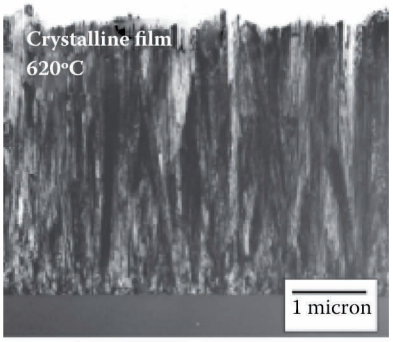
These properties explain how polysilicon also has emerged as the central structural/mechanical material in surface micromachining, and a closer look at the influence of deposition methodology on its material characteristics is warranted. In this case study we tackle polysilicon deposition by LPCVD and compare it with PECVD and sputtered polysilicon.

Undoped Polysilicon

The properties of low-pressure CVD (LPCVD) undoped polysilicon films are determined by the nucleation and growth of the silicon grains. In most cases, the Si wafers to deposit polysilicon on are placed vertically and closely spaced together in quartz boats. Stress depends on crystal structure. LPCVD Si films, grown slightly below the crystallization temperature (about 600°C for LPCVD), initially form an amorphous solid that subsequently may crystallize during the deposition process.[56,68] The CVD method results in amorphous films when the deposition temperatures are well below the melting temperature of Si (1410°C). The subsequent transition from amorphous to crystalline depends on atomic surface mobility and deposition rate. At low temperatures, surface mobility is low, and nucleation and growth are limited. Newly deposited atoms become trapped in random positions and, once buried, require a substantial amount of time to crystallize as solid-state diffusion is significantly lower than surface mobility. That is why, for low temperature deposition, amorphous layers only start to crystallize after sufficient time at temperature in the reactor. Working at temperatures between 580 and 591.5°C, Guckel et al.[68] produced mostly amorphous films. However, Krulevitch,[56] working at only slightly higher temperatures (605°C) and probably leaving the films longer in the LPCVD setup, produced crystallized films. On crossing the transition temperature between amorphous and crystalline growth (see Figures 7.61 and 7.38), crystalline growth immediately initiates at the substrate due to the increased surface mobility, which allows adatoms to find low-energy, crystalline positions from the start of the deposition process.

The deposition temperature at which the transition from amorphous to a crystalline structure occurs depends on many parameters, such as deposition rate, partial pressure of hydrogen, total pressure, presence of dopants, and presence of impurities (O, N, or C).[67] In the crystalline regime, numerous nucleation sites form, resulting in a transition zone of a multitude of small grains at the film/substrate interface to columnar crystallites on top, as shown in the schematic of a 620–650°C columnar film in Figure 7.62. In this figure, a transition zone of small, randomly oriented grains is sketched near the SiO_2 layer. The rate of crystallization is faster here than the deposition rate. Columnar grains ranging between 0.03 and 0.3 μm in diameter form on top of the small grains.[67] The columnar coarse grain structure arises from a process of growth competition among the small grains, during which those grains preferentially oriented for fast vertical growth survive at the expense of misoriented, slowly growing grains.[69,70] The lower the deposition temperature, the smaller the initial grain size will be. At 700°C, films also are columnar; however, the grains are cylindrical, extending through the thickness of the entire film, and there is no transition zone near the





Effect of temperature:

Amorphous to crystalline:	570°C
Equi-axed grains:	600°C
Columnar grains:	625°C
(110) Crystal orientation	625 – 650°C
(100) Crystal orientation	650 – 700°C

FIGURE 7.61 From amorphous to polycrystalline Si.

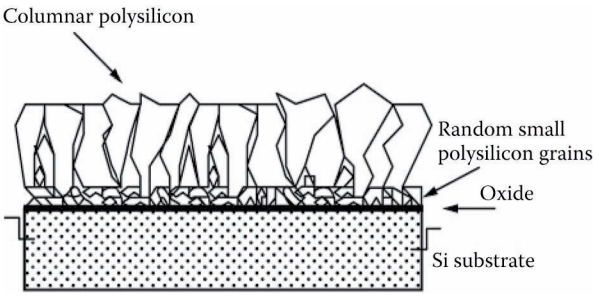


FIGURE 7.62 Schematic of compressive polysilicon formed at 620–650°C. The columnar coarse-grain structure arises from a process of grain growth competition among the small grains, during which those grains preferentially oriented for fast vertical growth survive at the expense of misoriented, slowly growing grains. (From Krulevitch, P. A. 1994. *Micromechanical investigations of silicon and Ni–Ti–Cu thin films.* PhD thesis, University of California, Berkeley.[56])

SiO_2 interface.[56] Temperature has to be very accurately controlled as grains grow with temperature, increasing surface roughness, causing loss of pattern resolution and stresses in MEMS.

Stress in polysilicon films was found to vary significantly with deposition temperature and silane pressure. Guckel et al.[68] found that their mainly amorphous films, deposited at temperatures less than 600°C, proved highly compressive with strain levels as high as –0.67%. At temperatures barely more than 600°C, Krulevitch reports tensile films, whereas for yet higher temperatures (≥620°C), the stress again turns compressive. Whereas films deposited at temperatures more than 630°C all turned out compressive, the magnitude of the compression decreased with increasing temperature. The stress gradient in the polysilicon films explains why compressive undoped and unannealed polysilicon beams tend to curl upward (positive stress gradient) when released from the substrate.[71]

Using high-resolution transmission electron microscopy, Guckel et al.[68] and Krulevitch[56] found a strong correlation between the material's microstructure and the exhibited stress. Guckel et al. found that in their mainly amorphous films, deposited at temperatures less than 600°C, a region near the substrate interface crystallized during growth with grains between 100 and 4000 Å. Krulevitch found that tensile, low-temperature films (605°C) have Si grains dispersed throughout the film thickness. These equiaxed crystals in an isotropic, higher-density film result in a tensile stress with a very small stress gradient. Krulevitch suggests that the compressive stress in the higher temperature compressive films (≥620°C) relates to the competitive growth mechanism of the columnar grains. As crystals grow vertically and in-plane they push on neighbors, resulting in compressive stress and a positive stress gradient. The same author concluded that thermal sources of stress are insignificant.

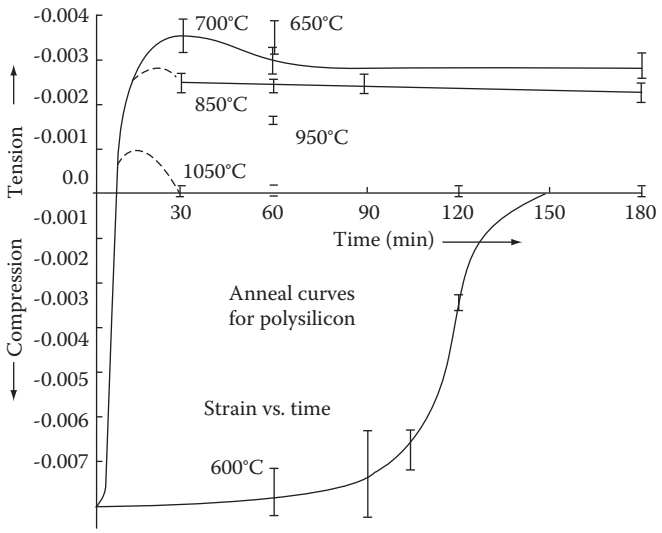


FIGURE 7.63 Anneal curves for polysilicon. Strain versus anneal time. Upper curves: low-temperature film. Lower curve: high-temperature film. (From Guckel, H., D. W. Burns, C. C. G. Visser, H. A. C. Tilmans, and D. Deroo. 1988. Fine-grained polysilicon films with built-in tensile strain. *IEEE Trans Electron Devices* 35:800–01. With permission.[74])

Importantly, Guckel et al. discovered that annealing in nitrogen or under vacuum converts the low-temperature films with compressive built-in strain (–0.007) to a tensile strain with controllable strain levels between 0 and +0.003 (see Figure 7.63). Annealing can reduce stresses by a factor of 10–100.

During anneal, no grain size increase was noticed (100–4000 Å), but a slight increase in surface roughness was measured. This type of polysilicon is referred to as *fine-grain polysilicon* (also Wisconsin poly-Si). Guckel et al. explain this strain field reversal as follows: as the amorphous region of the film crystallizes, it attempts to contract, but because of the substrate-constrained newly crystallized region, a tensile stress results. Higher temperature films, during an anneal, also become less strained, but the strain remains compressive (see lower curve in Figure 7.63). Moreover, in this case grain size increases, and the surface turns considerably rougher. The latter is called *coarse-grain polysilicon*. Fine-grain polysilicon, with its tensile strain, is preferable; however, it cannot be doped to as low a resistivity as coarse-grain polysilicon. Hence fine-grain polysilicon should be considered a structural material rather than an electronic material.

The above picture is further complicated by the controlling nature of the substrate. For example, depositing amorphous Si at even lower temperatures of 480°C from disilane (Si_2H_6) and crystallizing it by subsequent annealing at 600°C demonstrated a large dependency of crystallite size on the underlying SiO_2 surface condition. Treating the surface with HF:H_2O or NH_4OH:H_2O_2:H_2O leads to polysilicon films with a large grain size, two or three times as large as without SiO_2 treatment, believed to be the consequence of nucleation rate suppression.[72]

Abe and Reed made low-strain polysilicon thin film by DC-magnetron sputtering and postannealing. The films showed very small regional stress and very smooth texture. The deposition rate was 193 Å/min, and the substrate was neither cooled nor heated. The average roughness was found to be comparable with the surface roughness of polished, bare silicon substrates.[73]

Summarizing, stress in polysilicon depends on the material's microstructure, with tension arising from the amorphous to crystalline transformation during deposition and compression from the competitive grain growth mechanism.

Polysilicon deposited at 600–650°C has a {110}-preferred orientation. At higher temperatures, the {100} orientation dominates. Dopants, impurities, and temperature influence this preferred orientation.[67] Drosd and Washburn[75] introduced a model explaining the experimental observation that regrowth of amorphous Si is faster for {100} surfaces, followed by {110} and {111}, which are 2.3 and 20 times slower, respectively. Interestingly, the latter also pinpoints the order of fastest to slowest etching of the crystallographic planes in alkaline etchants. As discussed in Chapter 4, Elwenspoek et al.[76] used this observation of symmetry between etching and growing of Si planes as an important insight to develop a new theory explaining anisotropy in etching. In Table 7.20 we compare the discussed coarse- and fine-grain polysilicon forms.

Doped Polysilicon

To produce micromachines, doped polysilicon is used far more frequently than undoped polysilicon. Dopants decrease the resistivity to produce conductors and control stress. Polysilicon can be doped by diffusion, implantation, or the addition of dopant gases during deposition (in situ doping).

Table 7.20 Comparison of Coarse-Grain and Fine-Grain Polysilicon

	Coarse-Grain Polysilicon	Fine-Grain Polysilicon
Temperature of deposition (°C)	620–650	570–591.5
Surface roughness	Rough >50 Å	Smooth <15 Å
Grain size	Undoped: 160–320 Å as deposited In situ P-doped: 240–400 Å	Very small grains
As deposited strain	–0.002 (compressive)	–0.007 (compressive)
Effect of high temperature anneal	Grains size increases Residual strain decreases but remains compressive Reduced bending moment	Grain size increases to 100 Å;[74] others have found 700–900 Å large variation in strain: from compressive to tensile
Dry and wet etch rate	Higher for doped material, depends on dopant concentration	Higher for doped material, depends on dopant concentration
Texture	<110> as deposited <311> in situ P-doped	No texture as deposited, depends on dopant concentration, <111> after 900–1,000°C anneal[77]

Doping polysilicon films in situ reduces the number of processing steps required for producing doped microdevices by eliminating the need for a subsequent high-temperature step associated with a diffusion or ion-implantation anneal and also provides more uniform doping throughout the film thickness. In situ doping of polysilicon is accomplished by maintaining a constant PH_3 to SiH_4 gas flow ratio of about 1 vol% in a hot-wall LPCVD setup. At this ratio, the phosphorous content in the film appears above the saturation limit, and the excess dopant segregates at the grain boundaries.[67] Phosphorus diffuses significantly faster in polysilicon than in single-crystalline silicon. The diffusion takes place primarily along the grain boundaries. The diffusivity in thin films of polysilicon (i.e., small equiaxed grains) is about 1 10^{-12} cm^2/s. The dopant concentration of in situ doped films is normally very high ($\sim 10^{20}$ cm^{-3}). At more than about 1 10^{21} cm^{-3}, the film resistivity reaches a plateau of 4×10^{-4} $\Omega \cdot cm$ because of the low mobility of electrons or holes. The maximum mobility for the highest phosphorus-doped polysilicon is about 30 cm^2/Vs.[78]

In situ phosphorus-doped polysilicon undergoes the same amorphous to crystalline growth transformation observed in the undoped film, with the material's microstructure depending on deposition temperature and deposition pressure. The temperature of transformation is lower for the doped films than for the undoped polysilicon and occurs between 580 and 620°C.[79,80] Thus, phosphorus doping enhances crystallization in amorphous silicon,[81] and because of passivation of the polysilicon surface by the phosphine gas, reduces the polysilicon deposition rate.[79] Decreases in deposition rate by as much as a factor of 25 have been reported.[82] Slower deposition rates allow more time for adatoms to find crystalline sites, resulting in crystalline growth at lower temperatures. From Table 7.20, we read that the grain size of phosphorus-doped polysilicon tends to be larger (240–400 Å) than for the undoped material and that {311} planes show up as a texture facet in the doped material. In contrast to in situ phosphine and arsine doping, which both decrease the deposition rate, diborane doping of polysilicon to make it p^+ accelerates the deposition rate.[67]

At lower deposition temperatures and higher pressures, the microstructure again consists of amorphous and crystalline regions, whereas at higher temperatures and lower pressures columnar films result and as deposited films exhibit compressive residual stress. The columnar films have a stress gradient that increases toward the film surface, as opposed to the gradient found in undoped columnar polysilicon. This gradient in stress most likely is to the result of nonuniform distribution of phosphorus throughout the film. Annealing at 950°C for 1 h results in the same stress and stress gradient for initially columnar and initially amorphous/crystalline films [i.e., $\sigma_f = -45$ MPa and $\Gamma = +0.2$ mm^{-1} (a linear stress gradient with physical dimensions 1/length), respectively].[56]

As in the case of undoped polysilicon, phosphorus-doped polysilicon films with smooth surfaces (fine grain) are obtained by depositing in situ doped films in the amorphous state and then

annealing.[77,83] Phosphorus-doped polysilicon oxidizes faster than undoped polysilicon. The rate of oxidation is determined by the dopant concentration at the polysilicon surface.[67] The addition of oxygen to polysilicon increases the film's resistivity, and the resulting coating, semi-insulating polysilicon (SIPOS) acts as a passivating coating for high-voltage devices in the IC industry. SISPOS has not yet emerged in surface micromachining.

Four drawbacks of in situ phosphorus doping are the complexity of the deposition process, slower deposition rates,[84] reduced film thickness uniformity,[82] and the cleaning of the reactor, which is more demanding for the doped process.

Diffusion is often a more effective method for doping of polysilicon than in situ doping, especially for very heavy dopings (e.g., polysilicon resistivities down to the 10^{-4} $\Omega \cdot$cm) of thick (e.g., 2 μm) films. However, diffusion is a high-temperature (e.g., 900–1,000°C) process. If the diffusion step is performed for long durations (a few hours) to achieve uniform doping throughout film thickness a few micrometers thick, it will destroy electronics that are fabricated on the wafer before polysilicon surface micromachining. If performed for too short a time, dopant distribution through the film thickness will not be sufficiently uniform, resulting in difficulties with mechanical property variations through the film thickness.

Ion implantation can also be used for doping polysilicon. The implantation energy is typically adjusted so that the peak of the concentration profile is at the center of the film thickness. However, the resistivity of implanted polysilicon films is not as low as that possible by diffusion.

As with undoped polysilicon, the intrinsic stresses in as-deposited doped polysilicon films are large (>500 MPa) and can result in warping or curling of released micromechanical structures. The values for the fracture stress of boron-, arsenic-, and phosphorus-doped polysilicon are 2.77 ± 0.08 GPa, 2.70 ± 0.09 GPa, and 2.11 ± 0.1 GPa, respectively, compared with 2.84 ± 0.09 GPa for undoped polysilicon. The lower value for phosphorus-doped material has been attributed to high surface roughness and with the large number of defects associated with extensive grain growth in highly phosphorus-doped films.[85] Maluf[86] reports that the operation of polysilicon at elevated temperatures is subject to long-term instabilities, drift, and hysteresis as a result of slow stress-annealing effects.

Several sensors using polysilicon piezoresistive sense elements have been demonstrated. The piezoresistive coefficient is an average over all the orientations in polycrystalline silicon. The gauge factor ranges between 20 and 40, which is about a factor of five smaller than in single-crystal Si, and quickly decreases as the doping exceeds 10^{19} cm^{-3}. At doping levels of 10^{20} cm^{-3}, the temperature coefficient of resistance of polysilicon is approximately 0.04% per °C compared with 0.14% per °C for crystalline Si. This is an advantage for the use of polysilicon, even though the gauge factor is low at these high doping levels.

PECVD and Sputtered Polysilicon

The quest for low-temperature polysilicon deposition processes is driven by the need for compatibility with prefabricated aluminum-metallized CMOS circuitry. This makes the 320°C PECVD and ~350°C sputter deposition methods of polysilicon especially interesting, although both methods result in amorphous films.

PECVD films, deposited in a 50-kHz parallel-plate diode reactor, can be doped in situ and crystallized by rapid thermal annealing (RTA; 1100°C, 100 s). It was shown that small-grained PECVD films annealed by RTA have good electrical properties and gauge factors between 20 and 30, similar to those reported for LPCVD polycrystalline silicon (see above).[87]

Honer et al.[88] introduced sputtered polysilicon to the MEMS community. Sputtered silicon can be used to fabricate polysilicon microstructures atop standard, aluminum metallized CMOS at temperatures less than 350°C. Films with stress values less than 100 MPa are routinely obtained.

An important benefit of both these low-temperature processes is that they are not only compatible with conventional oxide sacrificial layers but also with certain organic sacrificial layers (e.g., polyimide). Organic layers can be removed in a dry oxygen plasma etch, thus avoiding the stiction and selectivity problems associated with wet etch releases and at

the same time alleviating the concerns over chemical attack on structural elements by HF.

A drawback for the sputtered boron-doped Si films is that to increase their conductivity (to 25 mho/□), they need to be clad in symmetric, 50-nm-thick layers of titanium-tungsten.

Kamins[78] provides a detailed study of polysilicon physical properties. Another good resource on polysilicon morphology and doping can be found at http://www.iue.tuwien.ac.at/22.0.html. Sharpe and Edwards[89] at Johns Hopkins have perhaps provided the most detailed polysilicon mechanical property studies.

Thin Film Characterization

In Volume III, Chapter 6 on metrology, we detail thin film characterization. For material structure we rely on scanning electron microscopy (SEM), transmission electron microscopy (TEM), scanning probe microscopies (STM, AFM), and x-ray diffraction (XRD). For chemical composition we use auger electron spectroscopy (AES), energy dispersive analysis of x-rays (EDAX or EDS), x-ray photoelectron spectroscopy (XPS), secondary ion mass spectrometry (SIMS), and Rutherford backscattering (RBS). Mechanical properties of thin films are studied using residual stress measurement and micro-/nanoindentation. Electrical properties are derived from resistance (four-point probe) and capacitance measurements. For optical properties we rely on ellipsometry, and for film thickness we rely on stylus profilometer and ellipsometry.

We also introduce a wide variety of surface micromachined devices that are used in characterization of thin films.

Surface Micromachining

Introduction

Bulk micromachining means that 3D features are etched into the bulk of crystalline and noncrystalline materials (see Chapter 4). In contrast, surface-micromachined features are built up, layer by layer, on the surface of a substrate (e.g., a single-crystal silicon wafer). Dry etching defines the surface features in the x, y plane, and wet etching releases them from the plane by undercutting. In surface micromachining,

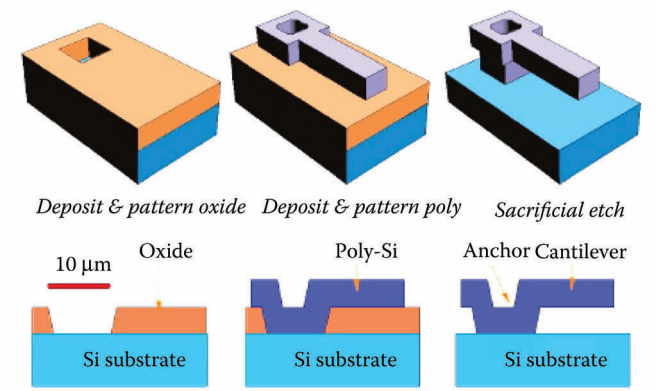


FIGURE 7.64 Surface micromachining of a polysilicon anchor.

shapes in the x, y plane are unrestricted by the crystallography of the substrate (see Figure 7.1). In Figure 7.64 a polysilicon anchor is made using typical surface micromachining techniques.

The nature of the deposition processes involved determines the limited height of surface-micromachined features (Hal Jerman, from EG&G's IC Sensors, called them *2.5D features*[90]). Specifically, low-pressure CVD (LPCVD) polycrystalline silicon (polysilicon) films generally are only a few micrometers high (low z), in contrast with wet bulk micromachining where only the wafer thickness limits the feature height. A low z may be a drawback for some mechanical sensors. For example, it would be difficult to fashion a large inertial mass for an accelerometer from thin polysilicon plates (a commercial surface-micromachined accelerometer, the ADLX05, has an inertial mass of only 0.3 μg). Not only do many parameters in the LPCVD polysilicon process need to be controlled very precisely, but subsequent high-temperature annealing (e.g., at temperatures of about 580°C) also is needed to transform the deposited amorphous silicon into polysilicon—the main structural material in surface micromachining. Even with the best possible process control, polysilicon has some material disadvantages over single-crystal Si. For example, it generates a somewhat smaller yield strength (values between 2 and 10 times smaller have been reported)[91,92] and has a lower piezoresistivity.[93] Moreover, because a single polysilicon grain diameter may constitute a significant fraction of the thickness of a mechanical member, the effective Young's modulus may exhibit variability from sensor to sensor.[94] An important

positive attribute of polysilicon is that its material properties, although somewhat inferior to single crystal, are far superior compared with those of metal films, and, most of all, because they are isotropic, design is rendered dramatically simpler than with single-crystal material. Dimensional uncertainties may be of greater concern than material issues. Although absolute dimensional tolerances obtained with lithography techniques can be submicrometer, relative tolerances are poor, perhaps 1% on the length of a 100-μm-long feature. The situation becomes critical with yet smaller feature sizes (see Figures 6.3 and 6.4). Although the relatively coarse dimensional control in the microdomain is not specific to surface micromachining, there is no crystallography to rely on for improved dimensional control, as in the case of wet bulk micromachining. Moreover, because the mechanical members in surface micromachining tend to be smaller, more postfabrication adjustment of the features is required to achieve reproducible characteristics. Finally, the wet process for releasing structural elements from a substrate tends to cause sticking of suspended structures to the substrate, or *stiction*, introducing another disadvantage associated with surface micromachining.

Some of the problems associated with surface micromachining mentioned above have been resolved by process modifications and/or alternative designs, and the technique has rapidly gained commercial interest, mainly because it is the most IC-compatible micromachining process developed to date. Moreover, in the past 10–15 years, processes such as silicon on insulator (SOI) (1),[95] hinged polysilicon (2),[96] Keller's molded milliscale polysilicon (3),[97] thick (10 μm and beyond) polysilicon (4),[98] LIGA and LIGA-like processes (5), and porous polysilicon (6) have further enriched the surface micromachining arsenal. Some preliminary remarks on each of these surface micromachining extensions follow.

1. Silicon crystalline features, anywhere between fractions of a micrometer to 100 μm high, can be readily obtained by surface micromachining of the epi silicon or fusion-bonded silicon layer of SOI wafers.[99] Structural elements made from these single-crystalline Si layers result in more reproducible and reliable sensors. SOI or epi-micromachining combines the best features of surface micromachining (i.e., IC compatibility and freedom in *x,y* shapes) with the best features of bulk micromachining (superior single-crystal Si properties). Moreover, SOI surface micromachining frequently involves fewer process steps and offers better control over the thickness of crucial building blocks. Given the poor reproducibility of mechanical properties and generally poor electronic characteristics of polysilicon films, SOI machining may surpass the polysilicon technology for fabricating high performance devices.
2. The fabrication of polysilicon planar structures for subsequent vertical assembly by mechanical rotation around micromachined hinges, introduced by Pister et al.[96] in 1991, dramatically increases the plethora of designs feasible with polysilicon. Today, erecting these polysilicon structures with the probes of an electrical probe station, or occasionally assembly by chance in an HF etch or DI water rinse,[100] represents a too-complicated or unreliable postrelease assembly method for commercial acceptance, and alternative self-assembly means are an urgent need.
3. While at the University of California, Berkeley, Keller, [97] now CEO of MEMS Precision Instruments (http://www.memspi.com), introduced a combination of surface micromachining and LIGA-like molding processes in the HEXSIL (HEXagonal honeycomb polySILicon) process, a technology enabling the fabrication of tall 3D microstructures without postrelease assembly. Using CVD processes, generally only thin films (≈2–5 μm) can be deposited on flat surfaces. If, however, these surfaces are the opposing faces of deep narrow trenches, the growing films will merge to form solid beams. Releasing of such polysilicon structures and the incorporation of electroplating steps expand the surface micromachining bandwidth in terms of choice of materials and accessible feature heights. In this fashion, high-aspect-ratio structures normally associated with LIGA can now also be made of CVD polysilicon.

4. Applying classical LPCVD to obtain polysilicon deposition is a slow process. For example, a layer of 10 µm typically requires a deposition time of 10 h. Consequently, most micromachined structures are based on layer thicknesses in the 2–5-µm range. Based on dichlorosilane (SiH_2Cl_2) chemistry, Lange et al.[98] developed a CVD process in a vertical epitaxy batch reactor with deposition rates as high as 0.55 µm/min at 1000°C. The process yields acceptable deposition times for thicknesses in the 10-µm range. The highly columnar polysilicon films are deposited on sacrificial SiO_2 layers and exhibit low internal tensile stress, making them suitable for surface micromachining.
5. Thick layers of polyimide and other new UV resists have also received a lot of attention as important new extensions of surface micromachining. Because of their transparency to exposing UV light, they can be transformed into tall surface structures with LIGA-like high-aspect ratios. They may also be electroplated and micromolded in any plastic of choice.
6. In Chapter 4 we discussed the transformation of single-crystal Si into a porous material with porosity and pore sizes determined by the current density, type, and concentration of the dopant, as well as the hydrofluoric acid concentration. A transition from pore formation to electropolishing is reached by increasing the current density and/or by decreasing the hydrofluoric acid concentration.[101,102] Porous silicon can also be formed under similar conditions from LPCVD polysilicon.[103]

In this section, we first review the main surface micromachining process steps and clarify the extensions of the surface micromachining technique listed above. Next we compare surface micromachining with bulk micromachining. In the surface micromachining examples at the end of this section, we first look at a lateral resonator. Resonators have found an important industrial application in accelerometers and gyros introduced by Analog Devices. The second example involves Texas Instruments' (TI's) Digital Micromirror Device™ (DMD), a chip now found in many projectors.

Historical Note

The first example of a surface micromachine for an electromechanical application consisted of an underetched metal cantilever beam for a resonant gate transistor made by Nathanson in 1967.[104] By 1970, a first suggestion for a magnetically actuated metallic micromotor emerged.[105] Because of fatigue problems, metals are not typically used as mechanical components. The surface micromachining method as we know it today was first demonstrated by Howe and Muller in the early 1980s and relied on polysilicon as the structural material.[106] These pioneers, and the late Guckel,[107] an early contributor to the field, produced free-standing LPCVD polysilicon structures by removing the oxide layers on which the polysilicon features were formed. Howe's first device consisted of a resonator designed to measure the change of mass on adsorption of chemicals from the surrounding air. However, this gas sensor does not necessarily represent a good application of a surface-micromachined electrostatic structure because humidity and dust foul the thin air gap of such a microstructure in a minimal amount of time. Later, mechanical structures, especially hermetically sealed mechanical devices, provided proof that the IC revolution could be extended to electromechanical systems.[104] In these structures, the height (z-direction) typically is limited to less than 10 µm, ergo the name *surface micromachining*.

Gabriel et al.[108] in 1989 made the first survey of possible applications of polysilicon surface micromachining. Microscale movable mechanical pin joints, springs, gears, cranks, sliders, sealed cavities, and many other mechanical and optical components have been demonstrated in the laboratory.[109,110] For awhile, in the early 1990s, it seemed that every MEMS group in the United States was trying to make surface-micromachined micromotors of the type shown in Figure 7.65.

Micromotors may lack practical use as of yet, but, just as the ion-sensitive field effect transistor (ISFET) galvanized the chemical sensor community in trying out new chemical sensing approaches, micromotors energized the micromachining research community to fervently explore miniaturization of a wide variety of mechanical sensors and actuators. Micromotors

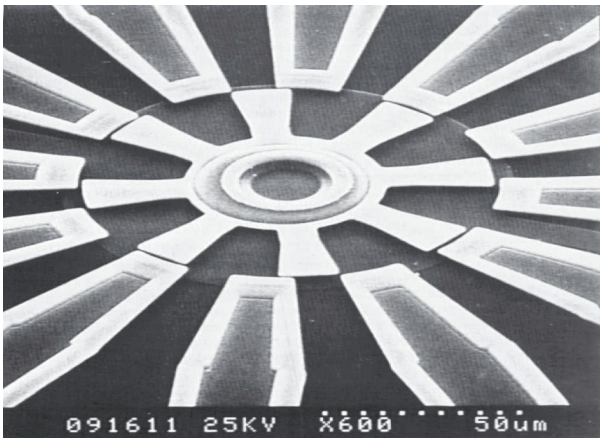


FIGURE 7.65 Surface micromachined motor. (M. Mehregany, 1990. PhD dissertation, MIT.)

also brought about the christening of the micromachining field into MEMS. In 1991, Analog Devices (Norwood, MA) announced the first commercial product based on surface micromachining, namely, the ADXL-50, a 50-g accelerometer for activating air-bag deployment.[111] The ADXL-50 chip is shown in Figure 7.66a. The SEM inset in this figure depicts the polysilicon micromachined mechanical structure. By 2001, Analog Devices was making 2 million surface-micromachined accelerometers a month (at $4 per device in volume). A second commercial success for surface micromachining was based on Texas Instruments' Digital Mirror Device™ (DMD). This surface-micromachined movable aluminum mirror is a digital light switch that precisely controls a light source for projection displays and hard copy applications.[112] A schematic of the tilting mirrors in a DMD chip is shown in Figure 7.66b. The commercial acceptance of this second application confirmed the staying power of surface micromachining.

Surface micromachining is also an established manufacturing process at Cronos (Research Triangle Park, NC; now a JDS Uniphase Company) and Robert Bosch (Stuttgart, Germany).

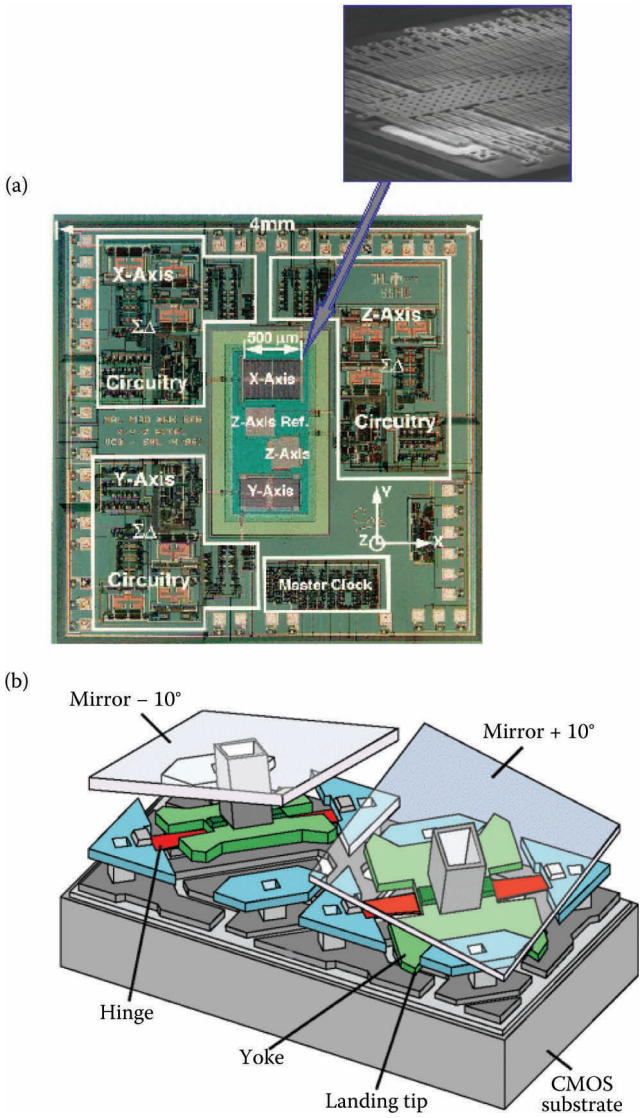


FIGURE 7.66 (a) The ADXL-50 chip from Analog Devices. The H-shaped structure in this photo is the polysilicon micromachined mechanical structure (http://www.analog.com/library/techArticles/mems/xlbckgdr4.html). (b) Schematic of the tilting mirrors in a DMD chip from TI.

Surface Micromachining Processes

Basic Process Sequence—Overview

A first simple surface micromachining process sequence for the creation of a freestanding polysilicon anchor was illustrated in Figure 7.1. We will now detail typical surface micromachining fabrication steps using the DARPA-sponsored multiuser MEMS process (MUMPs®). The baseline process of the MUMPs program is a three-layer polysilicon (Poly 0, Poly 1, and Poly 2) surface micromachining process ("polysurf") known as *PolyMUMPs*. The process includes eight lithography levels and seven physical layers. A physical layer represents a layer of material deposited during the fabrication process and is usually represented in mixed-case letters. A lithographic level is used to pattern a physical layer and is always represented in CAPITAL letters. The physical layers illustrated in Figure 7.67 are two mechanical and one electrical layer of polysilicon,

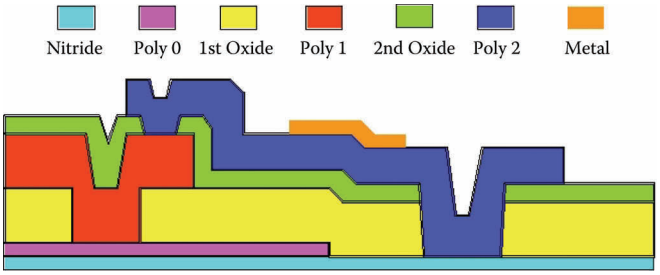


FIGURE 7.67 The seven physical: two mechanical (Poly 1 and Poly 2) and one electrical layer (Poly 0) of polysilicon, two sacrificial oxide layers, one metal conduction and one electrical isolation nitride layer.

two sacrificial oxide layers, one metal conduction, and one electrical isolation nitride layer (also buffer or isolation layer). A sacrificial oxide layer of fast etching silicon dioxide is used to define the separation between mechanical layers and between mechanical layer and the substrate. The nitride layer is used for electrical isolation between substrate and electrical surface layers. The Poly 0 layer is the electrical polylayer used as a ground plane or for electrode formation below the first mechanical layer (Poly 1). The first sacrificial oxide layer will establish a gap between Poly 1 and substrate/nitride after release etching. The second sacrificial oxide layer will make for the gap between a second mechanical layer (Poly 2) and the first mechanical polysilicon (Poly 1) after release. For sacrificial oxide one often uses phosphosilicate glass (PSG), a phosphorous-containing silicon dioxide layer generated by CVD and used for its fast etching properties. A metal layer provides electrical connection to the package.

The eight lithography levels are:

1. POLY ZERO: Defines the polysilicon zero features (Poly 0)
2. ANCHOR 1: Opens points of contact between first polysilicon and substrate (nitride or Poly 0)
3. DIMPLE: Generates "bumps" in undersurface of Poly 1 to minimize stiction
4. POLY 1: Defines first polysilicon features
5. POLY 1-POLY 2-VIA: Opens points of contact between first and second polysilicon
6. ANCHOR 2: Opens points of contact between second polysilicon and substrate/nitride
7. POLY 2: Defines second polysilicon features
8. METAL: Defines location of metal features

Remarks:

- Dimples are small, shallow features on the underside of the lower polysilicon layer to minimize the area of contact between the polysilicon and the substrate.
- A lithography level may or may not correspond with a physical layer: e.g., Poly 1 = POLY 1, but second oxide is patterned by both ANCHOR 2 and POLY 1-POLY 2-VIA.
- Release is the last step of the process where the sacrificial layers are removed by submersion into HF.
- Stiction is the sticking effect between polysilicon and the substrate that occurs during the removal of the sacrificial oxide. Many attempts to limit stiction, including dimples, special release chemicals, and processes, are tried, some successfully (see below).

Since its inception in December 1992, the multiuser MEMS process (MUMPs) quickly became a well-established, commercial program that provides customers with cost-effective access to surface micromachining for prototyping activities. It caused many researchers to focus on design and testing of new MEMS concepts rather than building yet another clean room. MUMPs was approaching its 40th run in December 2000. Another source for surface micromachining prototyping is at Sandia, with the ultraplanar multilevel MEMS technology (SUMMiT), e.g., SUMMiT-IV and SUMMiT-V (http://www.mems.sandia.gov/tech-info/summit-v.html) introduced in 1998 – (four and five polysilicon level processes). SUMMiT is reviewed further below.

Process Steps for Making a Polysilicon Micromotor

The following is a description of the PolyMUMPs process applied to the fabrication of a surface-micromachined electrostatic micromotor. A blanket n^+ diffusion of the silicon substrate (e.g., with $POCl_3$), defining a ground plane, often outlines the very first step in surface micromachining (Step I in Figure 7.68), followed by a passivation step of the substrate, for example, with 0.15-μm-thick LPCVD low stress nitride (sometimes on top of a 0.5-μm

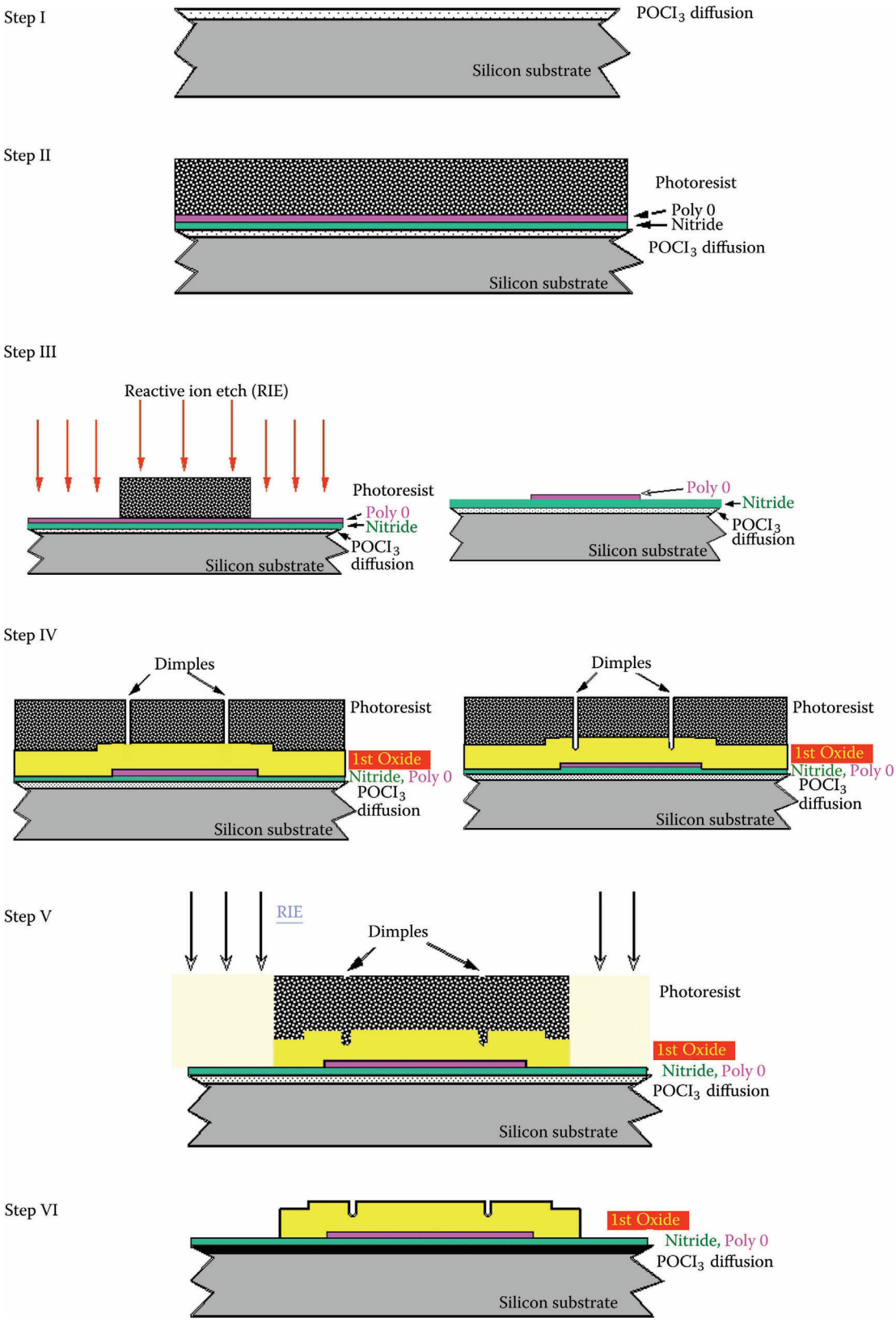


FIGURE 7.68 The PolyMUMPs process applied to the fabrication of a surface-micromachined micromotor.

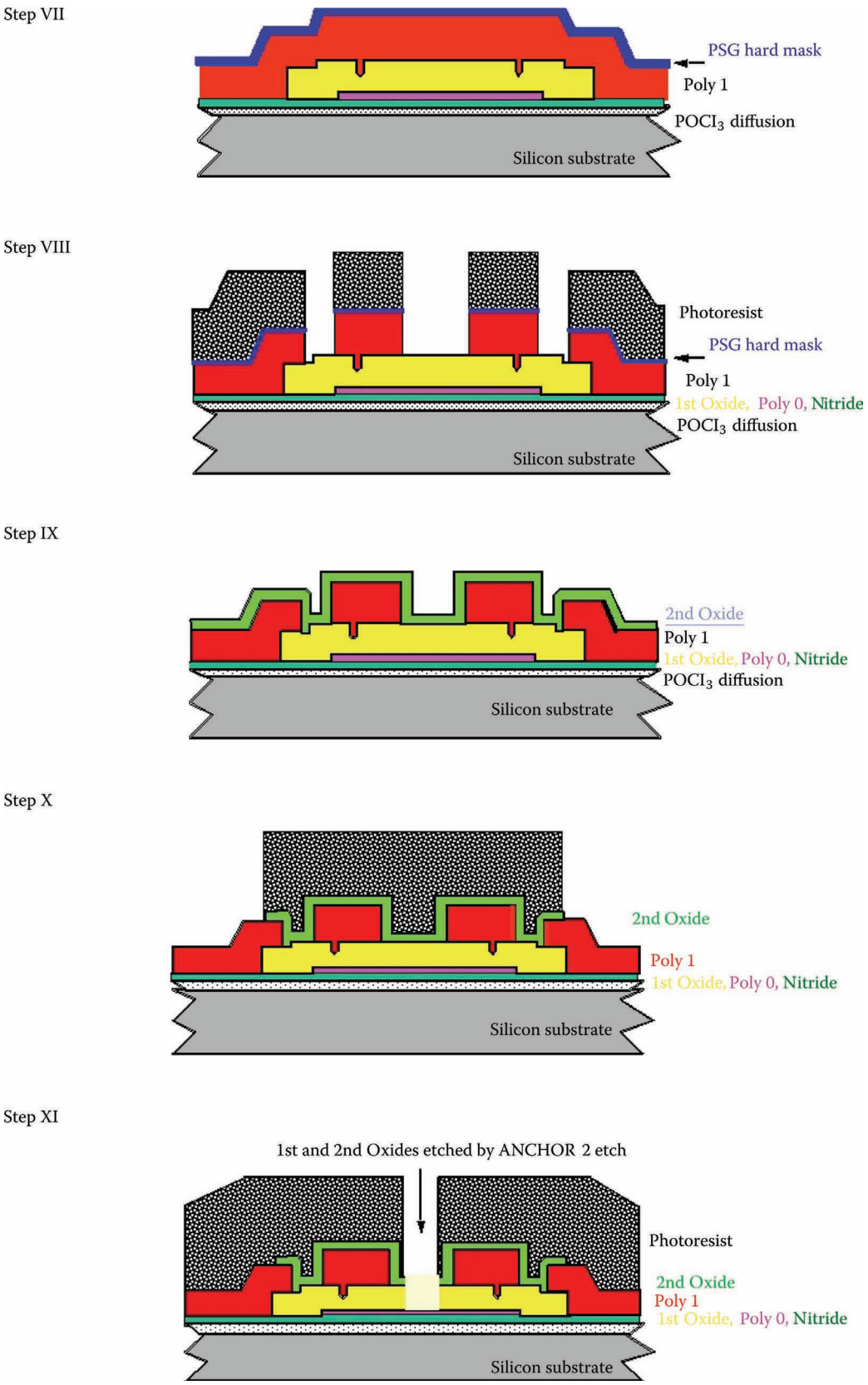


FIGURE 7.68 (*Continued*)

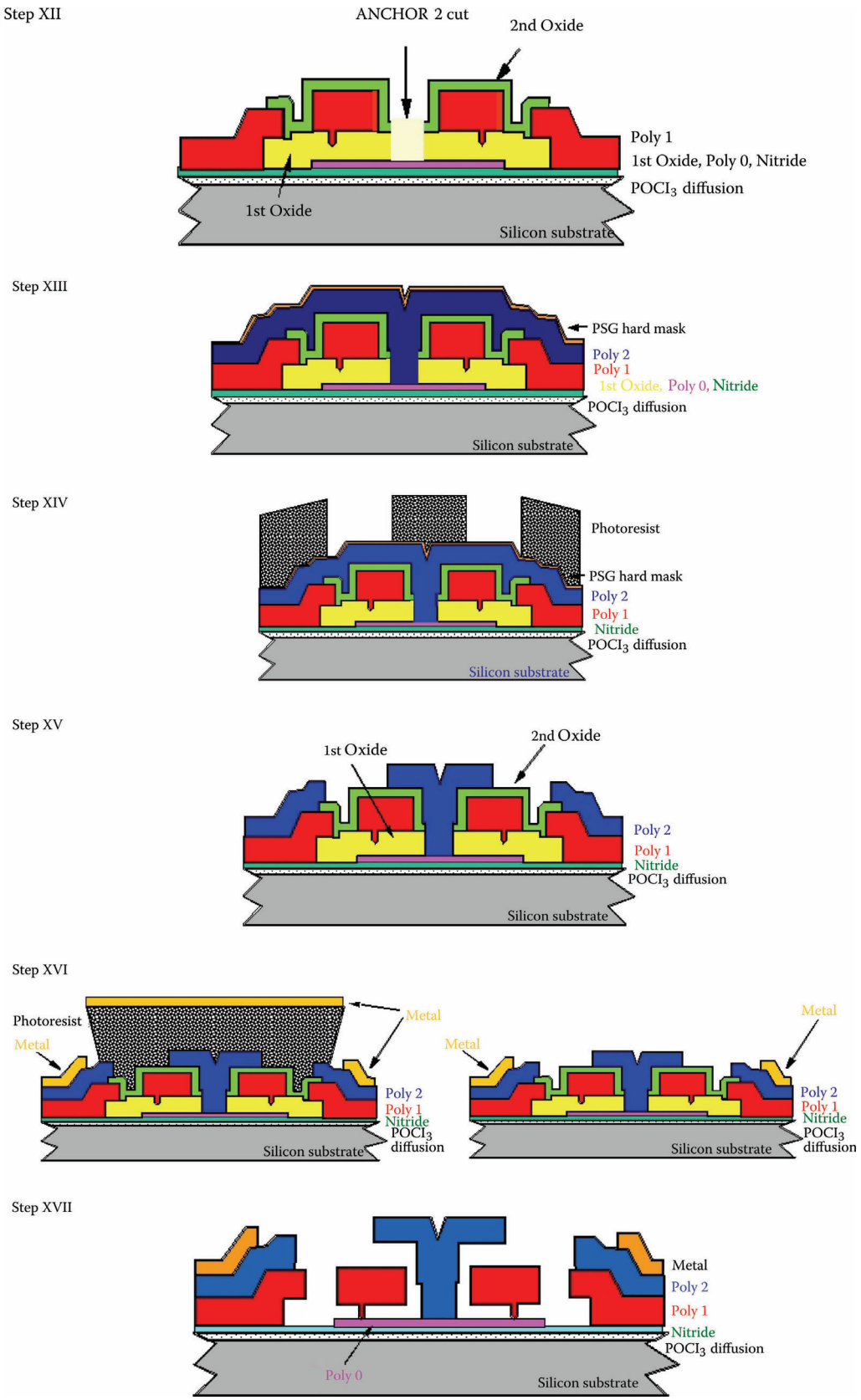


FIGURE 7.68 (*Continued*)

thermal oxide). This is followed by a blanket deposition of polysilicon (Poly 0), and the wafer is spin-coated with a photoresist for patterning (Step II). The photoresist is exposed using the first mask level (POLY 0), and the image is developed. The unwanted polysilicon is removed by RIE etching, transferring the POLY 0 pattern onto the wafer, and the photoresist is stripped in a solvent (Step III).

Step IV involves the deposition of the first sacrificial oxide (PSG) layer (2.0 μm thick) on the wafer by low temperature CVD. The wafer is coated with photoresist, and a second mask level (DIMPLE) is exposed and developed. The dimples are etched into the first oxide by a combination of RIE and buffered oxide etch (BOE) etching. In Step V, the dimple photoresist is stripped, and a new layer of photoresist is applied for the third mask level (ANCHOR 1). The first oxide is patterned and then processed with RIE to remove the oxide from the anchor area. The photoresist is stripped in a solvent bath (Step VI). The wafer is now ready for Poly 1 processing. The ANCHOR 1 level defines where Poly 1 will be attached to the substrate, and the thickness of the first oxide defines how far above the substrate (either nitride or Poly 0) Poly 1 will sit after release. After the wafer is cleaned, the first polysilicon layer (Poly 1) is deposited using LPCVD. An additional thin PSG layer is deposited on top of the Poly 1, and the wafer is annealed at high temperature to reduce the residual stress and dope the Poly 1 (Step VII). We will come back to this stress reduction matter and doping of a polysilicon layer sandwiched between two PSG further below. The wafer is now again coated with photoresist, and the fourth mask (POLY 1) is patterned. The wafer is RIE etched, stopping on the first oxide, using both the photoresist and thin PSG (top) layer as masks for the Poly 1 layer (Step VII). In Step IX, the second sacrificial oxide layer (0.75 μm) is deposited by low temperature CVD, conformally coating the topography on the wafer and defining the separation of the first polylayer from the second polysilicon. The wafer is coated with photoresist, and the fifth mask (POLY 1-POLY 2-VIA) is patterned and RIE etched. This defines the contact regions between Poly 1 and Poly 2 (Step X). The photoresist is stripped, and the wafer is recoated. The sixth mask (ANCHOR 2) is patterned, and both first oxide and second oxide are etched in one step (Step XI). This defines the region where Poly 2 will contact the substrate (either nitride or Poly 0). The photoresist is stripped, and the wafer is ready for the Poly 2 deposition (Step XII). The second polysilicon layer is deposited followed by a thin PSG layer. The structure is annealed to reduce the Poly 2 stress and to dope the Poly 2. Because the CVD film is conformal, all the holes will be filled, to varying degrees, with Poly 2 (Step XIII). The wafer is coated with photoresist, and the seventh mask (POLY 2) is patterned. Both the photoresist and thin PSG (hard mask) will mask the RIE etch (Step XIV). After the Poly 2 layer is etched and the wafer is stripped, the basic mechanical structure is complete (Step XV). A lift-off template is used to deposit the metal layer on the wafer. Photoresist is patterned using the eighth (and ninth) masks (METAL), and the metal is deposited, adhering to the Poly 2, where exposed, and breaking continuity as it goes over the photoresist step (lift-off profile), and the photoresist and remnant metal are removed by rinsing in solvent. The structure is now completed and ready for releasing (Step XVI). A 2.5-min soak in straight 49% HF removes all the sacrificial oxide layers and releases the moveable mechanical parts. The rotor is now free to spin about the center pin, and the stator poles are fixed and electrically active (Step XVII). The chips are rinsed in DI water, followed by soaking in isopropyl alcohol and baking in a convection oven.

Process Steps Details

Pattern Transfer to Buffer/Isolation Layer Suppose the buffer/isolation passivation layer itself needs to be patterned, perhaps to make a metal contact pad to the underlying Si substrate. Then, the appropriate fabrication step is a pattern transfer to the thin isolation film as shown in Figure 7.69. In this case, the isolation layer is a 1-μm-thick thermal SiO_2 film, and we illustrate a wet pattern transfer with a 1-μm resist layer. Typically, an isotropic etch such as buffered HF, also called a *buffered oxide etch* [e.g., BHF or BOE (5:1), which is five parts NH_4F and one part concentrated HF], is used (unbuffered HF attacks the photoresist). This solution etches SiO_2 at a rate of 100 nm/min, and the creation of the opening to the underlying substrate takes about 10 min. The etch progress may be monitored optically (color

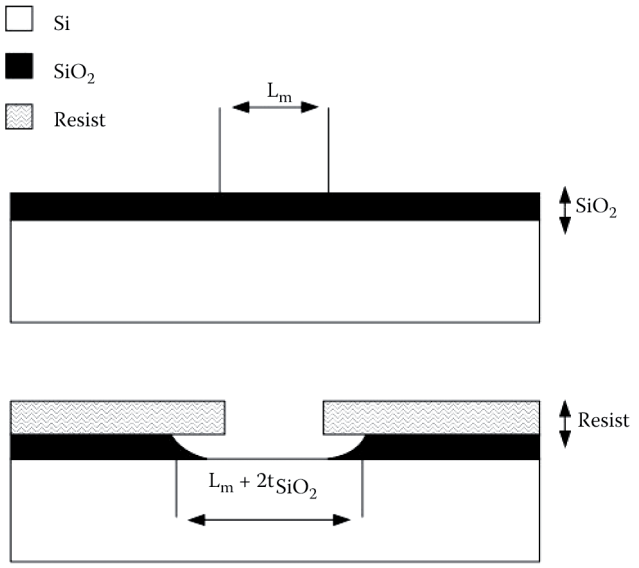


FIGURE 7.69 Wet etch pattern transfer to a thin thermal SiO_2 film for the fabrication of a contact pad to the Si substrate.

change) or by observing the hydrophobic/hydrophilic behavior[113] of the etched layer. With a resist opening, L_m, the undercut typically measures the same thickness as the oxide thickness, t_{SiO2}. In other words, the contact pad will have a size of $L_m + 2t_{SiO2}$. The undercut worsens with loss of photoresist adhesion during etching. Using an adhesion promoter such as HMDS (hexamethyldisilazane) proves useful in such cases. A new bake after 5 min of etching is a good procedure to maintain the resist integrity. After the isotropic etch, the resist is stripped in a piranha etch bath. This strong oxidizer grows about 3 nm of oxide back in the cleared window. To remove the oxide resulting from the piranha, a dip in diluted BHF suffices. After cleaning and drying, the substrate is ready for contact metal and base material deposition. Applying a dry etch (e.g., CF_4-H_2 plasma) to open up the window in the oxide would eliminate undercutting of the resist but requires a longer setup time. With an LPCVD low-stress nitride or a nitride on top of a thermal SiO_2, an often-used combination for etching the buffer/isolation layer is a dry etch (e.g., SF_6 plasma) followed by 5:1 BHF.[114]

Deposition of Structural Material For the best step coverage of a structural material over underlying topography, chemical vapor deposition (CVD) is preferred. If a physical deposition method must be used, sputtering is preferred over evaporation (line-of-sight deposition technique), which leads to the poorest step coverage. In the latter case, edge taper could be introduced advantageously.

The most widely used structural material in surface micromachining is polysilicon (poly-Si or simply "poly"). Polysilicon is deposited by low pressure (25–150 Pa) chemical vapor deposition (LPCVD) in a furnace (a polychamber) at about 600°C. The undoped material is usually deposited from pure silane, which thermally decomposes according to Reaction 7.2. Typical process conditions may consist of a temperature of 605°C, a pressure of 550 mTorr (73 Pa), and a silane flow rate of 125 sccm. Under those conditions, a normal deposition rate is 100 Å/min. To deposit a 1-μm film will take about 90 min. Sometimes the silane is diluted by 70–80% nitrogen. The silicon is deposited at temperatures ranging from 570–580°C for fine-grained polysilicon to 620–650°C for coarse-grained polysilicon. The characteristics of these two types of polysilicon materials were compared above (see Table 7.20). Furnace annealing of the polysilicon film at 1050°C in nitrogen for 1 h is commonly used to reduce stress stemming from thermal expansion coefficient mismatch and nucleation and growth of the polysilicon film. Low-temperature deposition of silicon at 320°C in PECVD and at ~350°C in sputter deposition is also possible, although this results in amorphous films.

To make parts of the microstructure conductive, dopants can be introduced in the polysilicon film by adding dopant gases to the silane gas stream, by drive-in from a solid dopant source, or by ion implantation. When doping from the gas phase, the dopant can be readily controlled in the range of 10^{19}–$10^{21}/cm^3$. Polysilicon deposition rates, in the case of gas-phase doping (in situ doping), may be significantly impacted. For example, decreases in polysilicon deposition rate by as much as a factor of 25 have been reported in phosphine and arsine doping. The effect is associated with the poisoning of reaction sites by phosphine and arsine.[115] The lower deposition rate of in situ phosphine doping can be mitigated by reducing the ratio of phosphine to silane flow by one third.[116] With the latter flow regime, deposition at 585°C (for a 2-μm-thick film), followed by 900°C rapid thermal annealing for 7 min, results in a polysilicon with low residual stress, negligible

stress gradient, and low resistivity.[116,117] In situ boron doping, in contrast to arsine and phosphine doping, accelerates the polysilicon deposition rate through an enhancement of silane adsorption induced by the boron presence.[118] Film thickness uniformity for doped films typically is less than 1%, and sheet resistance uniformity is less than 2%. Alternatively, polysilicon may be doped from PSG films sandwiching the undoped polysilicon film as we alluded to when discussing Steps VII and XIII in Figure 7.68. By annealing such a PSG sandwich at 1050°C in N_2 for 1 h, the polysilicon is symmetrically doped by diffusion of dopant from the top and the bottom layers of PSG. Symmetric doping results in a polysilicon film with a moderate compressive stress. The resulting uniform grain texture avoids gradients in the residual stress, which would cause bending moments, warping microstructures on release. Finally, ion implantation of undoped polysilicon, followed by high-temperature dopant drive-in, also leads to conductive polysilicon. This polysilicon has a moderate tensile stress, with a strain gradient that causes cantilevers to deflect toward the substrate.[119] The polysilicon is now ready for patterning by RIE in, e.g., CF_4-O_2 plasma.

Although the mechanical properties are less understood than for single-crystal Si, microstructures based on polysilicon as a mechanical member have been commercialized swiftly.[111,119]

Surface micromachining of thin single-crystalline Si layers in epitaxial layers or fusion-bonded and etched back Si, as well as surface micromachining involving resists such as polyimides, is also very common and is discussed separately below. Other structural materials used in surface micromachining include aluminum, SiO_2, silicon nitride, silicon oxynitride, diamond, SiC, sputtered silicon, GaAs, tungsten, α-Si:H, Ni, W, Al, etc. More information on the material properties of these materials in thin film form is provided in the section on Surface Micromachining Modifications and in Volume III, Chapter 5 (Selected Materials and Processes for MEMS and NEMS).

Base Layer (also Spacer or Sacrificial Layer) Deposition A thin LPCVD phosphosilicate glass (PSG) layer (e.g., 2 μm thick) is a preferred base, spacer, or sacrificial layer material. Adding phosphorous to SiO_2 to produce PSG enhances the etch rate in HF.[120,121] Other advantages to using doped SiO_2 include its utility as a solid-state diffusion dopant source to make subsequent polysilicon layers electrically conductive and helping to control window taper (planarization). As deposited phosphosilicate displays a nonuniform etch rate in HF and must be densified, typically carried out in a furnace at 950°C for 30 min to 1 h in a wet oxygen ambient. The etch rate in BHF can be used as the measure of the densification quality. The base window etching stops at the buffer isolation layer, often a Si_3N_4/SiO_2 layer that also forms the permanent passivation of the device. Windows in the base layer are used to make anchors onto the buffer/isolation layer for mechanical structures.

The edges of the etched windows in the base may need to be tapered to minimize coverage problems with subsequent structural layers, especially if these layers are deposited with a line-of-sight deposition technique. An edge taper is introduced through an optimization of the plasma etch conditions (see Chapter 3), through the introduction of a gradient of the etch rate, or by reflow of the etched spacer. Viscous flow at higher temperature smoothes the edge taper. The ability of PSG to undergo viscous deformation at a given temperature primarily is a function of the phosphorous content in the glass; reflow profiles get progressively smoother the higher the phosphorous concentration—reflecting the corresponding enhancement in viscous flow.[122] Doped oxides used as diffusion sources contain 5–15 wt% of the dopant. Doped oxides for passivation or interlevel insulation contain 2–8 wt% phosphorus to prevent the diffusion of ionic impurities to the device. PSG used for the reflow process contains 6–8 wt% phosphorus and reflows at temperatures between 950 and 1100°C. Oxides with lower phosphorus concentrations will not soften and flow, but higher phosphorus concentrations can give rise to deleterious effects because phosphorus can react with atmospheric moisture to form phosphoric acid, which can consequently corrode the aluminum metallization. Addition of boron to PSG further reduces the reflow temperature without exacerbating this corrosion problem [undoped SiO_2 is often used above

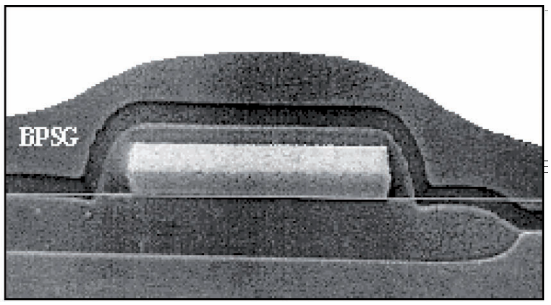


FIGURE 7.70 Edge taper reflow of BPSG layer.

and below PSG or borophosphosilicate glass (BPSG) to prevent corrosion of Al]. BPSG typically contains 4–6 wt% P and 1–4 wt% B and reflows at 800°C. In Figure 7.70 we depict the reflow of a BPSG spacer after patterning in a dry etch; the SEM shows the BPSG oxide layer after an 800°C reflow step, showing smooth topography over the step.

Ion implantation of PSG creates a rapidly etching, damaged PSG layer.[123–125] The steady-state taper is a function of the etch rate in PSG etchants (e.g., BHF).

Selective Etching of Spacer Layer

Selective Etching To create movable micromachines, the microstructures must be freed from the spacer layers. The challenge in freeing microstructures by undercutting is evident from Figure 7.71. After patterning the polysilicon by RIE in, e.g., SF_6 plasma, it is immersed in an HF solution to remove the underlying sacrificial layer, releasing the structure from the substrate. The release step is the source of several technological woes. Commonly, a layer of sacrificial phosphosilicate glass, between 1 and 2000 μm long and 0.1–5 μm thick, is etched in concentrated, dilute, or buffered HF (BHF = BOE). The spacer etch rate, R_s, should be faster than the attack on the microstructural element, R_m, and that of the insulator layer, R_i. For this type of complete undercutting, only wet etchants can be used. Etching narrow gaps and undercutting wide areas with BHF can take hours. The removal rate is this slow because the sacrificial layer is only a few micrometers thick, and the reaction becomes quickly diffusion limited. As a consequence, the depth of sacrificial layer dissolved under the structure will increase slowly with the etching time as $d_{\text{release}} \propto \sqrt{t_{\text{etch}}}$ Simply said, releasing a structure twice as wide will take four times more time. However, if the etching lasts too long, the chemical may start attacking the device structural material, too. Detailed studies on the etching mechanism of oxide spacer layers were undertaken by Monk et al.[126,127] They found that the etching reaction shifts from kinetic controlled to diffusion controlled as the etch channel becomes longer. This affects mainly large-area structures because diffusion limitations were observed only after approximately 200 μm of channel etching or 15 min in concentrated HF. Eaton and Smith[128] developed a release etch model that is an extension of the work done by Monk et al.[126,127] and Liu et al.[129] To shorten the etch time, extra apertures in the microstructures are sometimes provided for additional access to the spacer layer (see Figure 7.72).

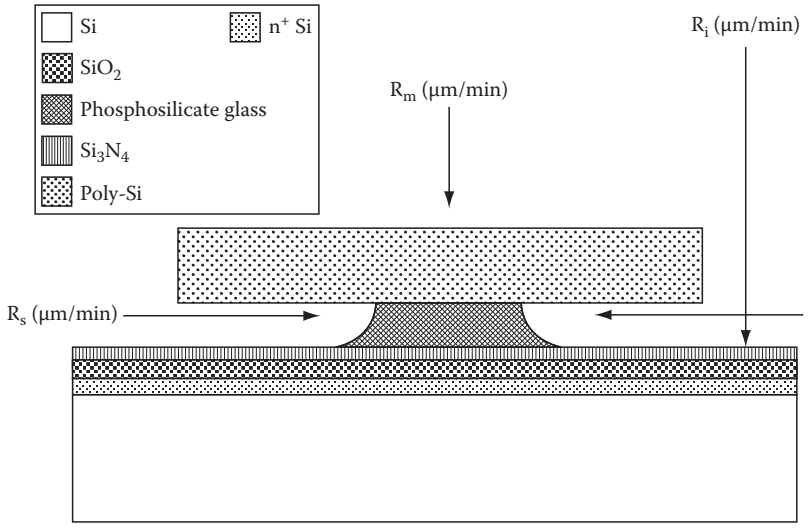


FIGURE 7.71 Selective etching of spacer layer.

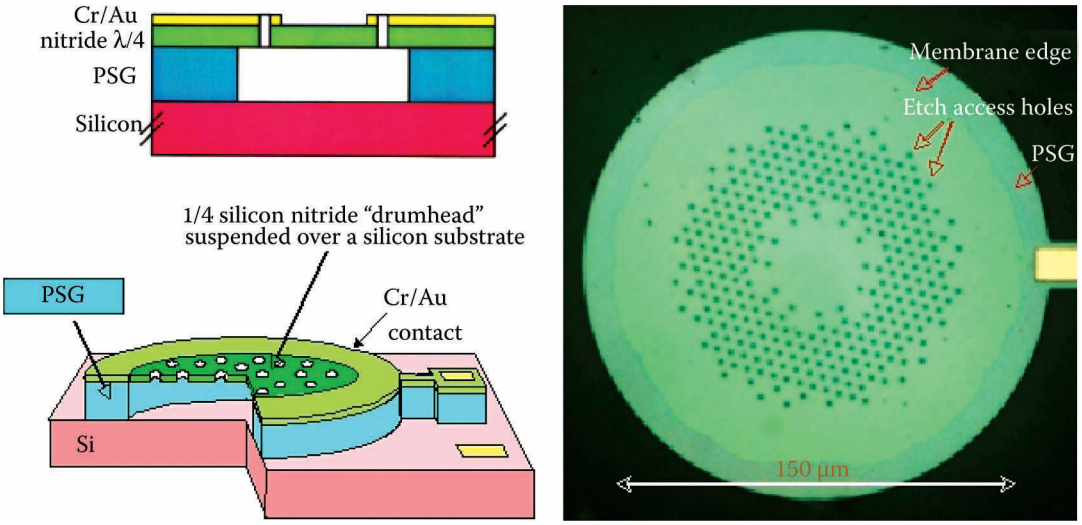


FIGURE 7.72 Analog optical modulator. The 1/4 silicon nitride "drumhead" suspended over a silicon substrate is equipped with access holes to etch the BSG faster. (Goossen, K., J. Walker, and S. Arney. 1994. Silicon modulator based on mechanically-active antireflection layer with 1 Mbit/sec capability. *IEEE Photonics Tech Lett* 6:1119.[130])

These access holes in the structural layer are typically spaced by roughly 30 μm, allowing for the HF to etch all the oxide beneath in less than 5 min.

The etch rate of PSG, the most common spacer material, increases monotonically with dopant concentration, and thicker sacrificial layers etch faster than thinner layers.[126]

The selectivity ratios for spacer layer, microstructure, and buffer layer are not infinite,[131] and in some instances even silicon substrate attack by BHF is observed under polysilicon/spacer regions.[132,133] It also has been shown that during an HF release step the mechanical properties of polysilicon, including residual stress, Young's modulus, and fracture strain, are affected.[134] Heavily phosphorous-doped polysilicon is especially prone to attack by BHF. In general, the Young's modulus and fracture strain of a thin polysilicon film decrease with increasing exposure time to HF and increasing HF concentration. Silicon nitride deposited by LPCVD etches much more slowly in HF than oxide films, making it a more desirable isolation film. When depositing this film with a silicon-rich composition, the etch rate is even slower (15 nm/min).[114] Eaton et al.[135] compared oxide and nitride etching in a 1:1 HF/H_2O and in a 1:1 HF/HCl solution and concluded that the HCl-based etch yielded both faster oxide etch rates (617 vs. 330 nm/min) and slower nitride etch rates (2 vs. 3.6 nm/min), providing a much greater selectivity of the oxide to silicon nitride (310 vs. 91!). The same authors also studied the optimum composition of a sacrificial oxide for the fastest possible etching in their most selective 1:1 HF/HCl etch. The faster sacrificial layer etch limited the damage to nitride structural elements. Their results are summarized in Table 7.21. A densified CVD SiO_2 was used as a control, and a 5%/5% borophosphosilicate glass (BPSG) was found to etch the fastest.

Using low-pressure vapor HF, Watanabe et al.[136] found high etch ratios of PSG and BPSG to thermal oxides of more than 2000, with the BPSG etching slightly faster than the PSG. We will see further that low-pressure vapor HF also leads to less stiction of structural elements to the substrate. In Table 7.22 we present etch rate and etch ratios for R_i and R_s in BHF (7:1) for a few selected materials.

Etching is followed by rinsing and drying. The problems with wet release continue during this drying and

TABLE 7.21 Etch Rate in 1:1 HF/HCl of a Variety of Sacrificial Oxides

Thin Oxide	Lateral Etch Rate (Å/min)
CVD SiO_2 (densified at 1,050 °C for 30 min)	6,170
Ion implanted and densified CVD SiO_2 (P, 8 10^{15}/cm^2, 50 keV)	8,330
Phosphosilicate (PSG)	11,330
5%/5% borophosphosilicate (BPSG)	41,670

Source: Adapted from Eaton, W. P., and J. H. Smith. 1995. *Smart structures and materials 1995: smart electronics (Proceedings of the SPIE).* San Diego: SPIE.[135]

Table 7.22 Etching of Spacer Layer and Buffer Layer in BHF (7:1)

	Material		
Property	LPCVD Si_3N_4	LP CVD SiO_2	LP CVD 7% PSG
Etch rate	7–12 Å/min	700 Å/min	~10,000 Å/min
Selectivity ratio	R_i = 1	R_s = 60–100	R_s ~800–1,200 Å/min

rinsing step. The meniscus created by the receding liquid/air interface tends to pull the structure against the substrate. This intimate contact gives rise to surface forces like Van der Walls force, which will irremediably pin one's structure to the substrate when the drying is complete, effectively destroying the device. This phenomenon is referred to as stiction (see Stiction below). Extended rinsing causes a native oxide to form on the surface of the polysilicon structure. Such a passivation layer often is desirable and can be formed more easily by a short dip in 30% H_2O_2.

Etchant-Spacer-Microstructure Combinations A wide variety of etchant, spacer, and structural material combinations has been used; a limited listing is presented in Table 7.23. One interesting case concerns polysilicon as the sacrificial layer. This was used, for example, in the fabrication of a vibration sensor at Nissan Motor Co.[137] In this case, polysilicon is etched in KOH from underneath a nitride/polysilicon/nitride sandwich cantilever. Also, a solution of HNO_3 and BHF can be used to etch polysilicon, but it proves difficult to control. Using aqueous solutions of NR_4OH, where R is an alkyl group, provides a better etching solution for polysilicon, with greater selectivity with respect to silicon dioxide and phosphosilicate glass. The relatively slow etch rate enables better process control,[138] and the etchant does not contain alkali ions, making it more CMOS compatible. With tetramethyl ammonium hydroxide (TMAH), the etch rate of CVD polysilicon, deposited at 600°C from SiH_4, follows the rates of the (100) face of single-crystal silicon and is dopant dependent. The selectivity of Si/SiO_2 and Si/PSG, at temperatures less than 45°C, is measured to be about 1000. Hence a layer of 500 Å of PSG can be used as the etch mask for 10,000 Å of polysilicon. In addition, porous Si has been used as a sacrificial layer in micromachining. The high surface area of this material makes for rapid etching in KOH.[139]

In Texas Instruments' Digital Micro Device (DMD) (Figure 7.66b) the structural layer is aluminum and the sacrificial layer is a polymer. The polymer is removed with oxygen plasma, and prolonged release time will only slightly affect the metal.

Figure 7.73 illustrates a fabrication sequence for a polyimide corrugated structure. The sacrificial Al in Step 4 may be etched away by a mixture of phosphoric acidic, acetic acid, and nitric acid (see Table 7.23).

Stiction

Stiction during Release The use of sacrificial layers enables the creation of very intricate movable polysilicon surface structures. An important limitation of such polysilicon shapes is that large-area structures tend to deflect through stress gradients or surface tension induced by trapped liquids and attach to the substrate/isolation layer during the final rinsing

Table 7.23 Etchants—Spacer and Microstructural Layer

Etchant	Buffer/Isolation	Spacer	Microstructure
Buffered HF (5:1,NH_4F/conc. HF)	LPCVD Si_3N_4/thermal SiO_2	PSG	Polysilicon[133]
RIE using CHF_3 BHF (6:1)	LPCVD Si_3N_4	LP CVD SiO_2	CVD Tungsten[140]
KOH	LPCVD Si_3N_4/thermal SiO_2	Polysilicon, porous Si (at room temperature)	Si_3N_4[141,142] SOI[139]
Ferric chloride	Thermal SiO_2	Cu	Polyimide[143]
HF	LPCVD Si_3N_4/thermal SiO_2	PSG	Polyimide[144]
Phosphoric/acetic acid/nitric acid (PAN or 5:8:1:1 water/phosphoric/acetic/nitric)	Thermal SiO_2	Al	PE CVD Si_3N_4 Nickel[145,146]
Ammonium iodide/iodine alcohol	Thermal SiO_2	Au	Ti[147]
Ethylene-diamine/pyrocathecol (EDP)	Thermal SiO_2	Polysilicon	SiO_2

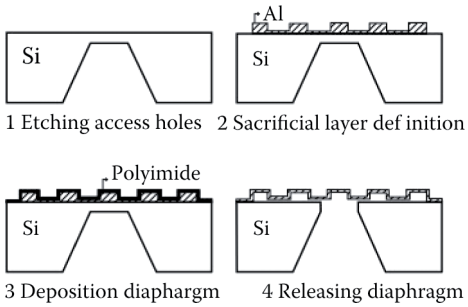


FIGURE 7.73 Schematic view of the fabrication process of a polyimide corrugated diaphragm. (From van Mullem, C. J., K. J. Gabriel, and H. Fujita. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91).* San Francisco: IEEE. With permission.[148])

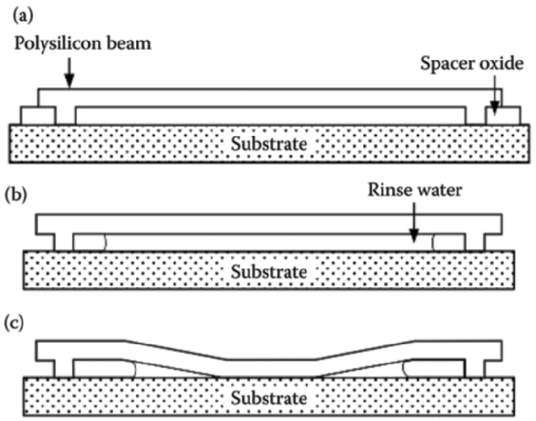


FIGURE 7.74 Stiction phenomenon in surface micromachining and the effect of surface tension on micromechanical structures. (a) Unreleased beam. (b) Released beam before drying. (c) Released beam pulled to the substrate by capillary forces as the wafer dries.

and drying step, a stiction phenomenon that may be related to hydrogen bonding, residual contamination, and the Casimir force (see also Volume I, Chapter 3). In Volume III, Chapter 7 on scaling laws, we will come to appreciate that surface forces become dominant in the microdomain, and not only stiction but also friction are of great concern to the micromachinist. One of the principal causes of malfunctioning in MEMS is stiction, namely, the collapse of movable elements into nearby surfaces, resulting in their permanent adhesion. This can occur during fabrication, especially as a result of capillary forces, or during operation (in-use stiction, see below).

The sacrificial layer removal with a buffered oxide etch (BOE) followed by a long, thorough rinse in deionized water and drying under an infrared lamp typically represent the last steps in the surface micromachining sequence. As the wafer dries, the surface tension of the rinse water pulls the delicate microstructure to the substrate where a combination of forces—van der Waals forces, hydrogen bonding, and Casimir forces—keeps it firmly attached (see Figure 7.74).[119] Once the structure is attached to the substrate by stiction, the mechanical force needed to dislodge it usually is large enough to damage the micromechanical structure.[125,141] The same phenomena are thought to be involved in room temperature wafer bonding (Volume III, Chapter 7). Van der Waals forces and hydrogen bonding are covered in Volume III, Chapter 3 on nanochemistry: bottom-up manufacturing. Casimir forces are quantum mechanical in nature (see Volume I, Chapter 3) and inversely proportional to the fourth power of the distance between the plates. One can quickly establish that two mirrors of 1 cm^2 separated by a distance of 1 μm have an attractive Casimir force of 130 nN (see Volume III, Chapter 8). Although this force appears small, at distances less than 1 μm the Casimir force becomes the dominant force between two neutral objects. At separations of 10 nm, the Casimir effect produces the equivalent of 1 atm! Obviously this is a force to reckon with in MEMS and NEMS devices (see Volume III, Chapter 8 on actuators). We will not further dwell on the mechanics of the stiction process here, but the reader should refer to the theoretical and experimental analysis of the mechanical stability and adhesion of microstructures under capillary forces by Mastrangelo et al.[149,150]

Strategies that are used to overcome stiction have tackled it at the design and fabrication level. An important fabrication idea has been to reduce the contact surface by introducing dimples under the structure. Creating stand-off bumps on the underside of a polysilicon plate[114,151] or adding meniscus-shaping microstructures to the perimeter of the microstructure are mechanical means to help reduce sticking[152] (see Figure 7.75).

Fedder et al.[153] used another mechanical approach to avoid stiction by temporarily stiffening the microstructures with polysilicon links. The very stiff polysilicon structures are not affected by liquid surface tension forces, and the links are severed afterward with a high current pulse once the potentially

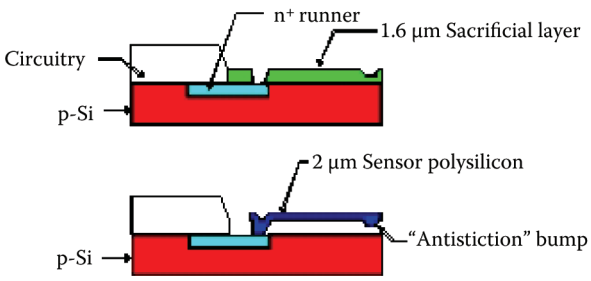


FIGURE 7.75 Stand-off bump or antistiction bump.

destructive processing is complete. Yet another mechanical approach to avoid stiction involves the use of sacrificial supporting polymer columns. A portion of the sacrificial layer is substituted by polymer spacer material, spun on after partial etch of the oxide glass. After completion of the oxide etch, the polymer spacer prevents stiction during evaporative drying. Finally, an isotropic oxygen plasma etches the polymer to release the structure.[154]

Ideally, to ensure high yields, contact between structural elements and the substrate should be avoided during processing. In a liquid environment, however, this may become impossible because of the large surface tension effects. Consequently, most solutions to the stiction problem involve reducing the surface tension of the final rinse solution by physicochemical means. Lober et al.,[131] for example, tried HF vapor, and Guckel et al.[155,156] used freeze-drying of water/methanol mixtures. Freezing and sublimating the rinse fluid in a low-pressure environment gives improved results by circumventing the liquid phase. Takeshima et al.[157] used t-butyl alcohol freeze-drying. Because the freezing point of this alcohol lies at 25.6°C, it is possible to perform freeze-drying without special cooling equipment. Supercritical drying results in increased microstructure yields.[158] With this technique, the rinse fluid is displaced with a liquid that can be driven into a supercritical phase under high pressure. This supercritical phase does not exhibit surface tension, which allows for the drying of released microstructures without sticking. Typically, CO_2 at 35°C and 1100 psi are used (see also Chapter 1).

Kozlowski et al.[159] substituted HF in successive exchange steps by the monomer divinylbenzene to fabricate very thin (500 nm) micromachined polysilicon bridges and cantilevers. The monomer was polymerized under UV light at room temperature and was removed in oxygen plasma. Analog Devices applied a proprietary technique involving only standard IC process technology in the fabrication of a microaccelerometer to eliminate yield losses due to stiction.[118]

Coating the structure with a nonsticking layer [e.g., fluorocarbon, hydrophobic self-assembled mono-layer (SAM)] has also proved successful, and this method, albeit more complex, has the added advantage to provide long-lasting protection (see below under "In-Use Stiction").

The two most popular methods today to avoid stiction are etching a polymer sacrificial layer in an O_2 plasma and using xenon difluoride (XeF_2) to etch sacrificial silicon (see also Chapter 3). Xenon difluoride has been used successfully to release very compliant or nanosized oxide structures where silicon was used as the sacrificial material. The process does not use plasma, making the chamber rather simple, and several manufacturers like XactiX (http://www.xactix.com) (in cooperation with STS) in the United States and PentaVacuum in Singapore offer tools to exploit the technology.

In-Use Stiction Stiction remains a fundamental reliability issue because of contact with adjacent surfaces after release. Stiction-free passivation that can survive the packaging temperature cycle is not known at present.[94] Attempted solutions are summarized below.

Adhesive or stiction energy may be minimized in a variety of ways: for example, by forming bumps on surfaces (see above) or roughening of opposite surface plates.[160] Also, self-assembled monolayer (SAM) coatings have been shown to reduce surface adhesion and to be effective at friction reduction in bearings at the same time.[161] Making the silicon surface very hydrophobic or coating it with diamond-like carbon are other potential solutions for preventing postrelease stiction of polysilicon microstructures. Man et al.[162] eliminate postrelease adhesion in microstructures by using a thin conformal fluorocarbon film. The film eliminates the adhesion of polysilicon beams up to 230 μm long even after direct immersion in water. The film withstands temperatures as high as 400°C, and wear tests show that the film

remains effective after 10^8 contact cycles. Along the same line, ammonium fluoride-treated Si surfaces are thought to be superior to the HF-treated surfaces because of a more complete hydrogen termination, which leads to a cleaner hydrophobic surface.[163] Gogoi et al.[164] introduced electromagnetic pulses for postprocessing release of stuck microstructures.

Wong et al.[165] at the Georgia Institute of Technology are mimicking one of nature's best nonstick surfaces—lotus plant leaves—to help create more reliable MEMS structures unaffected by water and even claim improved biocompatible surfaces able to prevent cells from adhering to implanted medical devices. As we discuss in Volume I, Chapter 6, the lotus plant's ability to repel water and dirt results from an unusual combination of a superhydrophobic (water-repelling) surface and a combination of micrometer-scale hills and valleys and nanometer-scale waxy bumps that create rough surfaces that do not give water or dirt a chance to adhere. With its superhydrophobicity and surface roughness, a lotus surface coating can prevent stiction.

Control of Film Stress

After reviewing typical surface micromachining process sequences, we are ready to investigate some of the mechanical properties of the fabricated mechanical members. Consider a straight lateral resonator as shown in Figure 7.76a. Electrostatic force is applied by a drive comb to a suspended shuttle. Its motion

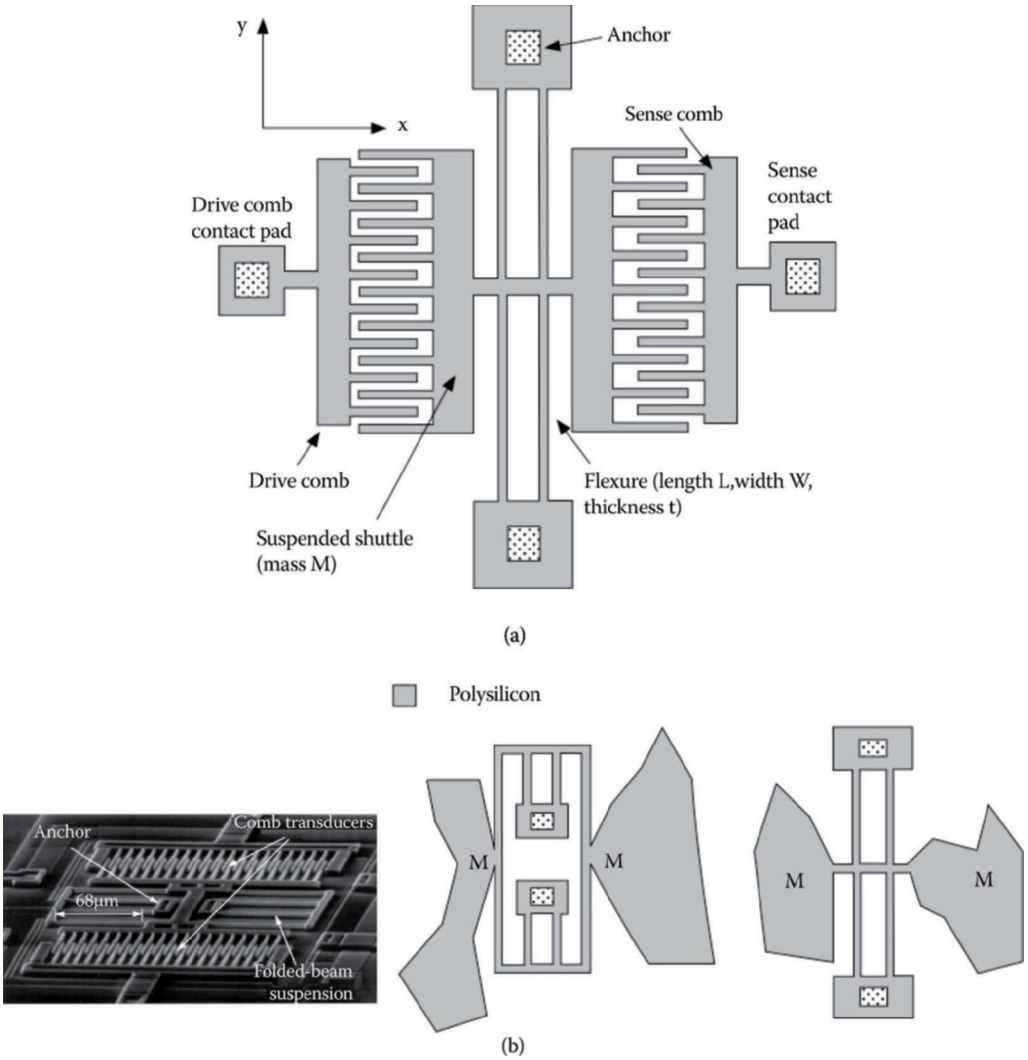


FIGURE 7.76 (a) Layout of a lateral resonator with straight flexures. (b) Folded flexures (left) to release stress are compared with straight flexures (right). *M* is shuttle mass. Scanning electron microscope picture represents a folded beam lateral resonator. (Nguyen, C. T. C., and R. T. Howe. 1993. *Electron devices meeting, 1993. IEDM '93. Technical Digest, International.* Washington DC: IEEE.[166])

is detected capacitively by a sense comb. For many applications, a tight control over the resonant frequency, f_0, is required. An analytical approximation for f_0 of this type of resonator can be deduced from Rayleigh's method:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{2EtW^3}{mL^3} + \frac{24\sigma_r tW}{5\,mL}} \qquad (7.29)$$

where

E = the Young's modulus of polysilicon
L, W, and t = the length, width, and thickness of the flexures, respectively
m = the mass of the suspended shuttle (of the order of 10^{-9} kg or less)

For typical values (L = 150 μm and W = t = 2 μm) and a small tensile residual stress, the resonant frequency f_0 is between 10 and 100 kHz.[94] For typical values of L/W, the stress term in Equation 7.29 dominates the bending term. Any residual stress, σ_r, obviously will affect the resonant frequency. Consequently, stress and stress gradients represent critical stages for microstructural design.

Ignoring the stress term the resonant frequency in Equation 7.29 scales as the square root of the ratio of the stiffness and mass (Figure 7.77) (see also scaling in Volume III, Chapter 7). Thus, frequency scales as L^{-1}, and hence micro- and nanoresonators can lead to very high frequencies.

One of the many challenges of any surface-micromachining process is to control the intrinsic stresses in the deposited films. Several techniques can be used to control film stress. Some we detailed before, but we list them again for completeness sake:

- Large-grained polysilicon films, deposited around 625°C, have a columnar structure and are always compressive. Compressive stress can cause buckling in constrained structures. Annealing at high temperatures, between 900 and 1150°C, in nitrogen significantly reduces the compressive stress in as-deposited polysilicon[167,168] and can eliminate stress gradients. No significant structural changes occur when annealing a columnar polysilicon film. The annealing process is not without danger in cases where active electronics are integrated on the same chip. Rapid thermal annealing might provide a solution (see "IC Compatibility" below).
- Undoped polysilicon films are in an amorphous state when deposited at 580°C or less. The stress and the structure of this low-temperature material depend on temperature and partial pressure of the silane. A low-temperature anneal leads to a fine-grained polysilicon with low tensile stress and very smooth surface texture.[155] Tensile rather than compressive films are a necessity if lateral dimensions of clamped structures are not to be restricted by compressive buckling.
- Phosphorous,[169–171] boron,[170,172,173] arsenic,[170] and carbon[174] doping have all been shown to affect the state of residual stress in polysilicon films. In the case of single-crystal Si, to compensate for strain induced by dopants, one can implant with atoms with the opposite atomic radius versus Si. Similar approaches would most likely be effective for polysilicon as well.
- Tang et al.[114,175] developed a technique that sandwiches a polysilicon structural layer between a top and bottom layer of PSG and lets the high-temperature anneal drive in the phosphorous symmetrically, producing low-stress polysilicon with a negligible stress gradient.
- Another stress reduction method is to vary the materials composition, something readily done in CVD processes. An example of this method is the Si enrichment of Si_3N_4, which reduces the tensile stress.[176,177]
- During plasma-assisted film deposition processes, stress can be influenced dramatically. In a physical deposition process such as sputtering, stress control involves varying gas pressure and substrate bias. In plasma-enhanced

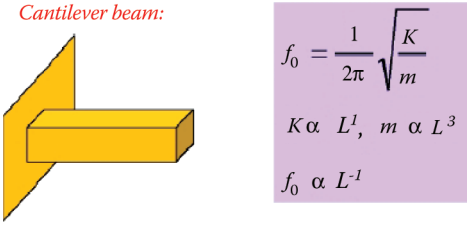


FIGURE 7.77 Scaling of the resonant frequency of simple cantilever beam.

chemical vapor deposition (PECVD), the RF power, through increased ion bombardment, influences stress. In this way, the stress in a thin film starts out tensile, decreases as the power increases, and finally becomes compressive with further RF power increase. PECVD equipment manufacturers are also working to build stress control capabilities into new equipment by controlling plasma frequency (see above). In CVD, stress control involves all types of temperature treatment programs.

- A clever mechanical design might facilitate structural stress relief.[178] By folding the flexures in the lateral resonator in Figure 7.76b and by the overall structural symmetry, the relaxation of residual polysilicon stress is possible without structural distortion. By folding the flexures, the resonant frequency, f_0, becomes independent of σ_r (see Equation 7.29). The springs in the resonator structure provide freedom of travel along the direction of the comb-finger motions (x) while restraining the structure from moving sideways (y), thus preventing the comb fingers from shorting out the drive electrodes. In this design, the spring constant along the y-direction must be higher than along the x-direction, i.e., $k_y >> k_x$. The suspension should allow for the relief of the built-in stress of the polysilicon film and the axial stress induced by large vibrational amplitudes. The folded-beam suspensions meet both criteria. They enable large deflections in the x-direction (perpendicular to the length of the beams) while providing stiffness in the y-direction (along the length of the beams). Furthermore, the only anchor points (see Figure 7.76b) for the whole structure reside near the center, thus allowing the parallel beams to expand or contract in the y-direction, relieving most of the built-in and induced stress.[114] Tang also modeled and built spiral and serpentine springs supporting torsional resonant plates. An advantage of the torsional resonant structures is that they are anchored only at the center, enabling radial relaxation of the built-in stress in the polysilicon film.[114] For some applications, the design approach with folded flexures is an attractive way to eliminate residual stress. However, a penalty for using flexures is increased susceptibility to out-of-plane warpage from residual stress gradients through the thickness of the polysilicon microstructure.[94]
- Corrugated structural members, introduced in the MEMS field by Jerman for bulk micromachined sensors,[179] also reduce stress effectively. In the case of a single-crystal Si membrane, stress may be reduced by a factor of 1,000–10,000.[180] One of the applications of such corrugated structures is the decoupling of a mechanical sensor from its encapsulation by reducing the influence of temperature changes and packaging stress.[180,181] Thermal stress alone can be reduced by a factor of 120.[182] Besides stress release, corrugated structures enable much larger deflections than do similar planar structures. This type of structural stress release was studied in some detail for single-crystal Si,[183] polyimide,[148] and LPCVD silicon nitride membranes,[184] but the quantitative influence of corrugated polysilicon structures still requires further investigation. A schematic view of the fabrication process of a polyimide corrugated diaphragm was shown in Figure 7.73.

Dimensional Uncertainties

The often-expressed concerns about run-to-run variability in material properties of polysilicon or other surface-micromachined materials are somewhat misplaced, Howe points out.[114] He contrasts the relatively large dimensional uncertainties inherent to any lithography technique with polysilicon quality factors of up to 100,000 and long-term (>3 years) resonator frequency variation of less than 0.02 Hz. We follow his calculations here to prove the relative importance of the dimensional uncertainties. The shuttle mass M of a resonator as shown in Figure 7.76 is proportional to the thickness (t) of the polysilicon film, and neglecting the residual stress term, Equation 7.29 reduces to:

$$f_0 \propto \left(\frac{W}{L}\right)^{\frac{3}{2}} \tag{7.30}$$

In case the residual stress term dominates in Equation 7.29, the resonant frequency is expressed as:

$$f_0 \propto \left(\frac{W}{L}\right)^{\frac{1}{2}} \tag{7.31}$$

The width-to-length ratio is affected by systematic and random variations in the masking and etching of the microstructural polysilicon. For 2-µm-thick structural polysilicon, patterned by a wafer stepper and etched in a reactive-ion etcher, a reasonable estimate for the variation in linear dimension of etched features, Δ is about 0.2 µm (10% relative tolerance).

From Equation 7.30 the variation Δ in lateral dimensions will result in an uncertainty δf_0 in the lateral frequency of:

$$\frac{\delta f_0}{f_0} \approx \frac{3}{2}\left(\frac{\Delta}{W}\right) \tag{7.32}$$

for a case where the residual stress can be ignored. With a nominal flexure width of W = 2 µm, the resulting uncertainty in resonant frequency is 15%. For the stress-dominated case, Equation 7.31 indicates that the uncertainty is now given as:

$$\frac{\delta f_0}{f_0} \propto \frac{1}{2}\left(\frac{\Delta}{W}\right) \tag{7.33}$$

The same 2-µm-wide flexure would now lead to a 5% uncertainty in resonant frequency. Interestingly, the stress-free case exhibits the most significant variation in the resonant frequency. In either case, resonant frequencies must be set by some postfabrication frequency trimming or other adjustment.

In Chapter 6 we draw further attention to the increasing loss of relative manufacturing tolerance with decreasing structure size (see Figure 6.3).

Sealing Processes in Surface Micromachining

Sealing cavities to hermetically enclose sensor structures is a significant attribute of surface micromachining. Sealing cavities in surface micromachines often embodies an integral part of the overall fabrication process and presents a desirable chip level, batch-packaging technique. The resulting surface packages (microshells) are much smaller than typical bulk micromachined ones. A sealed cavity made with epi-micromachining is shown in Figure 7.82 further below, and more details on packaging using surface micromachining techniques are provided in Volume III, Chapter 4.

IC Compatibility

Putting detection and signal conditioning circuits right next to the sensing element enhances the performance of the sensing system, especially when dealing with high impedance sensors. A key benefit of surface micromachining, besides small device size and single-sided wafer processing, is its compatibility with CMOS processing. IC compatibility implies simplicity and economy of manufacturing. In the examples at the end of this chapter, we will discuss, for example, how Analog Devices used a mature 4-µm BiCMOS process to integrate electronics with a surface-micromachined accelerometer.

To develop an appreciation of integration issues involved in combining a CMOS line with surface micromachining, we highlight Yun's[185] comparison of CMOS circuitry and surface micromachining processes in Table 7.24A. Surface micromachining processes are similar to IC processes in several aspects. Both processes use similar materials, lithography, and etching techniques. CMOS processes involve at least 10 lithography steps where lateral small feature size plays an important role. Some processing steps, such as gate and contact patterning, are critical to the functionality and performance of the CMOS circuits. Furthermore, each processing step is strongly correlated with other steps. Change in any one of the processing steps will lead to modifications in a number of other steps in the process. In contrast, surface micromachining appears relatively simple. It usually consists of two to six masks, and the feature sizes are much larger. The critical processing steps, such as structural polysilicon, often are self-aligned, which eliminates lithographic alignment. The CMOS process is mature, generic, and fine-tuned, whereas surface micromachining strongly depends on the application and still needs maturing.

Table 7.24B presents the critical temperatures associated with the LPCVD deposition of a

Table 7.24 Surface Micromachining and CMOS

A. Comparison of CMOS and Surface Micromachining

	CMOS	Surface Micromachining
Common features	Silicon-based processes Same materials Same etching principles	
Process flow	Standard	Application specific
Vertical dimension	~1 μm	~1–5 μm
Lateral dimension	<1 μm	2–10 μm
Complexity (number of masks)	>10	2–6

B. Critical Process Temperatures for Microstructures

	Temperature (°C)	Material
LP CVD deposition	450	Low-temperature oxide (LTO)/PSG
LP CVD deposition	610	Low-stress polysilicon
LP CVD deposition	650	Doped polysilicon
LP CVD deposition	800	Nitride
Annealing	950 1,050	PSG densification Polysilicon stress annealing

Source: Yun, W. 1992. *A surface micromachined accelerometer with integrated CMOS detection circuitry*. PhD thesis. University of California, Berkeley.[185]

variety of frequently used materials in surface micromachining. Polysilicon is used for structural layers, and thermal SiO_2, LPCVD SiO_2, and PSG are used as sacrificial layers; silicon nitride is used for passivation. The highest temperature process in Table 7.24B is 1050°C and is associated with the annealing step to release stress in the polysilicon layers. Doped polysilicon films deposited by LPCVD under conventional IC conditions usually are in a state of compression that can cause mechanically constrained structures such as bridges and diaphragms to buckle. The annealing step above at about 1000°C promotes crystallite growth and reduces the strain. If a polysilicon microstructure is built after the CMOS active electronics have been implemented (a so-called *post-CMOS procedure*), temperatures greater than 950°C must be avoided because junction migration will take place at those temperatures. This is especially true with devices incorporating shallow junctions where migration might be a problem at temperatures as low as 800°C. The degradation of the aluminum metallization presents a bigger problem. Aluminum typically is used as the interconnect material in the conventional CMOS process. At temperatures of 400–450°C, the aluminum metallization will start suffering. Anneal temperatures (densification of the PSG and stress anneal of the polysilicon) only account for some of the concerns; in general, several compatibility issues must be considered: 1) deposition and anneal temperatures, 2) passivation during micromachining etching steps, and 3) surface topography.

Yun[185] compared three possible approaches to build integrated microdynamic systems: pre-, mixed, and post-CMOS microstructural processes. He concluded that building up the microstructures after implementation of the active electronics offers the best results.

To avoid problems with microstructure topography, which commonly includes step heights of 2–3 μm, the CMOS module is fabricated before the microstructure module in a post-CMOS process. Although this solves topography problems, it introduces constraints on the CMOS. In a post-CMOS process, the electronic circuitry is passivated to protect it from the subsequent micromachining processes. The standard IC processing may be performed at a regular IC foundry, whereas the surface micromachining occurs as an add-on in a specialized sensor fabrication facility. LPCVD silicon nitride (deposited at 800°C; see Table 7.24B) is stable in HF solutions

and is the preferred passivation layer for the IC during the long-release etching step. PECVD nitride can be deposited at around 320°C, but it displays relatively poor step coverage, whereas pinholes in the film allow HF to diffuse through and react with the oxide underneath. LPCVD nitride is conformably deposited. Although it shows fewer pinholes, circuitry needs to be able to survive the 800°C deposition temperature.

Aluminum metallization must be replaced by another interconnect scheme to increase the post-CMOS temperature ceiling higher than 450°C. Tungsten, which is refractory, shows low resistivity, and has a thermal expansion coefficient matching that of silicon, is an obvious choice. One problem with tungsten metallization is that tungsten reacts with silicon at about 600°C to form WSi_2, implying the need for a diffusion barrier. The process sequence for the tungsten metallization developed at the University of California, Berkeley is shown in Figure 7.78. A diffusion barrier consisting of $TiSi_2$ and TiN is used. The TiN film forms during a 30-s sintering step to 600°C in N_2. Rapid thermal annealing with its reduced time at high temperatures (10 s to 2 min) and high ramp rates (~150°C/s) allows very precise process control and a dramatic reduction of thermal

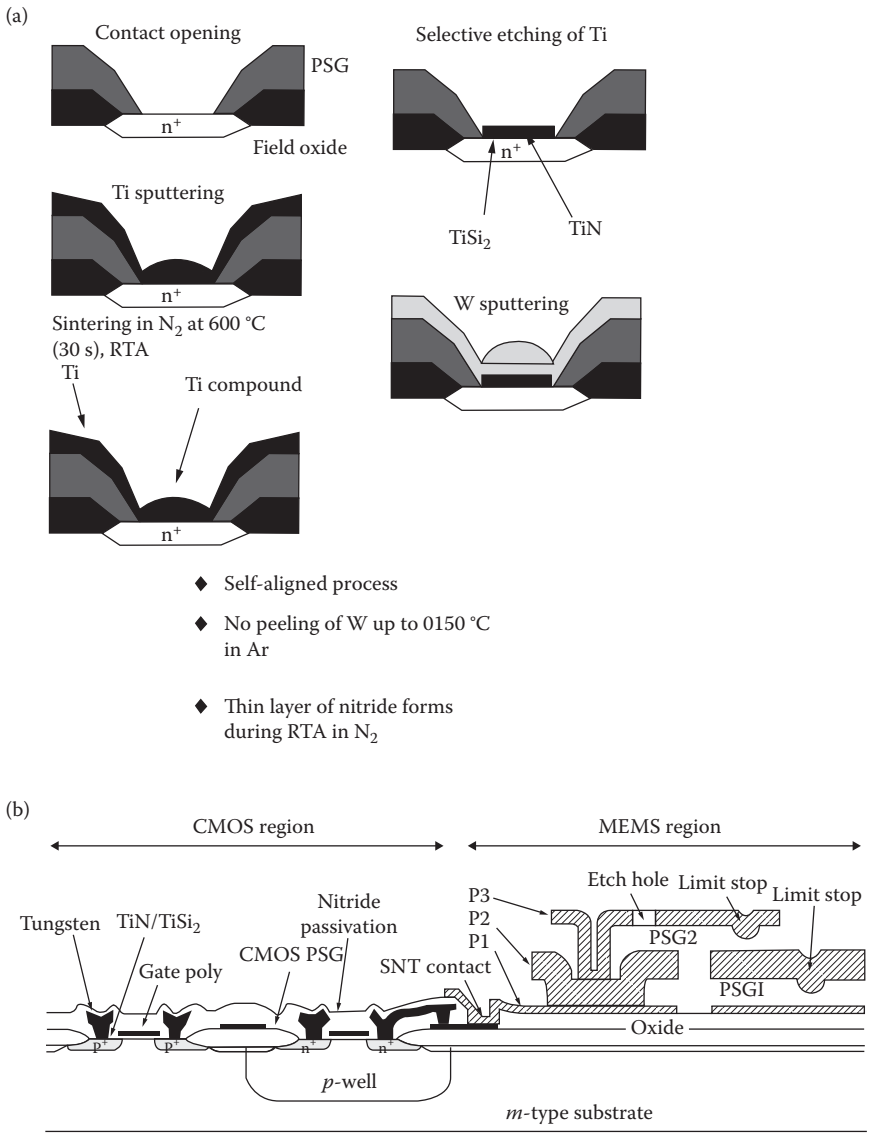


FIGURE 7.78 (a) Tungsten metallization process in a modified CMOS process. (Adapted from Yun, W. 1992. *A surface micromachined accelerometer with integrated CMOS detection circuitry.* PhD thesis. University of California, Berkeley. With permission.[185]) (b) CMOS electronics combined with postprocessed MEMS device.

budgets, reducing duress for the active on-chip electronics. Titanium silicide is formed at the interface of titanium and silicon, whereas titanium nitride forms simultaneously at the exposed surface of the titanium film. The $TiSi_2$/TiN forms a good diffusion barrier against the formation of WSi_2 and at the same time provides an adhesion and contact layer for the W metallization.

To avoid the junction migration in a post-CMOS process, rapid thermal annealing is used for both the PSG densification and polysilicon stress anneal: 950°C for 30 s for the PSG densification and 1000°C for 60 s for the stress anneal of the polysilicon. Alternatively, fine-grained polysilicon can be used because it yields a controlled tensile strain with low-temperature annealing.[186]

Despite some advantages, the post-CMOS process with tungsten metallization is not the preferred implementation. Hillock formation in the W lines during annealing and high contact resistance remains problematic.[116] The heavily doped structural and sacrificial layers may also affect the finely tuned CMOS fabrication sequence. Most importantly, the use of tungsten for circuit interconnects is not consistent with mainstream IC technologies, where aluminum interconnects predominate. Given that IC manufacturers have already invested enormous resources into the development of multilevel aluminum interconnect technologies, and further given the inferior resistivity of tungsten versus aluminum, the described tungsten-based post-CMOS process, although useful as a demonstration tool, is not likely to be adapted in industry.[187] A post-CMOS process for high-aspect-ratio integrated single-crystal silicon microstructures was described in Example 3.3.

The mixed CMOS/micromachining approach implements a processing sequence that puts the processes in a sequence to minimize performance degradation for both electronic and mechanical components. According to Yun, this requires significant modifications to the CMOS fabrication sequence. Nevertheless, Analog Devices relied on such an interleaved process sequence to build the first commercially available integrated microaccelerometer (see Example 7.1). The modifications required on a standard BiCMOS line are minimal: to facilitate integration of the IC and to surface micromachine the thickness of deposited microstructural films, the line is limited to 1–4 μm. Relatively deep junctions permit thermal processing for the sensor polysilicon anneal, and interconnections to the sensor are made only via n^+ underpasses. Therefore, no metallization is present in the sensor area. This industrial solution remains truer to the traditional IC process experience than the post-CMOS procedure. Howe[116] recently detailed another example of such a mixed process In Howe's scenario, the micromachining sequence is inserted after the completion of the electronic structures but before contact etching or aluminum metallization. By limiting the polysilicon annealing to 7 min at 900°C, only minor dopant redistribution is expected. Contact and metallization lithography and etching are more complex now because of the severe topography of the polysilicon microstructural elements. These processes, however, have their own associated limitations: mixed processes often require longer, more expensive development periods for new product lines.

The pre-CMOS approach is to fabricate microstructural elements before any CMOS process steps. At first glance, this seems like an attractive approach because no major modifications would be needed for process integration. However, because of the vertical dimensions of microstructures, step coverage is a problem for the interconnection between the sensor and the circuitry (the latest approach introduced by Howe faces the same dilemma). Passivation of the microstructure during the CMOS process can also become problematic. Furthermore, the heavily doped structural layers can affect the fine-tuned CMOS fabrication sequences, such as gate oxidation. Consequently, this approach is only used for special applications.[185] Precircuit processes may place limitations on foundry-based fabrication schemes because circuit foundries may be sensitive to contamination from MEMS foundries.

A unique pre-CMOS process (micromechanics first) that overcomes the planarity issues of building the MEMS before the CMOS has been developed at Sandia National Labs (http://mems.sandia.gov/tech-info/summit-v.html).[188] The SUMMiT-V™ fabrication process is a batch fabrication process using conventional IC processing tools. Using this

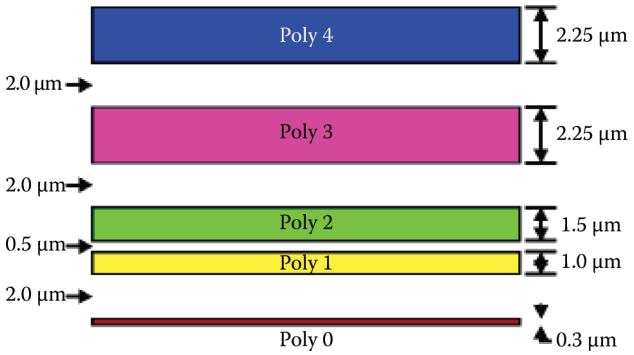


FIGURE 7.79 Sandia's Ultraplanar Multilevel MEMS Technology (SUMMiT-V™) (http://mems.sandia.gov/tech-info/summit-v.html).

technology, high-volume, low-cost production can be achieved. Topography issues are mitigated using chemical-mechanical polishing (CMP) to achieve planarization (see Chapter 1 on lithography), and stress is maintained at low levels using a proprietary process. The process uses 14 individual masks, approximately the same quantity as in many CMOS IC processes (Figure 7.79).

In this approach, micromechanical devices are fabricated in a trench etched in a Si epilayer. After the mechanical components are complete, the trench is filled with oxide, planarized using CMP, and sealed with a nitride membrane. The flat wafer with the embedded micromechanical devices is then processed by means of conventional CMOS processing. Additional steps are added at the end of the CMOS process to expose and release the embedded micromechanical devices and to hook them up with the CMOS. A cross-section of a structure made with this technology is shown in Figure 7.80.

The SPIE "Smart Structures and Materials 1996" meeting in San Diego, CA, had two complete sessions dedicated to the crucial issue of integrating electronics with polysilicon surface micromachining.[189] Research aimed at achieving truly modular merged circuits + microstructures technology is still ongoing.

Surface Micromachining Modifications

Silicon on Insulator—SOI Surface Micromachining

Introduction SOI is an exciting new approach to both IC chip making and MEMS. In SOI, bulk silicon wafers are replaced with wafers that have three layers: a thin surface layer of silicon (from a few hundred angstroms to several micrometers thick), an underlying layer of insulating material, and a support or "handle" silicon wafer. The insulating layer, usually made of silicon dioxide, is called the *buried oxide* or *BOX* and is typically a few thousand angstroms thick. As we will see below, this thin buried oxide layer, sandwiched between two layers of Si, can be achieved in several different ways.

When transistors are built within the thin top silicon layer, they switch signals faster (up to 10 GHz), run at lower voltages, and are much less vulnerable to signal noise from background cosmic ray particles. Furthermore, on an SOI wafer each transistor is isolated from its neighbor by a complete layer of silicon dioxide, which makes them immune to "latch-up" problems, and they can be spaced closer together than transistors built on bulk silicon wafers. Thus, building circuits on SOI allows for more compact

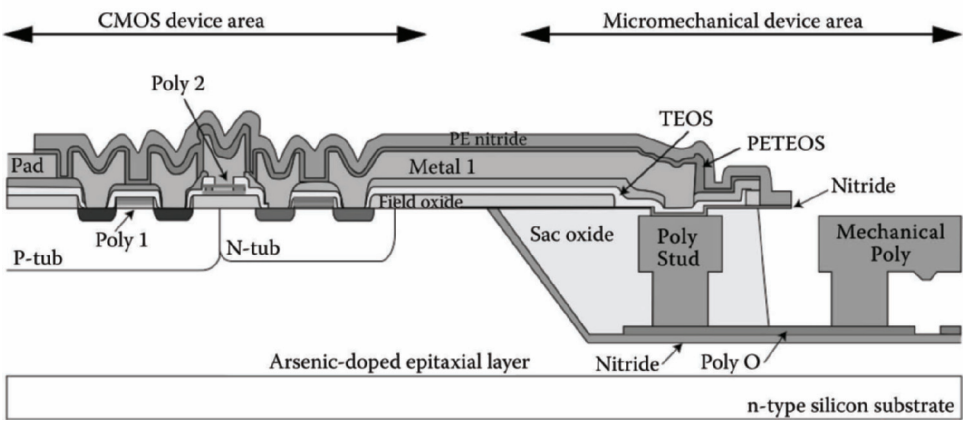


FIGURE 7.80 Schematic cross-section of the embedded micromechanics approach to Sandia's CMOS/MEMS integration.

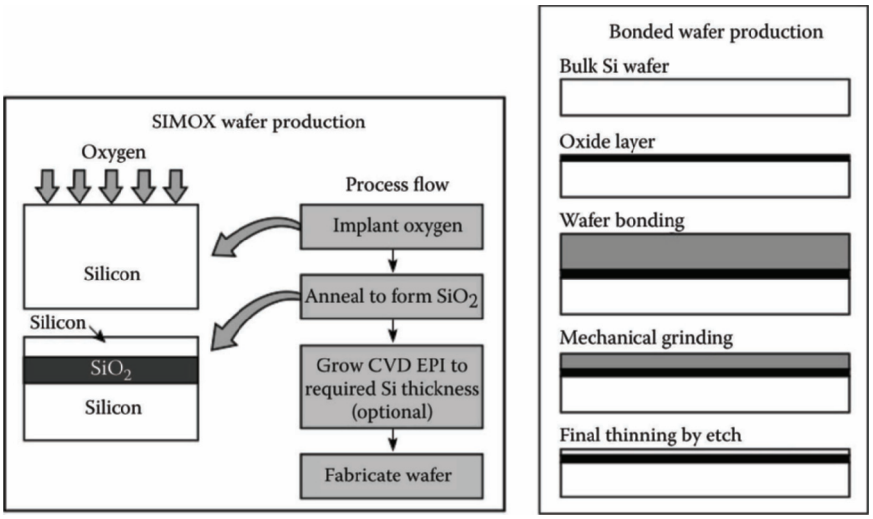


FIGURE 7.81 How SOI wafers are made.

chip designs, resulting in smaller IC devices (with higher production yield) and more chips per wafer (increasing fabrication productivity).

In MEMS, silicon crystalline features, anywhere between fractions of a micrometer to 100 μm high, can be readily obtained by surface micromachining of the epi silicon or fusion-bonded silicon layer of SOI wafers.[99] Structural elements made from these single-crystalline Si layers result in more reproducible and reliable sensors than those made from polysilicon. SOI or epi-micromachining combines the best features of surface micromachining (i.e., IC compatibility and freedom in *x*,*y* shapes) with the best features of bulk micromachining (superior single-crystal Si properties). Moreover, SOI surface micromachining frequently involves fewer process steps and offers better control over the thickness of crucial building blocks. Given the poor reproducibility of mechanical properties and generally poor electronic characteristics of polysilicon films, SOI machining may surpass the polysilicon technology for fabricating high-performance devices.

SOI Surface Micromachining Below, we present a short review of the three most popular methods for producing SOI wafers, introduce some uses of SOI in MEMS, and compare SOI versus polysilicon surface micromachining.

SOI Wafer Fabrication Techniques Three major techniques currently are applied to produce SOI wafers (Figure 7.81) (see also under "Epitaxy" above): SIMOX (Separated by IMplanted OXygen), the Si fusion bonded (SFB) wafer technique, and zone-melt recrystallized (ZMR) polysilicon. With SIMOX, standard Si wafers are implanted with oxygen ions and then annealed at high temperatures (1300°C). The oxygen and silicon combine to form a silicon oxide layer beneath the silicon surface. The oxide layer's thickness and depth are controlled by varying the energy and dose of the implant and the anneal temperature. In some cases, a CVD process deposits additional epitaxial silicon on the top silicon layer. Attempts have also been made to implant nitrogen in Si to create abrupt etch stops [in the SIMON process (Separation by IMplantation Of Nitrogen]. At high-enough energies, the implanted nitrogen is buried 0.5–1 μm deep. At a high-enough dose, the etching in that region stops. It is not necessary to implant a stoichiometric amount of nitrogen; a dose lower by a factor of two to three suffices. After implantation, it is necessary to anneal the wafer because the implantation destroys the crystal structure at the surface of the wafer.

The Si fusion bonded wafer process starts with an oxide layer (typically about 1 μm) grown on a standard Si wafer. That wafer is then bonded to another wafer, with the oxide sandwiched between. For the bonding, no mechanical pressure or other forces are applied. The sandwich is annealed at 1100°C for 2 h in a nitrogen ambient, which creates a strong bond between the two wafers. One of the wafers is then

ground to a thickness of a few micrometers using mechanical and CMP.

A third process for making SOI structures is to recrystallize polysilicon (e.g., with a laser, an electron beam, or a narrow strip heater), which has been deposited on an oxidized silicon wafer. This process is called *zone melting recrystallization* (ZMR) and is used primarily for local recrystallization but has not yet been explored much for use in micromachining applications.

The crystalline perfection of conventional silicon wafers in SFB and ZMR is completely maintained in the SOI layer because the wafers do not suffer from implant-induced defects. By using plasma etching, wafers with a top Si layer thickness of as little as 1000–3000 Å with total thickness variations of less than 200 Å can be made.[190] SOI top Si layers of 2 µm thick are more standard.[191] A tremendous amount of effort is spent in the IC industry on controlling the SOI thin Si layer thickness, which benefits any narrow tolerance IC and micromachine.

Etched-back fusion-bonded Si wafers and SIMOX are used extensively to build both ICs and micromachines, and both are commercially available. In 1989 TI introduced its commercial 64K static random access memory (SRAM) based on SOI, and AMD's 90 nm is also based on SOI technology (introduced in late 2004). One vendor is SOITEC (http://www.soitec.com/en/products).

SOI Use in MEMS

Overview Kanda[192] reviews different types of SOI wafers in terms of their micromachining and IC applications. Working with SOI wafers, he points out, offers several advantages over bulk Si wafers: fewer process steps are needed for feature isolation; parasitic capacitance is reduced; and power consumption is lowered. In the IC industry, SOI wafers are principally used for high-speed CMOS ICs (<10 GHz), smart power ICs (voltages up to 100 V), 3D ICs, and radiation-hardened devices.[193]

An important use of silicon on insulator wafers is in the production of high-temperature sensors. Compared with p-n junction isolation, which is limited to about 125°C, much higher temperature (up to 300°C) devices are possible based on the dielectric insulation of SOI. A wide variety of SOI surface-micromachined structures have been explored, including pressure sensors, accelerometers, torsional micromirrors, light sources, and optical choppers.[95,99] Often in those micromachining applications, SOI wafers are used to produce an etch stop. The silicon fusion bonded (SFB) method offers the more versatile MEMS approach because of the associated potential for thicker single-crystal layers and the option of incorporating buried cavities, facilitating micromachine packaging. Sensors manufactured by means of SFB now are commercially available.[194] The SIMOX approach is less labor intensive and holds better membrane thickness control. An important expansion of the SOI technique is selective epitaxy. The latter enables a wide range of new mechanical structures and enables novel etch-stop methods,[195] as well as electrical and/or thermal separation and independent optimization of active sensor and readout electronics.[196] In most cases, SOI machining involves dry anisotropic etching to etch a pattern into the Si layer on top of the insulator. These structures then are released by etching the sacrificial buried SiO_2 insulator layer, which displays a thickness with very high reproducibility (400 ± 5 nm) and uniformity (<±5 nm), especially in the case of SIMOX. Etched free cantilevers and membranes consist of single-crystalline silicon with thicknesses ranging from micrometers and submicrometers (SIMOX) up to hundreds of micrometers (SFB).

The increasing use of SOI technology for the fabrication of MEMS devices has brought with it its own set of unique problems that must be overcome for applications where CD control is important. For example, the etching of deep trenches in SOI wafers necessitates a fast etch process, followed by a slow etch process as soon as the buried dioxide layer is reached.[197] The main problem in plasma processing of SOI is the notching at the oxide interface, which results from the deflection of the incident ions by charging up the insulating oxide surface. The notch width increases with time, and it is here that etch uniformity begins to play an important role because overetching times to clear the wafer can become significant. For example, a 400-µm-deep etch with +5% uniformity requires a >10% overetch to ensure that all features etch to the required depth. Thus, some areas of the wafer will be subject to a 40-µm

overetch, and precise control of the profile during overetching becomes essential.

Creating Buried Cavities in Silicon by Silicon Fusion Bonded Micromachining Silicon fusion bonding enables the formation of thick single-crystal layers with cavities built in. An example is shown in Figure 7.82.[99] The device pictured involves two 4-in. <100> wafers: a handle wafer and a wafer used for the SOI surface. The p-type (3–7 Ωcm) handle wafer is thermally oxidized at 1100°C to obtain a 1-µm-thick oxide. Thermal oxidation enables thicker oxides than the ones formed in SIMOX by ion implantation and avoids the potential implantation damage in the working material. To make a buried cavity, the oxide is patterned and etched. To produce yet deeper cavities, the Si handle wafer may be etched as well (as in Figure 7.82). In the case shown, the top wafer consists of the same p-type substrate material as the handle wafer with a 2–30-µm-thick n-type epitaxial layer. The epitaxial layer determines the thickness of the final mechanical/structural element. The epitaxial layer is fusion bonded to the cavity side of the handle wafer (2 h at 1100°C). The top wafer is then partially thinned by grinding and polishing (Figure 7.82a). An insulator is deposited and patterned on the back side of the handle wafer to etch access holes to the insulator. After the insulator at the bottom of the etch hole is removed by a buffered oxide etch (BOE), aluminum is sputtered and sintered to make contact to the n-type epilayer for the electrochemical etch back of the remaining p-type material (Figure 7.82b). The final single-crystal silicon thickness is uniform to within ±0.05 µm (standard deviation) and does not require a costly, high-accuracy polish step.

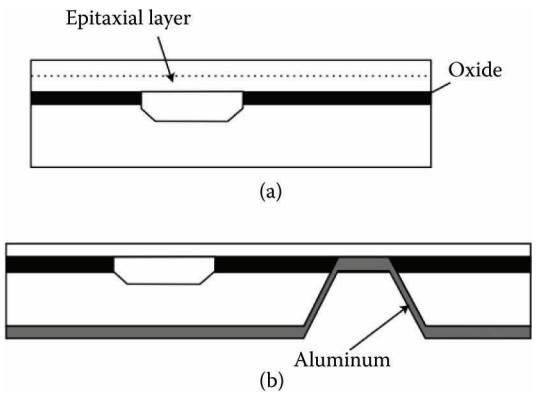


FIGURE 7.82 (a) Wafer sandwich after grind-and-polish step. (b) Wafer after electrochemical etch-back in KOH, buried oxide removal, and aluminum deposition. [From Noworolski, J. M., E. Klaassen, J. Logan, K. Petersen, and N. Maluf. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95).* Stockholm, Sweden: IEEE. With permission.[99]]

Draper Laboratory is using SOI processes in the development of inertial sensors, gyros, and accelerometers as an alternative to devices fabricated by the dissolved wafer process (see Example 4.1). The main advantage is that the former consists of an all-Si process rather than a Si/Pyrex sandwich; an all Si device will exhibit less thermal stress.[198]

Fabricating Pressure Sensors in SIMOX Surface Micromachining Both capacitive and piezoresistive pressure sensors were microfabricated from SIMOX wafers.[95] Figure 7.83 illustrates the process sequence by Diem et al.[95] for fabricating an absolute capacitive pressure sensor. The 0.2-µm silicon surface layer of the SIMOX wafer is thickened with doped epi-silicon to 4 µm. An access hole is RIE etched in the Si layer, and vacuum cavity and electrode gap are obtained by etching the SiO_2 buried layer. Because the buried thick oxide layer exhibits a very high reproducibility and homogeneity over the whole wafer (0.4 µm ± 5 nm), the resulting vacuum cavity and electrode gap after etching also are very well controlled. The small gap results in relatively high capacitance values between the free membrane and bulk substrate (20 pf/mm^2). Diaphragm diameter, controlled by the SiO_2 etching, is up to several hundred micrometers (±2 µm). The etching hole is hermetically sealed under vacuum by plasma CVD deposition of nonstressed dielectric layer plugs.

With the above scheme, Diem et al. realized an absolute pressure sensor with a size of less than 1.5 mm^2. The temperature dependence of a capacitive sensor is mainly a result of the temperature coefficient of the offset capacitance. Therefore, temperature compensation is needed for high accuracy sensors. A drastic reduction of the temperature dependence is obtained by a differential measurement, especially if the reference capacitor resembles the sensing capacitor. A reference capacitor is designed with the membrane blocked by several plugs for pressure

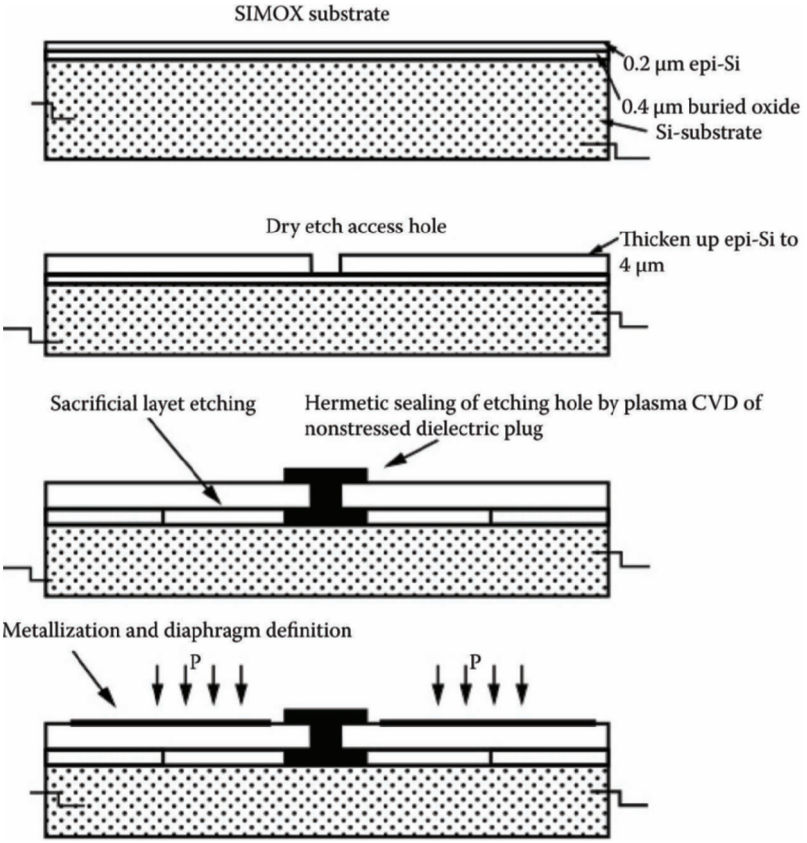


FIGURE 7.83 Process sequence of a SIMOX absolute capacitive pressure sensor by Diem et al. [From Diem, B., M. T. Delaye, F. Michel, S. Renard, and G. Delapierre. 1993. *Seventh international conference on solid-state sensors and actuators (Transducers '93).* Yokohama, Japan: IEEE. With permission.[95]]

insensitivity. The localization and the number of plugs are modeled by finite element analysis (FEA; ANSYS software was used) to get a deformation lower than 1% of the active sensor's deformation. Even without temperature calibration, the high output of the differential signal results in an overall output error better than ±2% over the whole temperature range (–40°C to +125°C) compared with 10% for nondifferential measurements. The temperature coefficient of the sensitivity is about 100 ppm/°C, which agrees with the theoretical variation of the Young's modulus of silicon. A piezoresistive sensor can be achieved by implanting strain gauges in the membrane. Although SIMOX wafers are more expensive than regular wafers, they come with several process steps embedded, and they make packaging easier.

Selective Epitaxy Surface Micromachining In the discussion on epitaxy, we drew attention to the potential of selective epitaxy for creating novel microstructures. The example in Figure 7.50 illustrates the selective deposition of epi-silicon on a Si substrate through a SiO_2 window. The same figure also demonstrates the simultaneous deposition of polysilicon on SiO_2 and crystalline epi-silicon on silicon, creating the basis for a structure featuring an epi-silicon anchor with polysilicon side arms.

Neudeck et al.[199,200] at Purdue and Gennissen et al.[195,196] in Twente proved that selective epitaxy can also be applied for automatic etch stop on buried oxide islands. Figure 7.84a demonstrates how epitaxial lateral overgrowth (ELO) can bury oxide islands. After removal of the native oxides from the seed windows, epi is grown for 20 min at 950°C and at 60 Torr using a $Si_2H_2Cl_2$-HCl-H_2 gas system. The epi growth front moves parallel to the wafer surface while growing in the lateral direction, leaving a smooth planar surface. During epi growth, the HCl prevents polynucleation on the nonsilicon areas. The epi quality is strongly dependent on the orientation of the seed holes

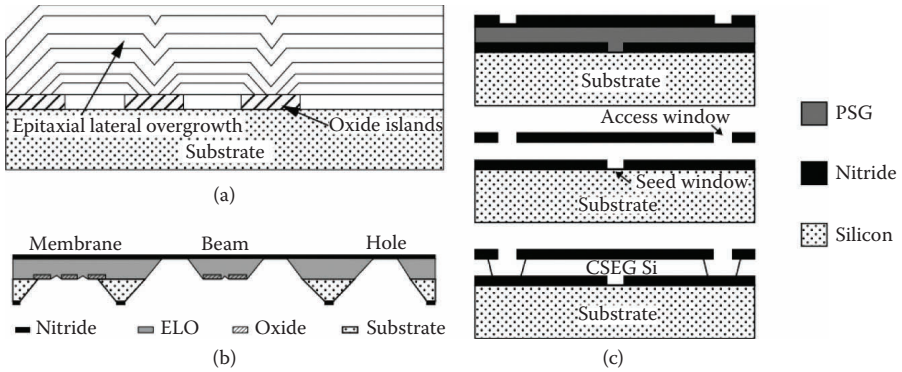


FIGURE 7.84 Micromachining with epi-silicon. (a) Lateral overgrowth process of epi-silicon (ELO, epitaxial lateral overgrowth) out of <100>-oriented holes in an oxide mask. (b) KOH etch stop on buried oxide islands or front side nitride. (c) Principle of confined selective epitaxial growth. (From Bartek, M., P. T. J. Gennissen, P. J. French, and R. F. Wolffenbuttel. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95).* Stockholm, Sweden: IEEE. With permission.[196])

in the oxide. Seed holes oriented in the <100> direction lead to the best epi material and surface quality. Selective epi's other big problem for fabrication remains sidewall defects.[201] The buried oxide islands stop the KOH etch of the substrate, enabling formation of beams and membranes as shown in Figure 7.84b. This technique might form the basis of many high-performance microstructures. The Purdue and Twente groups also work on confined selective epitaxial growth (CSEG), a process pioneered by Neudeck et al.[200] In this process a micromachined cavity is formed above a silicon substrate with a seed contact window to the silicon substrate and access windows for epi-silicon (Figure 7.84c).[196] Low-stress, silicon-rich nitride layers act as structural layers to confine epitaxial growth; PSG is used as sacrificial material. This confined selective epitaxial growth technique allows electrical and/or thermal isolation separation, as well as independent optimization of active sensor and readout electronic areas.

SOI versus Polysilicon Surface Micromachining

The power of polysilicon surface micromachining mainly lays in its CMOS compatibility. When deposited on an insulator, both polysilicon and single crystal layers enable higher operating temperatures (>200°C) than bulk micromachined sensors featuring p-n junction isolation only (130°C max).[202] An additional benefit for SOI-based micromachining is IC compatibility combined with single-crystal Si performance excellence. The maximum gauge factor (see Volume I, Equation 4.93) of a polysilicon piezoresistor is about 30, roughly 15 times larger than that of a metal strain gauge but only one third that of an indiffused resistor in single-crystal Si.[203] Higher piezoresistivity and fracture stress would seem to favor SOI for sensor manufacture. However, there is an important counterargument: the piezoresistivity and fracture stress in polysilicon are isotropic, a major design simplification. Moreover, by laser recrystallization, the gauge factor of polysilicon might increase to greater than 50,[204] and by appropriate boron doping the temperature coefficient of resistance (TCR) can reach 0 versus a TCR of, e.g., 1.7×10^{-3} K^{-1} for single-crystal p-type silicon. Neither technical nor cost issues will be the deciding factor in determining which technology will become dominant in the next few years. Micromachining is very much a hostage to trends in the IC industry; promising technologies such as GaAs and micromachining do not necessarily take off, in no small part because of the invested capital in some limited sets of standard silicon technologies. On this basis, SOI surface micromachining is the favored candidate; SOI extends silicon's technological relevance and experiences, increasing investment from the IC industry and thus benefiting SOI micromachining.[192]

Based on the above, we believe that SOI micromachining not only introduces an improved method of making many simple micromachines, but it also will probably become the favored

approach of the IC industry. A summary of SOI advantages is listed below:

- IC industry use in all type of applications such as MOS, bipolar digital, bipolar linear, power devices, BiCMOS, CCDs, and heterojunction bipolar[205]
- Batch packaging through embedded cavities
- CMOS compatibility
- Substrate industrially available at increasingly lower costs (about $200 today)
- Excellent mechanical properties of the single-crystalline surface layer
- Freedom of shapes in the *x*-*y* dimensions and continually improving dry etching techniques, resulting in larger aspect ratios and higher features
- Freedom to choose a very well-controlled range of thicknesses of epi surface layers
- SiO_2 buried layer as sacrificial and insulating layer and excellent etch stop
- Dramatic reduction of process steps as the SOI wafer comes with several "embedded" process steps
- High-temperature operation

A tutorial on SOI can be found at http://www.freescale.com/webapp/sps/site/overview.jsp?code=TM_SOI.

Hinged Polysilicon

One way to achieve high vertical structures with surface micromachining is to build large, flat structures horizontally and then rotate them on a hinge to an upright position. Pister et al.[194–203] developed the polysilicon hinges shown in Figure 7.85A; on these hinges, long structural polysilicon features (1 mm and beyond) can be rotated out of the plane of a substrate. To make these hinged structures, a 2-µm-thick PSG oxide layer (PSG-1) is deposited on the Si substrate as the sacrificial material, followed by the deposition of the first polysilicon layer (2-µm-thick Poly 1). Photolithography and dry etching to form the desired structural elements, including hinge pins to rotate them, pattern the polysilicon structural layer. Following the deposition and patterning of Poly 1, another layer of sacrificial material (PSG-2) of 0.5-µm thickness is deposited (Figure 7.85Aa). Contacts are made through both PSG layers to the Si substrate, and a second layer of polysilicon is deposited and patterned (Poly 2), forming a staple to hold

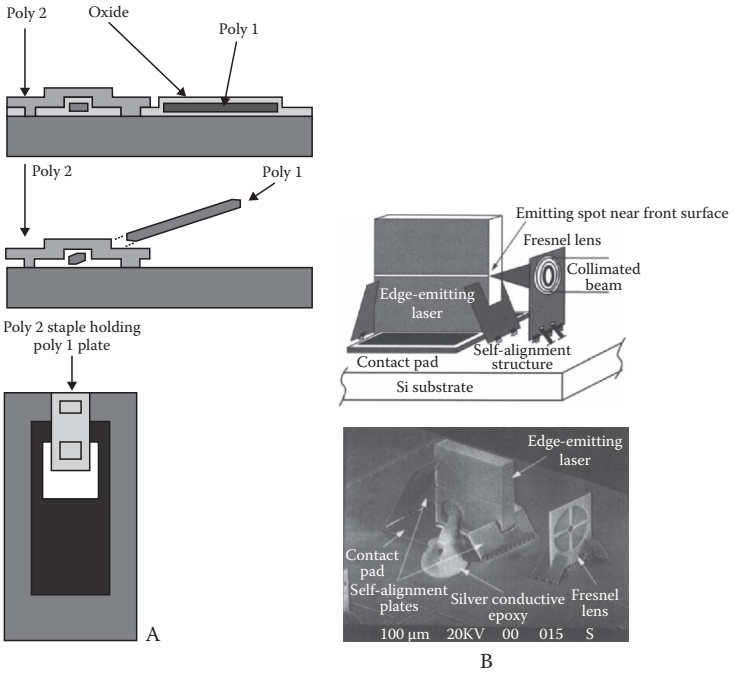


FIGURE 7.85 Microfabricated hinges. (A) Cross-section, side view, and top view of a single-hinged plate before and after the sacrificial etch. SEM of a hinge (MCNC). (B) Schematic (top) and SEM micrograph of the self-aligned hybrid integration of an edge-emitting laser with a micro-Fresnel lens. (From Lin, L. Y., S. S. Lee, M. C. Wu, and K. S. J. Pister. 1995. *Proceedings: IEEE micro electro mechanical systems (MEMS '95)*. Amsterdam, the Netherlands: IEEE. With permission.[206])

the first polysilicon layer hinge to the surface (Figure 7.85Ab). The first and second layers of poly are separated everywhere by PSG-2 for the first polysilicon layer to freely rotate off the wafer surface when the PSG is removed in a sacrificial etch. After the sacrificial etch, the structures are rotated in their respective positions (Figure 7.85Ac). This is accomplished in an electrical probe station by skillfully manipulating the movable parts with the probe needles. Once the components are in position, high friction in the hinges tends to keep them in the same position. To obtain more precise and stable control of position, additional hinges and supports are incorporated. To provide electrical contact to the vertical polysilicon structures, one can rely on the mechanical contact in the hinges, or polysilicon beams (cables) can be attached from the vertical structure to the substrate.

Pister's research team made a wide variety of hinged microstructures, including hot wire anemometers, a box dynamometer to measure forces exerted by embryonic tissue, a parallel-plate gripper,[207] a microwindmill,[208] a microoptical bench (MOB) for free-space integrated optics,[206] and a standard CMOS single piezoresistive sensor to quantify the single heart cell contractile forces of a rat.[209] One example from this group's efforts is illustrated in Figure 7.85B, showing an SEM photograph of an edge-emitting laser diode shining light onto a collimating micro-Fresnel lens.[206] The micro-Fresnel lens in the SEM photo is surface micromachined in the plane and erected on a polysilicon hinge. The lens has a diameter of 280 μm. Alignment plates at the front and the back sides of the laser are used for height adjustment of the laser spot so that the emitting spot falls exactly onto the optical axis of the micro-Fresnel lens. After assembly, the laser is electrically contacted by silver epoxy. Although this hardly outlines standard IC manufacturing practices, excellent collimating ability for the Fresnel lenses has been achieved. The eventual goal of this work is a microoptical bench (MOB), in which microlenses, mirrors, gratings, and other optical components are prealigned in the mask layout stage using computer-aided design. Additional fine adjustment would be achieved by on-chip microactuators and micropositioners such as rotational and translational stages. Erecting these polysilicon structures with the probes of an electrical probe station and, occasionally, assembly by chance in the HF etch or deionized water rinse represent too complicated or too unreliable postrelease assembly methods for commercial acceptance. At the University of Illinois (Urbana), the Micro Actuator, Sensors, and Systems Group headed by Dr. Chang Liu uses magnetic actuation to position a large array of hinged microstructures in parallel. A magnetic material, Permalloy, is electroplated and integrated into the hinged microstructures. The result may be improved manufacturability and stability of these delicate structures.

Friction in polysilicon joints is high because friction is proportional to the surface area (s^2) and becomes dominant over inertial forces (s^3) in the microdomain (see Volume III, Chapter 7). Such joints are not suitable for microrobotic applications. Although attempts have been made to incorporate polysilicon hinges in such applications,[210] plastically deformable hinges make more sense for microrobot machinery involving rotation of rigid components. Noting that the external skeleton of insects incorporates hard cuticles connected by elastic hinges, Suzuki et al.[146] fabricated rigid polysilicon plates (E = 140 GPa) connected by elastic polyimide hinges (E = 3 GPa) as shown in Figure 7.86. Holes in the polysilicon plates shorten the PSG etch time compared with plates without holes. The plates without holes remain attached to the substrate, whereas the ones with holes are completely freed. Using electrostatic actuators, the structure shown in Figure 7.85 can be made to flap like the wing of a butterfly. By

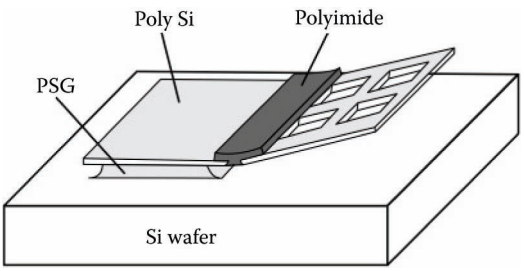


FIGURE 7.86 Flexible polyimide hinge and polysilicon plate (butterfly wing). (From Kim, Y. W., and M. G. Allen. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91).* San Francisco: IEEE. With permission.[143])

applying an AC voltage of 10 kHz, resonant vibration of such a flapping wing was observed.[144]

Hoffman et al.[211] demonstrated aluminum plastically deformable hinges on oxide movable thin plates. Oxide plates and Al hinges were etched free from a Si substrate by using XeF_2, a vapor phase etchant exhibiting excellent selectivity of Si over Al and oxide. According to the authors, this process, because of its excellent CMOS compatibility, might open the way to designing and fabricating sophisticated integrated CMOS-based sensors with rapid turnaround time (see also Chapter 3).

Milliscale Molded Polysilicon Structures

The manual assembly of fabricated microparts complicates assembly of tall 3D features in the hinged polysilicon approach we just described, which is similar to building a miniature boat in a bottle. Keller, while at the University of California, Berkeley, came up with an elegant alternative for building tall, high-aspect-ratio microstructures in a process that does not require postrelease assembly steps.[97] The technique involves deep dry etching of trenches in a Si substrate, deposition of sacrificial and structural materials in those trenches, and demolding of the deposited structural materials by etching away the sacrificial materials. CVD processes can typically only deposit thin films (~1–2 μm) on flat surfaces. If, however, these surfaces are the opposing faces of deep narrow trenches, the growing films will merge to form solid beams. In this fashion, high-aspect-ratio structures that would normally be associated with LIGA now also can be made of CVD polysilicon. The procedure is illustrated in Figure 7.87.[97,212] The first step is to etch deep trenches into a silicon wafer. The depth of the trenches equals the height of the desired beams and is limited to about 100 μm with aspect

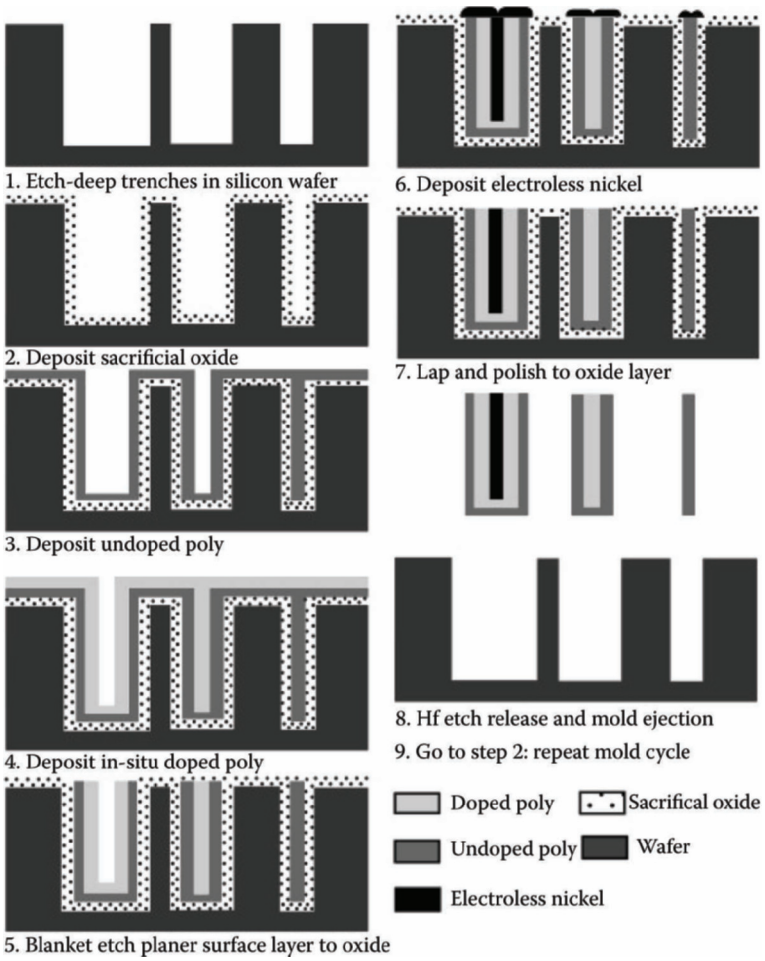


FIGURE 7.87 Schematic illustration of HEXSIL process. The mold wafer may be part of an infinite loop. (Courtesy of Dr. C. Keller, MEMS Precision Instruments.)

ratios of about 10 (e.g., a 10-μm-diameter hole with a depth of 100 μm). For trench etching, Keller uses a Cl_2 plasma etch with the following approximate etching conditions: flow rates of 200 sccm for He and 180 sccm for Cl_2, a working pressure of 425 mTorr, a power setting of 400 W, and an electrode gap of 0.8 cm. The etch rate for Si in this mode equals 1 μm/min. Thermal oxide and CVD oxide act as masks with 1 μm of oxide needed for each 20 μm of etch depth. Before the Cl_2 etch, a short 7-s SF_6 pre-etch removes any remaining native oxide in the mask openings. During the chlorine etch, a white sidewall passivating layer must be controlled to maintain perfect vertical sidewalls. After every 30 min of plasma etching, the wafers are submerged in a silicon isotropic etch long enough to remove the residue.[213] Beyond 100 μm, severe undercutting occurs, and the trench cross-section becomes sufficiently ellipsoidal to prevent molded parts from being pulled out. Advances in dry cryogenic etching are continually improving attainable etch depths, trench profiles, and minimum trench diameter. We can expect continuous improvements in the tolerances of this novel technique. After plasma etching, an additional 1 μm of silicon is removed by an isotropic wet etch to obtain a smoother trench wall surface. Alternatively, to smooth sidewalls and bottom of the trenches, a thermal wet oxide is grown and etched away. The sacrificial oxide in step 2 is made by CVD phosphosilicate glass (PSG) (at 450°C, 140 Å/min), CVD low-temperature oxide (LTO at 450°C), or CVD polysilicon (580°C, 65 Å/min). The latter is completely converted to SiO_2 by wet thermal oxidation at 1100°C. The PSG needs an additional reflowing and densifying anneal at 1000°C in nitrogen for 1 h. This results in an etching rate of the sacrificial layer of ~20 mm/min in 49% HF. The mold shown in Figure 7.87 displays three different trench widths and can be used to build integrated micromachines incorporating doped and undoped polysilicon parts, as well as metal parts. The remaining volume of the narrowest trench after oxide deposition is filled completely with the first deposition of undoped polysilicon (Poly 1) in step 3. The undoped poly will constitute the insulating regions in the micromachine. Undoped CVD polysilicon was formed in this case at 580°C, with a 100-sccm silane flow rate, and a 300-mTorr reactor pressure, resulting in a deposition rate of 0.39 μm/h. The deposited film under these conditions is amorphous or very fine grained. Generally speaking, CVD polysilicon films conform well to the underlying topography on the wafer and show good step coverage. With trenches of an aspect ratio in excess of 10, some thinning of the film on the sidewalls occurs. Because the narrowest trenches are completely filled in by the first deposition, they cannot accept material from later depositions. The trenches of intermediate width are lined with the first material and then completely filled in by the second deposition. In the case illustrated, the second deposition (step 4) consists of in situ doped polysilicon and forms the resistive region in the micromachine under construction. To prevent diffusion of *P* from the doped poly deposited on top of the narrow undoped beams, a blanket etch in step 5 is used to remove the doped surface layer before the anneal of the doped poly. The third deposition, in step 6 of the example case, consists of electroless nickel plating on polysilicon surfaces but not on oxide surfaces and results in the conducting parts of the micromachine. By depositing structural layers in order of increasing conductivity, as done here, regions of different conductivity can be separated by regions of narrow trenches containing only nonconducting material. Lapping and polishing in step 7 with a 1-μm diamond abrasive in oil planarizes the top surface, readying it for HF etch release and mold ejection in step 8. Annealing of the polysilicon is required to relieve the stress before removing the parts from the wafer so they remain straight and flat. In step 8, the sacrificial oxide is dissolved in 49% HF. A surfactant such as Triton X100 is added to the etch solution to facilitate part ejection by reducing surface adhesion between the part and the mold. The parts are removed from the wafer, and the wafer may be returned to step 2 for another mold cycle. An example micromachine, resulting from the described process, is the thermally actuated tweezers shown in Figure 7.88. These HEXSIL tweezers measure 4 mm long, 2 mm wide, and 80 μm tall. The thermal expansion beam to actuate the tweezers consists of the in situ doped polysilicon; the insulating parts are made from the undoped polysilicon material. Nickel-

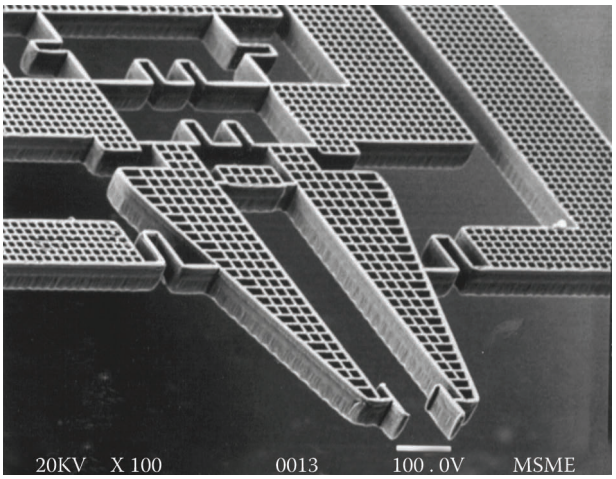


FIGURE 7.88 SEM micrograph of HEXSIL tweezers: 4 mm long, 2 mm wide, and 80 µm tall. Lead wires for current supply are made from nickel-filled polysilicon beams; in situ phosphorus-doped polysilicon provides the resistor part for actuation. The width of the beam is 8 mm: 2 mm polysilicon, 4 mm nickel, and 2 mm polysilicon. (Courtesy of Dr. C. Keller, MEMS Precision Instruments.)

filled polysilicon beams are used for the current supply leads. It is possible to combine the HEXSIL process with classical polysilicon micromachining, as illustrated in Figure 7.89a, where HEXSIL forms a stiffening rib for a membrane filter fashioned by surface micromachining of a surface polysilicon layer. The surface polysilicon is deposited after HEXSIL. A critical need in HEXSIL technology is controlled mold ejection. Keller et al.[213] have experimented with HEXSIL-produced bimorphs, making the structure spring up after release. Keller at MEMS Precision Instruments is exploiting HEXSIL technology commercially (http://www.memspi.com). Keller did his early HEXSIL work at the University of California, Berkeley and later incorporated the Bosch deep RIE process (see Chapter 3) to create tweezers structures similar to the one shown in Figure 7.88, which can now be fabricated in single-crystal silicon (see Figure 7.89b). MEMS Precision Instruments today makes microtweezers with one of three processes: HEXSIL, deep RIE Bosch process, or Sandia's SUMMiT-V process.

Thick Polysilicon

Applying classical LPCVD to obtain polysilicon deposition is a slow process. For example, a layer of 10 µm typically requires a deposition time of 10 h.

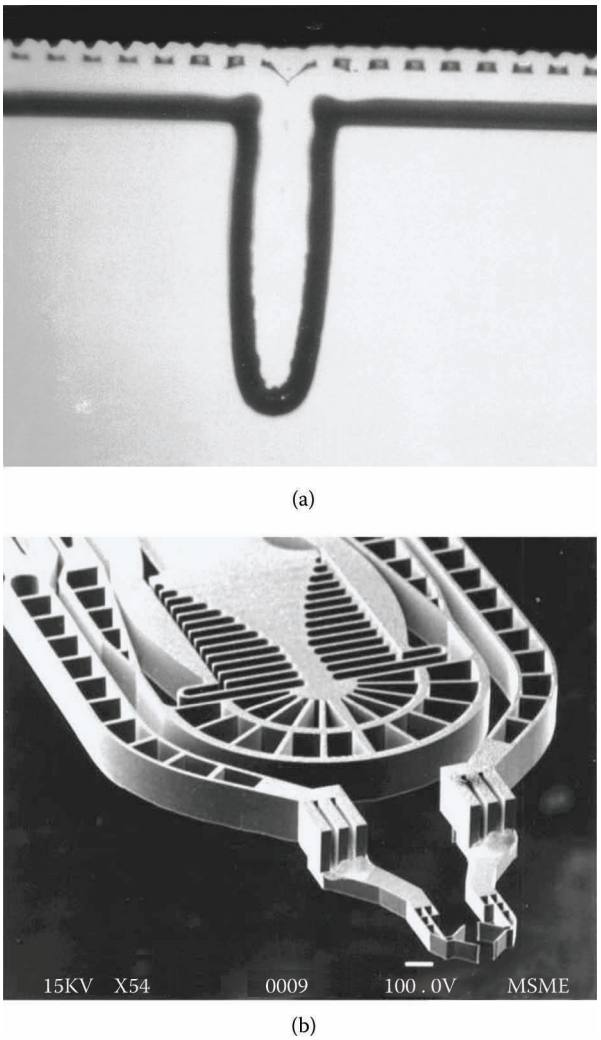


FIGURE 7.89 (a) SEM micrograph of surface-micromachined membrane filter with a stiffening rib (50 mm high). Original magnification 1000×. (b) Single-crystal microtweezers made by the Bosch deep RIE process. (Courtesy of Dr. C. Keller, MEMS Precision Instruments.)

Consequently, most micromachined structures are based on layer thicknesses in the 2–5-µm range. Basing their process on dichlorosilane (SiH_2Cl_2) chemistry, Lange et al.[98] developed a CVD process in a vertical epitaxy batch reactor with deposition rates as high as 0.55 µm/min at 1000°C. The process yields acceptable deposition times for thicknesses in the 10-µm range (20 min). The highly columnar polysilicon films are deposited on sacrificial SiO_2 layers and exhibit low internal tensile stress, making them suitable for surface micromachining. The surface roughness comprises about 3% of the thickness, which might preclude some applications.

Kahn et al.[214] made mechanical property test structures from thick undoped and in situ B-doped

polysilicon films. The elastic modulus of the B-doped polysilicon films was determined as 150 ± 30 GPa. The residual stress of as-deposited undoped thick polysilicon was determined as 200 ± 10 MPa.

This "thick epi-polysilicon surface micromachining" has been used to fabricate accelerometers with a seismic mass thickness well beyond what regular micromachining can achieve (http://www.sensorscan.com/sensoren/sensoren/accelerometer_1.htm).

Resists as Structural Elements and Molds in Surface Micromachining

Introduction Deep UV photoresists were covered in Chapter 1. In this section, we briefly reiterate some of the material covered there in the context of surface micromachining. Deep UV photoresists enable the molding of high-aspect-ratio microstructures in a wide variety of moldable materials, or they are used directly as structural elements in the case of permanent photoresists.

UV Depth Lithography

Polyimide Surface Structures Because of their transparency to exposing UV light, polyimide surface structures can be made very high and exhibit LIGA-like high-aspect ratios. By using multiple coats of spun-on polyimide, thick suspended plates are possible. Moreover, composite polyimide plates can be made, depositing and patterning a metal film between polyimide coats. Polyimide surface microstructures are typically released from the substrate by selectively etching an aluminum sacrificial layer, although Cu and PSG (e.g., in the polyimide flapping wing in Figure 7.86) have been used as well (see also Table 7.23).

Polyimides, because they are easily deformed, usually do not qualify as mechanical members but have been used, for example, in plastically deformable hinges (see Figure 7.86).[144,211] An early result in polyimide surface micromachining was obtained at SRI International, where polyimide pillars (spacers) about 100 µm in height were used to separate a Si wafer, which was equipped with a field emitter array, from the display glass plate of a field emission flat panel display.[215] The concept of a field emission display is illustrated in Figure 7.90a. The flat panel display and an SEM picture of the pillars are shown in Figure 7.90b. The field-emitter tip arrays are positioned between these supporting polyimide posts. The Probimide 348 FC formulation of Ciba-Geigy was used. This viscous precursor formulation (48% by weight of a polyamic ester, a surfactant for wetting, and a sensitizer) with a 3500 cs viscosity was applied to the Si substrate and formed into a film of a 125-µm thickness by spinning. A 30- to 40-min prebake at about 100°C removed the organic solvents from the precursor. The mask with the pillar pattern was then aligned to the wafer coated with the precursor and subjected to about 20 min of UV radiation. After driving off moisture by another baking operation, the coating, still warm, was spray developed (QZ 3301 from Ciba-Geigy), revealing the desired spacer matrix. By baking the polyimide at 100°C in a high vacuum (10^{-9} Torr), the pillars shrunk to about 100 µm, whereas the polyimide became more dense and exhibited greater structural integrity.

Frazier et al.[216] obtained a height-to-width aspect ratio of about 7 with polyimide structures. Ultraviolet was used to produce structures with heights in the range of 30–50 µm. At greater heights, the verticality of the sidewalls became relatively poor. Spun-on thickness in excess of 60 µm in a single coat was obtained for both Ciba-Geigy and Dupont commercial UV-exposable, negative-tone polyimides. Using a G-line mask aligner, exposure energy of 350 mJ/cm^2 was sufficient to develop a pattern with the Ciba-Geigy QZ 3301 developer. Allen and his team combined polyimide insert molds with electrodeposition to make a wide variety of metal structures.[217]

Other Resists Used for UV Deep Lithography Surface Micromachining Besides polyimides, research on Novolak-type resists also is leading to higher 3D features. Lochel et al.[206] use Novolak, positive tone resists of high viscosity (e.g., AZ 4000 series, Hoechst). They deposit, in multiple-coating process, layers up to 200 µm thick in a specially designed spin coater that incorporates a corotating cover. The subsequent deep UV lithography yields patterns with aspect ratios up to 10, steep edges (more than 88°), and a minimum feature size down to 3 µm. By combining this resist technology with sacrificial

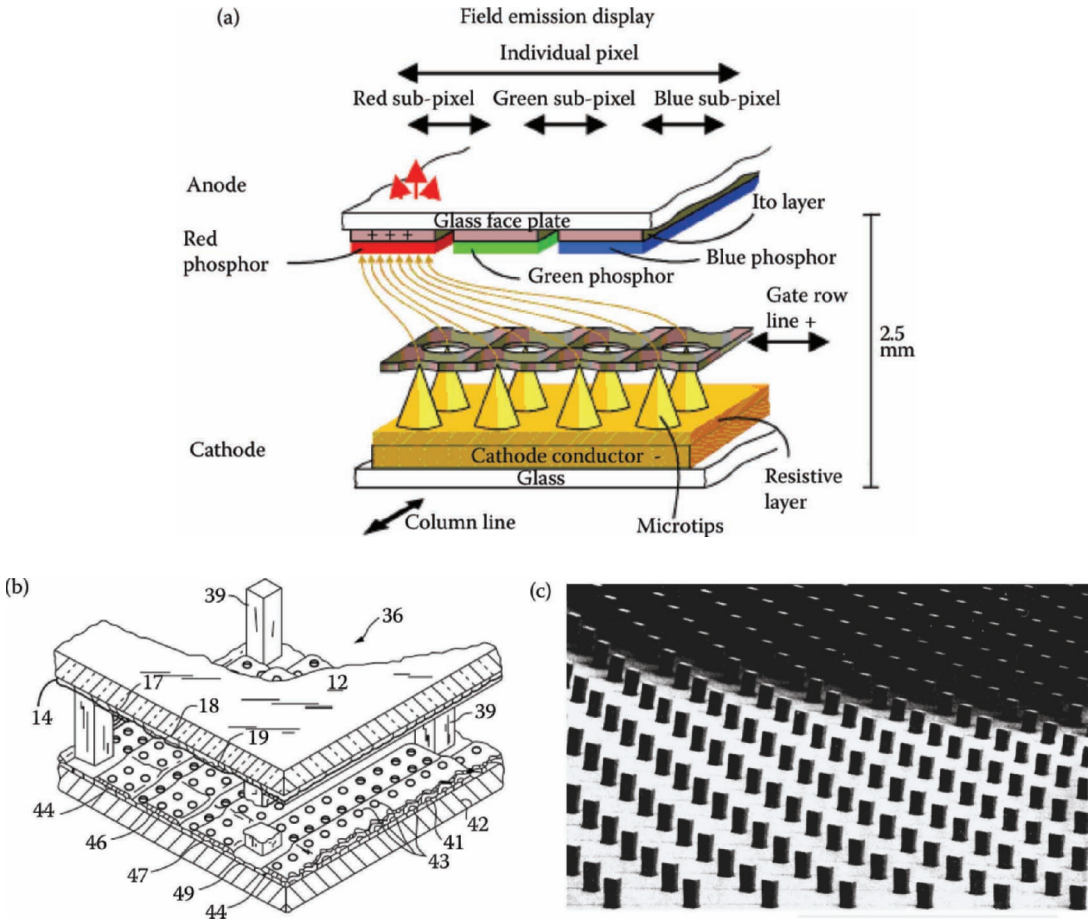


FIGURE 7.90 Polyimide structural elements. (a) Principle of a field emission display. (b) Micro-machined flat panel display. Number 39 represents one of the spacer pillars in the matrix of polyimide pillars (100 mm high). The spacer array separates the emitter plate from the front display plate. (c) SEM of the spacer matrix. (Courtesy of Dr. I. Brodie.) The height of the pillars is similar to what can be accomplished with LIGA. Because only a simple UV exposure was used, this polyimide is referred to as poor man's LIGA process, or pseudo-LIGA. (From I. Brodie et al., US Patent 4,923,421, 1990; and Brodie, I., H. R. Gurnick, C. E. Holland, and H. A. Moessner. 1990. *Method for providing polyimide spacers in a field emission panel display.*[215])

layers and electroplating, a wide variety of 3D microstructures result.

Along the same line, researchers at IBM experimented with Epon SU-8 (Shell Chemical), an epoxy-based, onium-sensitized, UV transparent negative photoresist used to produce high-aspect-ratio (>10:1) features and straight sidewalled images in thick film (>200 μm) using standard lithography.[218,219] SU-8 imaged films were used as stencils to plate Permalloy for magnetic motors.[218]

Patterns generated with these thick resist technologies should be compared with LIGA-generated patterns, not only in terms of aspect ratio but also in terms of sidewall roughness and sidewall run-out. Such a comparison will determine which surface machining technique to use for the job.

Porous Polysilicon

In Chapter 4, we discussed the transformation of single-crystal Si into a porous material with porosity and pore sizes determined by the current density, type, and concentration of the dopant, and the hydrofluoric acid concentration. A transition from pore formation to electropolishing is reached by increasing the current density and/or by decreasing the hydrofluoric acid concentration.[101,102] Porous silicon can also be formed under similar conditions from LPCVD polysilicon.[103] In this case, pores roughly follow the grain boundaries of the polysilicon. Figure 7.91 illustrates the masking and process sequence to prepare a wafer to make thin layers of porous Si between two insulating layers of low

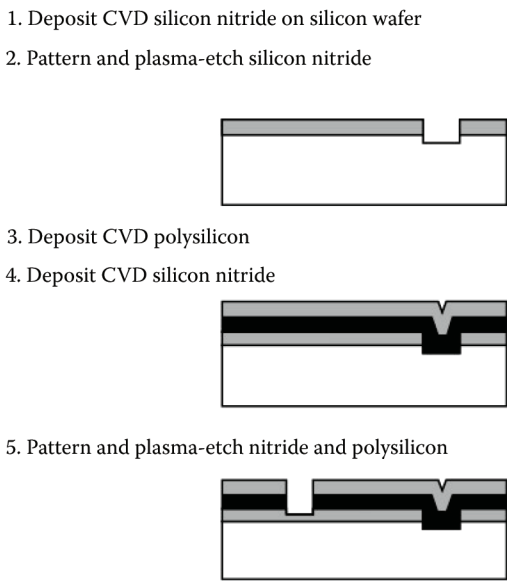


FIGURE 7.91 Masking and process sequence preparing a wafer for laterally grown porous polysilicon. (From Field, L. 1987. EECS/ERL Res. summary. University of California, Berkeley. With permission.[220])

stress silicon nitride.[220] The wafer, after the process steps outlined in Figure 7.91, is put in a Teflon test fixture, protecting the back from HF attack. An electrical contact is established on the back of the wafer, and a potential is applied with respect to a platinum-wire counter electrode immersed in the same HF solution. Electrolytes consisting of 5–49 wt% HF and current densities from 0.1–50 A cm^{-2} are used. The advance of a pore-etching front, growing parallel to the wafer surface, may be monitored using a line-width measurement tool. The highest observed rate of porous-silicon formation is 15 µm min^{-1} (in 25 wt% HF). In the electropolishing regime, at the highest currents, the etch rate is diffusion limited, but the reaction is controlled by surface reaction kinetics in the porous Si growth regime.

By changing the conditions from pore formation to electropolishing and back to porous Si, an enclosed chamber may be formed with porous polysilicon walls (plugs) and "floor and ceiling" silicon nitride layers. Sealing of the cavities by clogging the microporous polysilicon was attempted by room temperature oxidation in air and in an H_2O_2 solution. Leakage through the porous plug persisted after those room temperature oxidation treatments, but with a Ag deposition from a 400-mM $AgNO_3$ solution and subsequent atmospheric tarnishing (48 h, Ag_2S) the chambers appeared to be sealed, as determined from the lack of penetration by methanol. This technology might open up possibilities for filling cavities with liquids and gases under low-temperature conditions. The chamber provided with a porous plug might also make a suitable on-chip electrochemical reference electrode with an electrolyte reservoir.

Hydrofluoric acid can penetrate thin layers of polysilicon either at foreign particle inclusion sites or at other critical film defects, such as grain boundaries (see above). This way, the HF can etch underlying oxide layers, creating, for example, circular regions of freestanding polysilicon, so-called *blisters*. The polysilicon permeability associated with blistering of polysilicon films has been applied successfully by Judy et al.[221,222] to produce thin-shelled hollow beam electrostatic resonators from thin polysilicon films deposited onto PSG. The possible advantage of using these hollowed structures is to obtain a higher resonator quality factor *Q*. The devices were made in such a way that the 0.3-µm-thick undoped polysilicon completely encased a PSG core. After annealing, the structures were placed in HF, which penetrated the polysilicon shell and dissolved away the PSG, eliminating the need for etch windows. It was not possible to discern the pathways through the polysilicon using TEM.

Lebouitz et al.[223] applied permeable polysilicon etch-access windows to increase the speed of creating microshells for packaging surface-micromachined components. After etching the PSG through the many permeable Si windows, the shell is sealed with 0.8 µm of low-stress LPCVD nitride.

Porous Si can be used very effectively as a sacrificial layer both in polysilicon surface micromachining and in SOI micromachining (see Table 7.23).[139] The high surface area of the porous Si results in rapid etching in KOH at room temperature. The porous layer can be formed directly under a deposited layer or under the epilayer in SOI surface micromachining.

Silicon is not the only material that can be converted into a porous sponge-like material. Most semiconductor materials can be modified this way, and in Volume III, Chapter 5 on selected MEMS and NEMS materials we discuss micromachining of

high-aspect-ratio structures in porous anodic aluminum oxide. The latter enables a class of MEMS applications such as high-temperature gas microsensors, vacuum microelectronics, RF-MEMS, nozzles, filters, membranes, and so on.[224]

Comparison of Bulk Micromachining with Surface Micromachining

Surface and bulk micromachining have many processes in common. Both techniques rely heavily on photolithography; oxidation; diffusion and ion implantation; LPCVD and PECVD for oxide, nitride, and oxynitride; plasma etching; use of polysilicon; and metallizations with sputtered, evaporated, and plated Al, Au, Ti, Pt, Cr, and Ni. Where the techniques differ is in the use of anisotropic etchants, anodic and fusion bonding, (100) versus (110) starting material, p^+ etch stops, double-sided processing and electrochemical etching in bulk micromachining, and the use of dry etching in patterning and isotropic etchants in release steps for surface micromachines. Combinations of substrate and surface micromachining also frequently appear. The use of polysilicon avoids many challenging processing difficulties associated with bulk micromachining and offers new degrees of freedom for the design of integrated sensors and actuators. Design freedom includes many more possible shapes in the *x,y* plane and the ease of integration of several sensors on one die (e.g., a two-axis accelerometer). The technology combined with sacrificial layers also allows the nearly indispensable further advantage of in situ assembly of the tiny mechanical structures because the structures are preassembled as a consequence of the fabrication sequence. Another advantage focuses on thermal and electrical isolation of polysilicon elements. Polycrystalline piezoresistors can be deposited and patterned on membranes of other materials, for example, on a SiO_2 dielectric. This configuration is particularly useful for high-temperature applications. The p-n junctions act as the only electrical insulation in the single-crystal sensors, resulting in high leakage currents at high temperatures, whereas current leakage for the polysilicon/SiO_2 structure virtually does not exist. The limits of surface micromachining are striking. CVD silicon usually caps at layers no thicker than 1–2 μm because of residual stress in the films and the slow deposition process (thick polysilicon needs further investigation).* A combination of a large variety of layers may produce complicated structures, but each layer is still limited in thickness. Also, the wet chemistry needed to remove the interleaved layers may require many hours of etching (except when using the porous Si option discussed above), and even then stiction often results.

The structures made from polycrystal silicon exhibit inferior electronic and slightly inferior mechanical properties compared with single-crystal silicon. For example, polysilicon has a lower piezoresistive coefficient (resulting in a gauge factor of 30 vs. 90 for single-crystal Si), and it has a somewhat lower mechanical fracture strength. Polysilicon also warps because of the difference of thermal expansion coefficient between polysilicon and single-crystal silicon. Its mechanical properties strongly depend on processing procedures and parameters.

Table 7.25 introduces a comparison of surface micromachining with wet bulk micromachining. The status depicted reflects the mid-1990s and only includes polysilicon surface micromachining. As discussed, SOI micromachining, thick polysilicon, hinged polysilicon, polyimide, and millimeter-molded polysilicon structures have dramatically expanded the application bandwidth of surface micromachining. In Chapter 10 on replication techniques-LIGA, we will see how x-ray lithography further expands the *z*-direction for new surface-micromachined devices with unprecedented aspect ratios and extremely low surface roughness. In Table 7.26 we compare physical properties of single-crystal silicon with those of polysilicon.

Although polysilicon can be an excellent mechanical material, it remains a poor electronic material. Reproducible mechanical characteristics are difficult and complex to consistently realize. Fortunately, SOI surface micromachining and other newly emerging surface micromachining techniques can alleviate many of the problems.[229]

* As a direct consequence, the inertial mass in a surface-micromachined accelerometer, such as the ADXL05 from Analog Devices, is only 0.3 μg, and the corresponding noise, dominated by Brownian noise, is 500 μg/√Hz. By contrast, the mass for a bulk-micromachined accelerometer can easily be made 100 μg.

Table 7.25 Comparison of Bulk Micromachining with Surface Micromachining

Bulk Micromachining	Surface Micromachining
Large features with substantial mass and thickness	Small features with low thickness and mass
Uses both sides of the wafer	Multiple deposition and etching required to build up structures
Vertical dimensions: one or more wafer thicknesses	Vertical dimensions are limited to the thickness of the deposited layers (~2 μm) leading to compliant suspended structures with the tendency to stick to the support
Generally involves laminating Si wafer to Si or glass	Surface micromachined device has its built-in support and is more cost effective
Piezoresistive or capacitive sensing	Capacitive and resonant sensing mechanisms
Wafers may be fragile near the end of the production	Cleanliness critical near end of process
Sawing, packaging, testing is difficult	Sawing, packaging, testing is difficult
Some mature products and producers	No mature products or producers
Not very compatible with IC technology	Natural but complicated integration with circuitry; integration is often required due to the tiny capacitive signals

Source: Adapted from Jerman, H. 1994. Hard copies of viewgraphs presented in Banff, Canada.[90]

Table 7.26 Comparison of Materials Properties of Si Single Crystal with Crystalline Polysilicon

Material Property	Single-Crystal Si (SCS)	Polysilicon
Thermal conductivity (W/cm°K)	1.57	Strong function of the grain structure of the film; 0.30–0.35 (for fine grains and double that for larger grains)
Thermal expansion (10^{-6}/°K)	2.33	2–2.8
Specific heat (cal/g°K)	0.169	0.169
Piezoresistive coefficients	n-Si (π_{11} = –102.2) p-Si (π_{44} = +138.1) e.g., gauge factor of 90	e.g., gauge factor of 30 (>50 with laser recrystallization)
Refractive index		4.1 at 600 nm
Mobility ($cm^2/V/s$)	Holes: 600 Electrons: 1500	Maximum for electrons: 30
Density (cm^{-3})	2.32	2.32
Fracture strength (GPa)	6	0.8–2.84 (Undoped polysilicon)
Dielectric constant	11.9	Sharp maxima of 4.2 and 3.4 eV at 295 and 365 nm, respectively
Residual stress	None	Depends on structure as deposited films are compressive
Temperature resistivity coefficient ($°C^{-1}$)TCR	0.0017 (p-type)	0.0012 nonlinear, + or – through selective doping, increases with decreasing doping level can be made 0!
Poisson ratio	0.262 max for (111)	0.23
Young's modulus (10^{11} N/m^2)	1.90 (111)	1.61
Resistivity at room temperature (ohm.cm)	Depends on doping	Strong function of the grain structure of the film. Plateaus at 4 10^{-4} above 1 10^{21} cm^{-3} P (always higher than for SCS)

Source: Based on Lin, L. 1993. *Selective encapsulations of MEMS: micro channels, needles, resonators, and electromechanical filters.* PhD thesis. University of California, Berkeley;[225] Adams, A. C. 1988. *VLSI technology.* Ed. S. M. Sze. New York: McGraw-Hill;[30] Kamins, T. 1988. *Polycrystalline silicon for integrated circuits.* Boston: Kluwer;[78] and Heuberger, A. 1989. *Mikromechanik.* Heidelberg: Springer Verlag.[227] (See also Volume I, Table 4.15.)

Examples

Example 7.9: Analog Devices Accelerometer

Both Robert Bosch GmbH (Stuttgart, Germany) and Analog Devices (Norwood, MA) offer surface-micromachined accelerometers based on lateral resonators. We will only review the Analog Devices ADXL accelerometer product family here. The ADXL-50 constituted the first commercially available surface-micromachined MEMS structures. Today, surface-micromachined accelerometers are incorporated in Ford and General Motors cars, among others, as well as inside joysticks for computer games, robots, watches, and shoes.

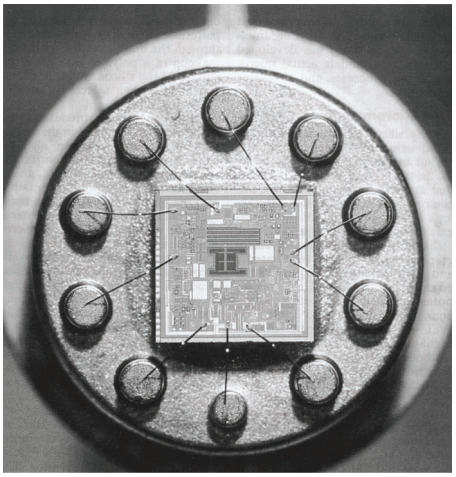


FIGURE 7.92 Analog Devices' ADXL-50 accelerometer with a surface-micromachined capacitive sensor (center), on-chip excitation, self-test, and signal-conditioning circuitry. (From Core, T. A., W. K. Tsang, and S. J. Sherman. 1993. Fabrication technology for an integrated surface-micromachined sensor. *Solid State Technol* 36:39–47. With permission.[119])

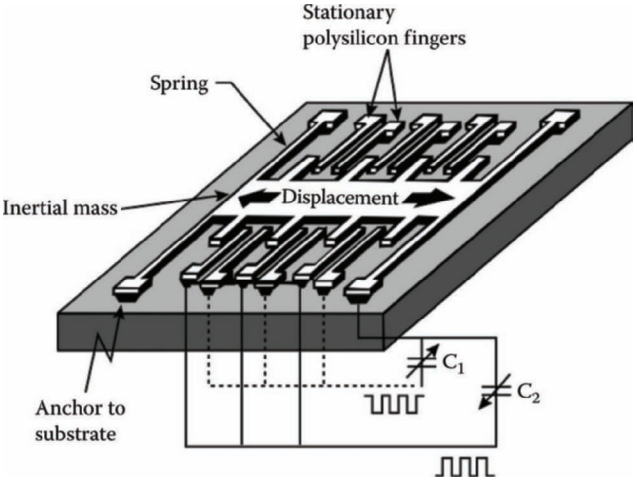


FIGURE 7.93 Illustration of the basic mechanical structure of Analog Devices' ADXL family of surface-micromachined accelerometers. A comb-like plate suspended from springs forms the inertial mass. Displacements of the mass are measured capacitively with respect to two sets of stationary finger-like electrodes. (Based on Maluf, N. 2000. *An introduction to microelectromechanical systems engineering.* Boston: Artech House.[3])

To facilitate integration of their surface-micromachined accelerometers with on-board electronics, Analog Devices opted for a mature 4-μm BiCMOS process.[119] BiCMOS is a manufacturing process for semiconductor devices that combines bipolar and CMOS to give the best balance between available output current and power consumption. Figure 7.92 presents a photograph of the finished accelerometer with on-chip excitation, self-test, and signal conditioning circuitry.

The suspended comb-like H-shaped structure in the center of the die is the sensitive element, and its primary axis of sensitivity lies in the plane of the die (*x-y* plane). In bulk micromachining, the sense axis is more often orthogonal to the plane of the die (*z*-axis). The polysilicon-sensing element of the ADLAX-50 only occupies 5% of the total die area and consists of three sets of 2-μm-thick polysilicon finger-like electrodes (Figure 7.93).

Two sets are anchored to the substrate, and a third set is suspended about 1 μm above the surface by means of two folded polysilicon beams acting as suspension springs. The fingers of the movable shuttle mass are interlaced with the fingers of the two fixed sets. The whole chip measures 500 × 625 μm and operates as an automotive airbag deployment sensor. The measurement accuracy is 5% over the ±50-g range. Deceleration in the axis of sensitivity exerts a force on the central mass that, in turn, displaces the interleaved capacitor plates, causing a fractional change in capacitance. The overall capacitance is small, typically on the order of 100 fF, and for the ADXL-05 (rated at ±5 g), with an inertial mass of 0.3 μg only, the capacitance change is as small as 100 aF.[197] These small capacitance changes necessitate on-chip integrated electronics to reduce the impact of parasitic sources. In operation, the ADXL family has a force-balance electronic control loop to prevent the mass from macroscopic movements, greatly improving output linearity because the center element never moves by more than a few nanometers. Applying a large-amplitude, low-frequency voltage, below the natural frequency of the sensor, allows one to compensate for accelerometer plate movement by external acceleration. At the same time, the sensing excitation frequency (1 MHz) is much higher than the resonant frequency, so that it does not produce an actuation force on the capacitor plates. As long as sense and actuation signals do not interfere, sense and actuation plates may be the same. Straight flexures were used for the layout of the lateral resonator shuttle mass (see also Figure 7.76a).

In the sensor design, n^+ underpasses connect the sensor area to the electronic circuitry, replacing the usual heat sensitive aluminum connect lines. Most of the sensor processing is

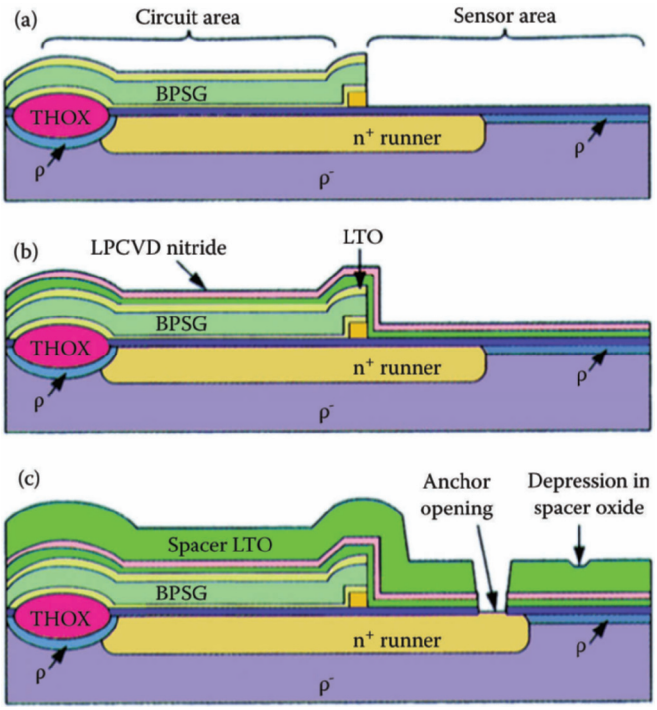


FIGURE 7.94 Preparation of IC chip for polysilicon. (a) Sensor area post-BPSG planarization and moat mask. (b) Blanket deposition of thin oxide and thin nitride layer. (c) Bumps and anchors made in LTO spacer layer. (From Core, T. A., W. K. Tsang, and S. J. Sherman. 1993. Fabrication technology for an integrated surface-micromachined sensor. *Solid State Technol* 36:39–47. With permission.[119])

inserted into the BiCMOS process right after the borophosphosilicate glass (BPSG) planarization. After planarization, a designated sensor region or moat is cleared in the center of the die (Figure 7.94a). A thin oxide is then deposited to passivate the n+ underpass connects, followed by a thin, low-pressure, vapor-deposited nitride to act as an etch stop (buffer layer) for the final polysilicon release etch (Figure 7.94b). The spacer or sacrificial oxide used is a 1.6-μm densified low-temperature oxide (LTO) deposited over the whole die (Figure 7.94c). In a first timed etch, small depressions that will form bumps or dimples on the underside of the polysilicon sensor are created in the LTO layer. These will limit stiction in case the sensor comes in contact with the substrate. A subsequent etch cuts anchors into the spacer layer to provide regions of electrical and mechanical contact (Figure 7.94c). The 2-μm-thick sensor polysilicon is then deposited, implanted, annealed, and patterned (Figure 7.95a). The relatively deep junctions of the BiCMOS process permit the polysilicon thermal anneal and brief dielectric densifications without resulting in degradation of the electronic functions. Next is the IC metallization, which starts with the removal of the sacrificial spacer oxide from the circuit area along with the LPCVD nitride and LTO layer. A low-temperature oxide is deposited on the polysilicon sensor part, and contact openings appear in the IC part of the die where platinum is deposited to form a platinum silicide (Figure 7.95b). The trimmable thin-film material, TiW barrier metal, and Al/Cu interconnect metal are sputtered on and patterned in the IC area. The circuit area is then passivated in two separate deposition steps. First, plasma oxide is deposited and patterned (Figure 7.95c), followed by a plasma nitride (Figure 7.96a) to form a seal with the earlier deposited LPCVD nitride. The nitride acts as an HF barrier in the subsequent long etch release. The plasma oxide left on the sensor acts as an etch stop for the removal of the plasma nitride. Subsequently, the sensor area is prepared for the final release etch.

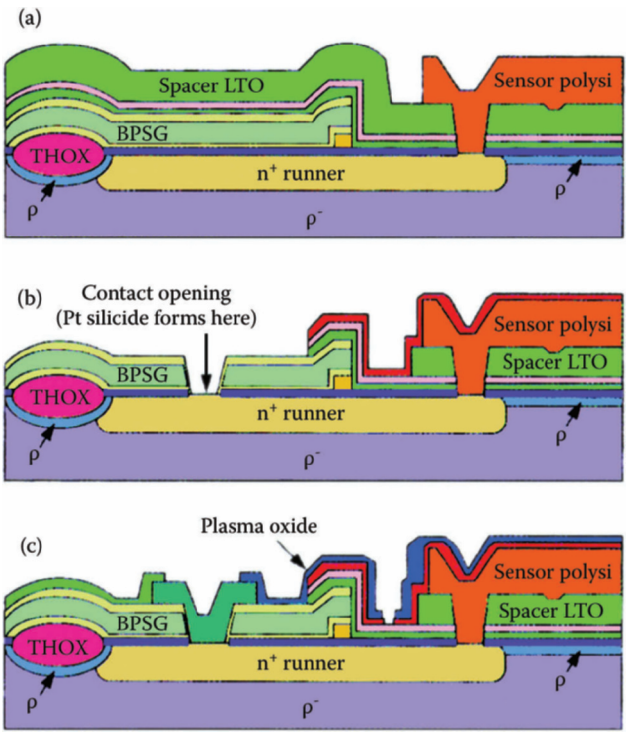


FIGURE 7.95 Polysilicon deposition and IC metallization. (a) Cross-sectional view after polysilicon deposition, implant, anneal, and patterning. (b) Sensor area after removal of dielectrics from circuit area, contact mask, and Pt silicide. (c) Metallization scheme and plasma oxide passivation and patterning. (From Core, T. A., W. K. Tsang, and S. J. Sherman. 1993. Fabrication technology for an integrated surface-micromachined sensor. *Solid State Technol* 36:39–47. With permission.[119])

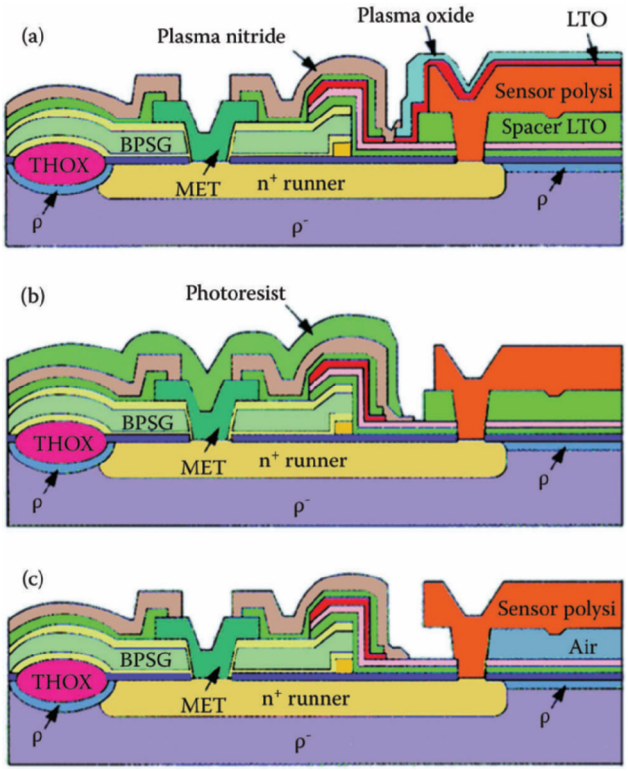


FIGURE 7.96 Prerelease preparation and release. (a) Postplasma nitride passivation and patterning. (b) Photoresist protection of the IC. (c) Freestanding, released polysilicon beam. (From Core, T. A., W. K. Tsang, and S. J. Sherman. 1993. Fabrication technology for an integrated surface-micromachined sensor. *Solid State Technol* 36:39–47. With permission.[119])

The undensified dielectrics are removed from the sensor, and the final protective resist mask is applied. The photoresist protects the circuit area from the long-term buffered oxide etch (Figure 7.96b). The final device cross-section is shown in Figure 7.96c.

Example 7.10: TI Micromirrors

In 1987 the first digital micromirror device (DMD™) was developed at TI by Dr. Larry Hornbeck [US Patent 4,615,595 (October 7, 1986)]. A typical DMD consists of a 2D array of optical switching elements (pixels) on a silicon substrate, as shown in Figure 7.97.

Two pixels are schematically illustrated in Figure 7.98 together with the underlying Si chip and circuitry. Each pixel is made up of a reflective aluminum micromirror supported from a central post. The central mirror post at the back of the mirror is mounted on a lower aluminum metal platform—the yoke. The yoke is suspended

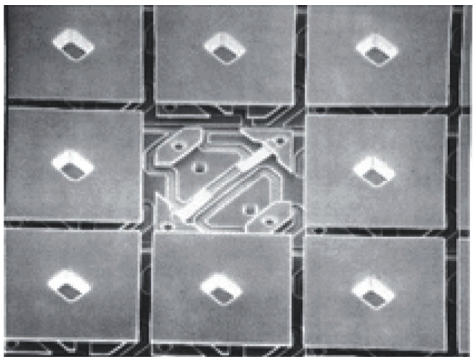


FIGURE 7.97 Texas Instruments DMD pixel array. One pixel has been removed to show the silicon chip below the reflective aluminum mirrors.

above the silicon substrate by thin compliant L-shaped hinges (made from a proprietary Al alloy) anchored to the underlying substrate by two stationary posts. The different components of an individual micromirror are illustrated in Figure 7.99. Two bias electrodes tilt the mirror either +10° or –10° by applying 24 V between one electrode or the other and the yoke. Off-axis illumination of the Al mirror reflects into the projection lens only when the micromirror is in its +10° state, producing a bright appearance or ON state. In the flat position and in the –10° state the pixel appears dark. In the fully deflected position, the yoke touches a landing

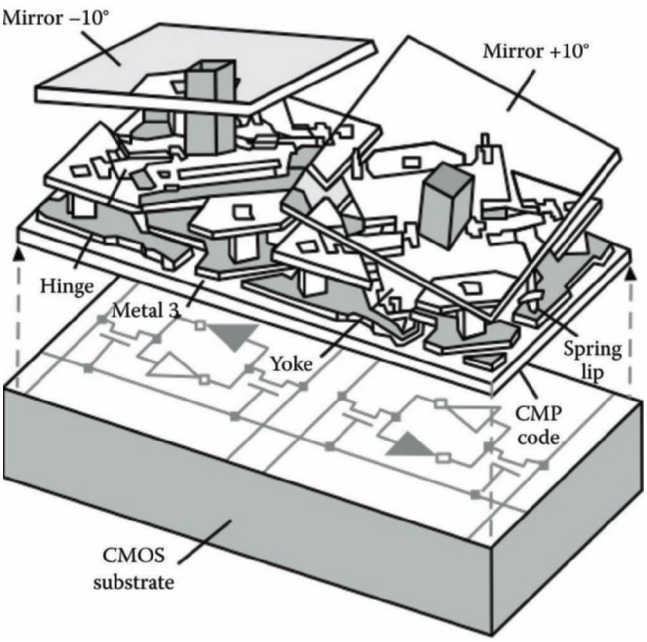


FIGURE 7.98 Schematic of two pixels in a Texas Instruments of a DMD. The mirrors are made transparent for clarity of the drawing. The +10° mirror is in the ON position, and the –10° is in the OFF position.

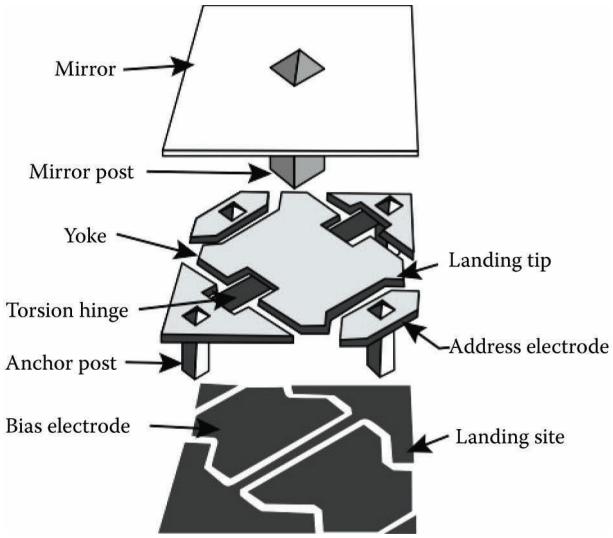


FIGURE 7.99 Illustration of the various components of a single DMD pixel. The basic structure consists of a bottom aluminum layer containing two electrodes, a middle aluminum layer containing a yoke suspended by two torsional hinges, and a top reflective aluminum mirror. An applied electrostatic voltage on a bias electrode deflects the yoke and the mirror toward that electrode. A pixel measures approximately 17 μm on a side. (Adapted from Maluf, N. 2000. *An introduction to microelectromechanical systems engineering.* Boston: Artech House.[3])

site with its landing tips, and because the landing site is biased at the same voltage, electrical shorting is prevented. Once the applied voltage is removed, the springy aluminum alloy hinges restore the micromirror to its initial position. A standard DMD microchip contains more than 442,000 switchable mirrors on a surface 5/8 in. wide. Mirrors are switched according to memory impulses stored in static-random-access memory (SRAM) cells beneath the tiny array, with mirrors tilting plus or minus 10° to reflect light into or away from an imaging lens. The mirrors can be independently cycled at 100,000 flips/s. Grays can be achieved by multiple mirrors or by flipping the mirror quickly between black and white states. Colors can be generated by having the incident light shine through a rotating disk divided into three colored segments: red, blue, and green. A separate image is generated for each color, timed to appear as the appropriate segment covers the light source. Printing and display technology can now enjoy the advantage of digital fidelity and digital stability.

In Figure 7.99, one single pixel is illustrated with its various components.[197]

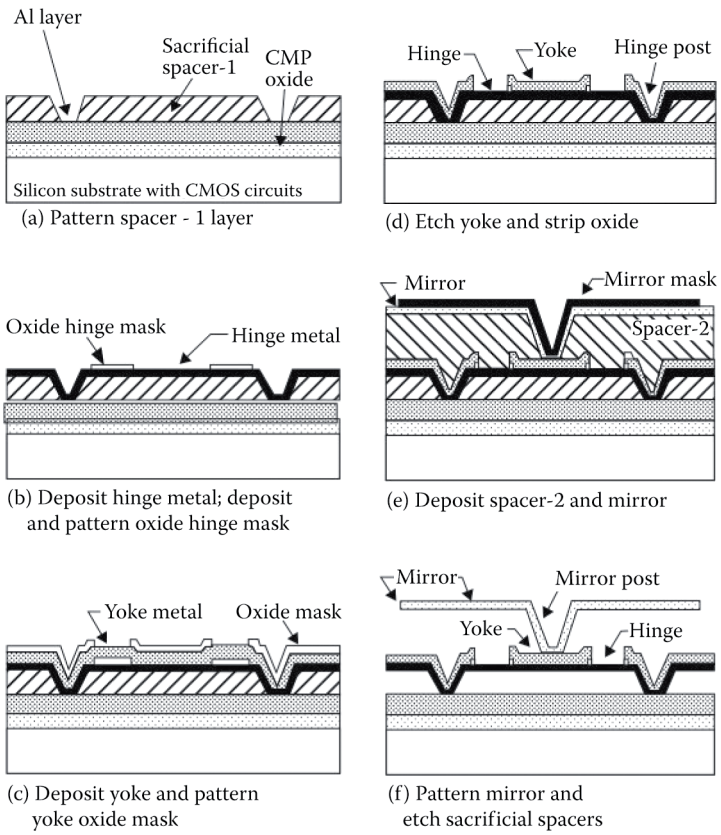


FIGURE 7.100 Fabrication steps of the Texas Instruments DMD. Steps (a) through (f) are explained in the text. (Based on Maluf, N. 2000. *An introduction to microelectromechanical systems engineering.* Boston: Artech House.[3])

The surface micromachining process to fabricate DMDs on wafers incorporating CMOS electronic address and control circuitry is illustrated in Figure 7.100. Because of the underlying active circuitry and the presence of Al metal for connectors and MEMS structures, all micromachining process steps are carried out at temperatures less than 400°C. After completing the CMOS circuitry, a thick oxide is deposited over the whole Si wafer and is chemomechanically polished to provide a flat surface to start building the mirror array. A sputter-deposited Al layer is patterned to provide bias and address electrodes, landing pads, and electrical interconnects to the underlying electronics. Hardened photoresist is used as the sacrificial material (Figure 7.100a). A proprietary Al alloy is sputter deposited to form the hinges for the mirror. It is the nature of this aluminum alloy that secures the mechanical integrity of the mirror actuation. Subsequently, the torsion hinge regions are protected by a patterned thin PECVD-deposited silicon dioxide (Figure 7.100b). In the next step, a thicker coat of another proprietary aluminum alloy is deposited to form the yoke structure; this new coat of Al buries the thin oxide hinge mask. A second PECVD oxide mask is deposited over this second level of Al metal and patterned in the shape of the yoke and anchor posts (Figure 7.100c). In a dry etch step, the exposed aluminum areas are removed down to the organic sacrificial resist except where the oxide hinge mask remains. In those regions, only the thick yoke metal is removed, stopping on the SiO_2 mask and so preserving the underlying hinge structure (Figure 7.100d). Both thin layers of PECVD oxide mask are stripped before a second layer of sacrificial resists is deposited, UV-hardened, and patterned. A third aluminum alloy is sputter deposited and defines the mirror and the central mirror post. Again a thin layer of PECVD silicon dioxide is deposited and patterned to define the mirror (Figure 7.100e). An oxygen plasma etch removes both sacrificial layers and releases the micromirrors (Figure 7.100f). Finally after release a special passivation step deposits a thin antistiction layer to prevent adhesion between the yoke tips and the landing pads. Because the weight of the micromirrors is insignificant, the DMD micromirrors can withstand 1500-G mechanical shocks. Optimization of the hinge metal alloy and fabrication processes have resulted in a mean time between failure of more than 100,000 h. Invented by Larry Hornbeck at TI, DMD is the key component in more than 17 projector brands and brought the first digital cinema to *Star Wars* and *Toy Story II*. A competing technology, the actuated mirror array (AMA; now Daewoo's TMA) invented by Gregory Um, is also a MEMS technology. Unlike DMD, TMA is built with piezo materials, and for high resolution its large array size and consequent cost still pose "sticky issues," but it achieves brightness 15% higher than any other projector. DMDs are also used in place of masks to map the human genome. At the University of Wisconsin, arrays for "gene expression analysis," usually produced with lithographic masks like displays, are being replaced with the same DMD used for projection display to make "virtual masks" containing nearly half a million features.

For his amazing piece of engineering, TI's Larry J. Hornbeck received the first Emmy Award ever bestowed for a projection display technology.

Questions

7.1: What is the mean free path (MFP)? How can you increase the MFP in a vacuum chamber? For metal deposition in an evaporation system, compare the distance between target and evaporation source with the working MFP. Which one has the smaller dimension? 1 atmosphere pressure = … mm Hg = … torr. What are the physical dimensions of impingement rate?

7.2: If we want to deposit a metal film on a substrate by resistance heated evaporation (as opposed to E-beam), what kind of metals are preferred? How is the thickness of a deposited thin film measured during evaporation?

7.3: Why is sputter deposition so much slower than evaporation deposition? Make a detailed comparison of the two deposition methods.

7.4: Develop the principal equation for the material flux to a substrate in a CVD process, and indicate how one moves from a mass transport limited to reaction-rate limited regime.

Explain why, in one case, wafers can be stacked close and vertically, while in the other a horizontal stacking is preferred.

7.5: Describe step coverage in CVD processes. Explain how gas pressure and surface temperature may influence the different coverage profiles.

7.6: Compare sputter deposition with evaporation for a simple metal such as Ag. Compare the two techniques for as many different parameters as you remember. Give examples where you would use one technique over the other.

7.7: CVD: (a) Show different types of step coverage and explain the most important parameters influencing each. (b) Describe the difference between PECVD and LPCVD.

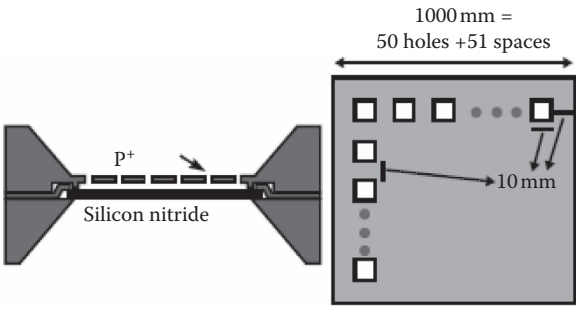
7.8: What are the characteristics of electron beam heated evaporation compared to resistance heated evaporation? Mark correct answers with an X.

( ) It is more complex.
( ) It is very versatile.
( ) It only works under lower temperatures.
( ) Everything a resistance heated evaporator can deposit can also be accommodated by electron beam heated evaporation.
( ) A magnetic field can be used to increase the temperature.
( ) It is not possible to provide a large evaporant surface area in electron beam heated evaporation.
( ) The adhesion between an evaporant and a substrate is accomplished by local reactions (sticking). Titanium and chromium are often used as a "glue" to improve the adhesion between an evaporant and a substrate.
( ) When the partial pressure of an evaporant vapor exceeds its equilibrium vapor pressure, it will condense.
( ) When a system is under equilibrium vapor pressure, the net transfer rate of material from one state to the other state is equal to 1.
( ) The cosine law is the underlying principle that a Kundsen cell can deposit a perfectly uniform coating inside a spherical glass jar.

7.9: Design at least three micromachines enabling the testing of thin film mechanical properties.

7.10: Detail several processes to create vacuum shells in a solid.

7.11: Consider a MEMS-based condenser acoustic transducer as shown in the figure below. The diaphragm is made of silicon nitride, and the back plate is perforated. A stress gauge was used to measure the stress in the silicon nitride and was found to be $1.5 \times 10^8 N/m^2$.



- Thickness of air gap = 5 mm
- Thickness of diaphragm = 5 mm
- Density of $Si_3N_4 = 3 \times 10^3 kg/m^3$

Total of 2500 square holes

(a) Write the differential equation governing the dynamic behavior of the movable diaphragm.
(b) Solve the differential equation under the 3 cases of overdamped, underdamped, and critically damped conditions.
(c) Calculate the mass of the diaphragm, the small signal capacitance, and the air-streaming resistance as per the layout shown if the device is to be operated in air and water.
(d) Calculate the cut-off frequency for the overdamped case in air and in water.

7.12: Discuss the advantages/disadvantages of using surface vs. bulk micromachining.

7.13: Define the following terms in 3–4 lines with examples where applicable:
LPCVD
PECVD
DRIE
Difference between RIE and plasma etching
LIGA process
SCREAM process

Silicon fusion bonding
Dissolved wafer process

7.14: Describe 5 methods you could implement to prevent stiction between surface micromachined components.

7.15: Design a comb drive for the MUMPs process.
(a) Show the mask layouts (top view) of the relevant layers.
(b) Calculate the maximum force, maximum displacement, and resonance frequency that you expect to achieve with this design.

Make sure the following constraints are satisfied:
- The total number of comb fingers should be less than 80.
- The device fits into a $200 \times 200\ \mu m^2$ square area.
- The maximum applied voltage is 30V.
- The device satisfies all design rules for MUMPs processing.

7.16: For pressure sensing diaphragms of the order of 100 µm square, why is it impractical to use capacitance to measure deflection?

7.17: List five methods you could use to release stress in a poly-Si comb resonator and describe the relative merits of each approach.

7.18: Take a look at the Sandia website at http://mems.sandia.gov/scripts/images.asp. Many of the surface micromachined gears have tiny holes in their surfaces. Why are these there?

References

1. Brodie, I., and J. J. Muray. 1982. *The physics of microfabrication.* New York: Plenum Press.
2. Murarka, S. P. 1988. *VLSI technology.* Ed. S. M. Sze. New York: McGraw-Hill.
3. Maluf, N. 2000. *An introduction to microelectromechanical systems engineering.* Boston: Artech House.
4. ERL. 1985. *SAMPLE version 1.6a user's guide.* Berkeley, CA: Electronics Research Laboratory (ERL), University of California, Berkeley.
5. Fichtner, W. 1988. *VLSI technology.* Ed. S. M. Sze. New York: McGraw-Hill.
6. Vossen, J. L., and W. Kern, eds. 1978. *Thin film processes.* Orlando, FL: Academic Press.
7. Maissel, L. I., and R. Glang, eds. 1970. *Handbook of thin film technology.* New York: McGraw-Hill.
8. Kinoshita, K., and M. J. Madou. 1984. Electrochemical measurement on Pt, Ir, and Ti oxides as pH probes. *J Electrochem Soc* 131:1089–94.
9. Studt, T. 1992. Better living through sputtering. *Res Dev* 54.
10. Comello, V. 1992. Tough coatings are a cinch with new PVD method. *Res Dev* January:59–60.
11. Editorial. 1994. *Brochure B-9870.* Zulte, Belgium: Innovative Sputtering Technology (IST).
12. Honer, K. A., and G. T. A. Kovacs. 2000. *Solid-state sensor and actuator workshop.* Hilton Head Island, SC.
13. Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure.* Weinheim, Germany: VCH Publishers.
14. Dutta, B., X. D. Wu, A. Inam, and T. Venkatesan. 1989. Pulsed laser deposition: a viable process for superconducting thin films? *Solid State Technol* February:106–10.
15. Inam, A. 1991. Pulsed laser takes the heat off HTS materials deposition. *Res Dev* February:90–92.
16. Cotell, C. M. 1993. Pulsed laser deposition and processing of biocompatible hydroxylapatite thin films. *App Surf Sci* 69:140–48.
17. Guo, T., P. Nikolaev, A. Thess, D. T. Colbert, and R. E. Smalley. 1995. Catalytic growth of single-walled nanotubes by laser vaporization. *Chem Phys Lett* 243:49–54.
18. Cotell, C. M., and K. S. Grabowski. 1992. Novel materials applications of pulsed laser deposition. *MRS Bull* 17:44–53.
19. Cotell, C. M., D. B. Chrisey, K. S. Grabowski, J. A. Sprague, and C. R. Gossett. 1992. Pulsed laser deposition of hydroxylapatite thin films on Ti-6Al-4V. *J Appl Biomater* 3:87–93.
20. Parker, E. H. C. 1985. *The technology and physics of molecular beam epitaxy.* New York: Plenum Press.
21. Comello, V. 1992. Silicon MBE research races toward production. *Res Dev* April:87–88.
22. Akedo, J., and M. Lebedev. 1999. Microstructure and electrical properties of lead zirconate titanate ($Pb(Zr_{52}/Ti_{48})O_3$) thick films deposited by aerosol deposition method. *Jpn J Appl Phys* 38:5397.
23. Jensen, K. F. 1989. *Microelectronics processing: chemical engineering aspects.* Eds. D. W. Hess, and K. F. Jensen. Washington, DC: American Chemical Society.
24. Ehrlich, D. J., R. M. J. Osgood, and J. Deutsch. 1982. Photodeposition of metal films with ultraviolet laser light. *J Vac Sci Technol* 21:23–32.
25. Wallenberger, F. T., and P. C. Nordine. 1993. Inorganic fibers and microstructures by laser assisted chemical vapor deposition. *Mater Technol* 8:198–202.
26. Granger, R. A. 1995. *Fluid mechanics.* New York: Dover Publications.
27. Wolf, S., and R. N. Tauber. 1987. *Silicon processing for the VLSI era.* Sunset Beach, CA: Lattice Press.
28. Lee, H. H. 1990. *Fundamentals of microelectronics processing.* New York: McGraw-Hill.
29. Coronell, D. G., and K. F. Jensen. 1994. Simulation of rarified gas transport and profile evolution in nonplanar substrate chemical vapor deposition. *J Electrochem Soc* 141:2545–51.
30. Adams, A. C. 1988. *VLSI technology.* Ed. S. M. Sze. New York: McGraw-Hill.
31. Hess, D. W., and K. F. Jensen, eds. 1989. *Microelectronics processing: chemical engineering aspects.* Washington, DC: American Chemical Society.
32. Sze, S. M., ed. 1988. *VLSI technology.* New York: McGraw-Hill.
33. Lee, J. G., S. H. Choi, T. C. Ahn, C. G. Hong, P. Lee, K. Law, M. Galiano, P. Keswick, and B. Shin. 1992. SA CVD: a new approach for 16 Mb dielectrics. *Semicond Int* May:115–20.
34. Huang, S., X. Cai, and J. Liu. 2003. Growth of millimeter-long and horizontally aligned single-walled carbon nanotubes on flat substrates. *J Am Chem Soc* 125:5636–37.

35. Compton, R. D. 1992. PECVD: a versatile technology. *Semicond Int* July:60–65.
36. Nguyen, S. V. 1999. High-density plasma chemical vapor deposition of silicon-based dielectric films for integrated circuits. *IBM J Res Dev* 43:109–26.
37. Burggraaf, P. 1993. The status of MOCVD technology. *Semicond Int* 16:80–83.
38. Peters, L. 1993. SOI takes over where silicon leaves off. *Semicond Int* March:48–51.
39. Anderson, T. J. 1989. *Microelectronics processing: chemical engineering aspects.* Eds. Hess, D. W., and K. F. Jensen. Washington, DC: American Chemical Society.
40. Borland, J., R. Wise, Y. Oka, M. Gangani, S. Fong, and Y. Matsumoto. 1988. Silicon epitaxial growth for advanced device structures. *Solid State Technol* January:111–19.
41. Rehrig, D. L. 1990. In search of precise epi thickness measurements. *Semicond Int* 13:90–95.
42. Singer, P. 1992. CVC builds cluster tool through partnering. *Semicond Int* 15:46.
43. Vinci, R. P., and J. C. Braveman. 1991. Mechanical testing of thin films. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
44. Bushan, B. 1996. *Proceedings: IEEE ninth annual international workshop on micro electro mechanical systems.* San Diego, CA.
45. Bushan, B., ed. 1995. *Handbook of micro/nanotribology.* Boca Raton, FL: CRC Press.
46. Schiotz, J., and K. W. Jacobsen. 2003. A maximum in the strength of nanocrystalline copper. *Science* 301:1357–59.
47. Schiotz, J., F. D. D. Tolla, and K. W. Jacobsen. 1998. Softening of nanocrystalline metals at very small grains. *Nature* 391:561.
48. Chokshi, A. H., A. Rosen, J. Karch, and H. Gleiter. 1989. On the validity of the hall-petch relationship in nanocrystalline materials. *Scripta Metallurgica* 23:1679–83.
49. Gleiter, H. 1989. Nanocrystalline materials. *Prog Mater Sci* 33:223–315.
50. Yu, D. Y. W., and F. Spaepen. 2004. The yield strength of thin copper films on Kapton. *J Appl Phys* 95:2991–97.
51. Ehmann, K. F., D. Bourell, M. L. Culpepper, T. J. Hodgson, T. R. Kurfess, M. Madou, K. Rajukar, and R. DeVor. 2007. *Micromanufacturing: international research and development.* New York: Springer.
52. Biebl, M., G. Brandl, and R. T. Howe. 1995. Young's modulus of in-situ phosphorus-doped polysilicon. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
53. Campbell, D. S. 1970. *Handbook of thin film technology.* Eds. L. I. Maissel, and R. Glang. New York: McGraw-Hill.
54. Greek, S., F. Ericson, S. Johansson, and J.-Å. Schweitz. 1995. In situ tensile strength measurement of thick-film and thin-film micromachined structures. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
55. Senturia, S. 1987. *Proceedings: IEEE micro robots and teleoperators workshop*. Hyannis, MA.
56. Krulevitch, P. A. 1994. *Micromechanical investigations of silicon and Ni–Ti–Cu thin films.* PhD thesis, University of California, Berkeley.
57. Pratt, R. I., G. C. Johnson, R. T. Howe, and D. J. J. Nikkel. Characterization of thin films using micromechanical structures. *Mat Res Soc Symp Proc* 276:197–202.
58. Judy, M. W., Y. H. Cho, R. T. Howe, and A. P. Pisano. 1991. *Proceedings: IEEE micro electro mechanical systems (MEMS '91)*. Nara, Japan.
59. Wu, T. H. T., and R. S. Rosler. 1992. Stress in PSG and nitride films as related to film properties and annealing. *Solid State Technol* May:65–71.
60. Hey, H. P. W., B. G. Sluijk, and D. G. Hemmes. 1990. Ion bombardment factor in plasma CVD. *Solid State Technol* April:139–44.
61. Chou, P. C., and N. J. Pagano. 1967. *Elasticity: tensor, dyadic, and engineering approaches.* New York: Dover Publications.
62. Guckel, H., D. W. Burns, H. A. C. Tilmans, C. C. G. Visser, D. W. DeRoo, T. R. Christenson, P. J. Klomberg, J. J. Sniegowski, and D. H. Jones. 1988. Processing conditions for polysilicon films with tensile strain for large aspect ratio microstructures. *Technical digest: 1988 solid state sensor and actuator workshop.* Hilton Head Island, SC.
63. Maseeh, F., and S. D. Senturia. 1990. *Technical digest: 1990 solid state sensor and actuator workshop*. Hilton Head Island, SC.
64. Hoffman, R. W. 1976. *Physics of nonmetallic thin films (NATO advanced study institutes series: series B, physics.* Eds. C. H. S. Dupuy, and A. A. Cachard. New York: Plenum Press.
65. Hoffman, R. W. 1975. Stresses in thin films: the relevance of grain boundaries and impurities. *Thin Solid Films* 34:185–90.
66. Iscoff, R. 1991. Hotwall LP CVD reactors: considering the choices. *Semicond Int* June:60–64.
67. Adams, A. C. 1988. *VLSI technology.* Ed. S. M. Sze. New York: McGraw-Hill.
68. Guckel, H., J. J. Sniegowski, T. R. Christenson, and F. Raissi. 1990. The application of fine-grained, tensile polysilicon to mechanically resonant transducers. *Sensors Actuators A* A21:346–51.
69. van der Drift, A. 1967. Evolutionary selection, a principle governing growth orientation in vapour-deposited layers. *Philips Res Rep* 22:267–88.
70. Matson, E. A., and S. A. Polysakov. 1977. On the evolutionary selection principle in relation to the growth of polycrystalline silicon films. *Phys Sta Sol (A)* 41:K93–K95.
71. Lober, T. A., J. Huang, M. A. Schmidt, and S. D. Senturia. 1988. Characterization of the mechanisms producing bending moments in polysilicon micro-cantilever beams by interferometric deflection measurements. *Technical digest: 1988 solid state sensor and actuator workshop*. Hilton Head Island, SC.
72. Shimizu, T., and S. Ishihara. 1995. Effect of SiO_2 surface treatment on the solid-phase crystallization of amorphous silicon films. *J Electrochem Soc* 142:298–302.
73. Abe, T., and M. L. Reed. 1996. *Ninth annual international workshop on micro electro mechanical systems.* San Diego, CA.
74. Guckel, H., D. W. Burns, C. C. G. Visser, H. A. C. Tilmans, and D. Deroo. 1988. Fine-grained polysilicon films with built-in tensile strain. *IEEE Trans Electron Devices* 35: 800–01.
75. Drosd, R., and J. Washburn. 1982. Some observation on the amorphous to crystalline transformation in silicon. *J Appl Phys* 53:397–403.
76. Elwenspoek, M., U. Lindberg, H. Kok, and L. Smith. 1994. *IEEE international workshop on micro electro mechanical systems (MEMS '94)*. Oiso, Japan.
77. Harbeke, G., L. Krausbauer, E. F. Steigmeier, and A. E. Widmer. 1983. LP CVD polycrystalline silicon: growth and physical properties of in-situ phosphorous doped and undoped films. *RCA Rev* 44:287–313.
78. Kamins, T. 1988. *Polycrystalline silicon for integrated circuits*. Boston: Kluwer.

79. Mulder, J. G. M., P. Eppenga, M. Hendriks, and J. E. Tong. 1990. An industrial LP CVD process for in situ phosphorus-doped polysilicon. *J Electrochem Soc* 137:273–79.
80. Kinsbron, E., M. Sternheim, and R. Knoell. 1983. Crystallization of amorphous silicon films during low pressure chemical vapor deposition. *Appl Phys Lett* 42: 835–37.
81. Lietoila, A., A. Wakita, T. W. Sigmon, and J. F. Gibbons. 1982. Epitaxial regrowth of intrinsic, ^{31}P-doped and compensated ($^{31P+11}$B-Doped) amorphous Si. *J Appl Phys* 53:4399–405.
82. Meyerson, B. S., and W. Olbricht. 1984. Phosphorous-doped polycrystalline silicon via LPCVD I. process characterization. *J Electrochem Soc* 131:2361–65.
83. Hendriks, M., and C. Mavero. 1991. Phosphorous doped polysilicon for double poly structures. Part I. morphology and microstructure. *J Electrochem Soc* 138:1466–70.
84. Kurokawa, H. 1982. P-doped polysilicon film growth technology. *J Electrochem Soc* 129:2620–24.
85. Biebl, M., and H. von Philipsborn. 1993. Fracture strength of doped and undoped panical filters. PhD thesis. University of California, Berkeley.
86. Maluf, N. 2000. An introduction to microelectromechanical systems engineering. Boston: Artech House.
87. Compton, R., D. 1992. PECVD: a versatile technology. *Semicond Int* July:60–65.
88. Honer, K. A., and G. T. A. Kovacs. 2000. *Solid-state sensor and actuator workshop.* Hilton Head Island, SC: Transducers Research Foundation.
89. Sharpe, W. N., B. Yuan, R. Vaidyanathan, and R. L. Edwards. 1997. *Tenth IEEE international workshop on microelectromechanical systems.* Nagoya, Japan: IEEE.
90. Jerman, H. 1994. Hard copies of viewgraphs presented in Banff, Canada.
91. Biebl, M., and H. von Philipsborn. 1993. *Fracture strength of doped and undoped panical filters.* PhD thesis, University of California, Berkeley.
92. Greek, S., F. Ericson, S. Johansson, and J.-Å. Schweitz. 1995. In situ tensile strength measurement of thick-film and thin-film micromachined structures. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
93. Le Berre, M., P. Kleinmann, B. Semmache, D. Barbier, and P. Pinard. 1996. *Smart electronics and MEMS: proceedings of the smart structures and materials 1996 meeting.* Eds. V. K. Varadan, and P. J. McWhorter. San Diego: SPIE.
94. Howe, R. T. 1995. Recent advances in surface micromachining. *Technical digest: 13th sensor symposium*. Tokyo, Japan.
95. Diem, B., M. T. Delaye, F. Michel, S. Renard, and G. Delapierre. 1993. SOI(SIMOX) as a substrate for surface micromachining of single crystalline silicon sensors and actuators. *Seventh international conference on solid-state sensors and actuators (Transducers '93)*. Yokohama, Japan: IEEE.
96. Pister, K. S. J. 1992. Hinged polysilicon structures with integrated CMOS TFTS. *Technical digest: 1992 solid state sensor and actuator workshop.* Hilton Head Island, SC.
97. Keller, C., and M. Ferrari. 1994. Milli-scale polysilicon structures. *Technical digest: 1994 solid state sensor and actuator workshop*. Hilton Head Island, SC.
98. Lange, P., M. Kirsten, W. Riethmuller, B. Wenk, G. Zwicker, J. R. Morante, F. Ericson, and J.-Å. Schweitz. 1995. Thick polycrystalline silicon for surface micromechanical applications: deposition, structuring and mechanical. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
99. Noworolski, J. M., E. Klaassen, J. Logan, K. Petersen, and N. Maluf. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden: IEEE.
100. Chu, P. B., P. R. Nelson, M. L. Tachiki, and K. S. J. Pister. 1995. Dynamics of polysilicon parallel-plate electrostatic actuators. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
101. Memming, R., and G. Schwandt. 1966. Anodic dissolution of silicon in hydrofluoric acid solutions. *Surf Sci* 4:109–24.
102. Zhang, X. G., S. D. Collins, and R. L. Smith. 1989. Porous silicon formation and electropolishing of silicon by anodic polarization in HF solution. *J Electrochem Soc* 136:1561–65.
103. Anderson, R. C., R. S. Muller, and C. W. Tobias. 1994. Porous polycrystalline silicon: a new material for MEMS. *J Microelectromech Syst* 3:10–18.
104. Nathanson, H. C., W. E. Newell, R. A. Wickstrom, and J. R. Davis. 1967. The resonant gate transistor. *IEEE Trans Electron Devices* ED–14:117–33.
105. Dutta, B., P. Dev, P. Dewilde, B. Sharma, and R. Newcomb. 1970. Integrated micromotors concepts. *International conference on microelectronic circuits and systems theory*. Sydney, Australia.
106. Howe, R. T., and R. S. Muller. 1982. Polycrystalline silicon micromechanical beams. *Spring meeting of the electrochemical society.* Montreal, Canada.
107. Guckel, H., and D. W. Burns. 1985. A technology for integrated transducers. *International conference on solid-state sensors and actuators.* Philadelphia.
108. Gabriel, K., J. Jarvis, and W. Trimmer. 1988. *Small machines, large opportunities: a report on the emerging field of microdynamics*. Murray Hill, NJ: National Science Foundation, published by AT&T Bell Laboratories.
109. Muller, R. S. 1987. *Proceedings: IEEE micro robots and teleoperators workshop*. Hyannis, MA: IEEE.
110. Fan, L.-S., Y. C. Tai, and R. S. Muller. 1987. Pin joints, gears, springs, cranks, and other novel micromechanical structures. *Fourth international conference on solid-state sensors and actuators (Transducers '87)*. Tokyo, Japan, June 2–5, 1987, pp. 849–52.
111. Editor. 1991. Analog devices combine micromachining with BICMOS. *Semicond Int*14:17.
112. Hornbeck, L. J. 1995. *Micromachining and microfabrication process technology.* Austin, TX.
113. Hermansson, K., U. Lindberg, B. Hok, and G. Palmskog. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
114. Tang, W. C. K. 1990. *Electrostatic comb drive for resonant sensor and actuator applications.* PhD thesis, University of California, Berkeley.
115. Adams, A. C. 1988. *VLSI technology.* Ed. S. M. Sze. New York: McGraw-Hill.
116. Howe, R. T. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
117. Biebl, M., G. T. Mulhern, and R. T. Howe. 1995. In situ phosphorus doped polysilicon for integrated MEMS. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
118. Fresquet, G., C. Azzaro, and J.-P. Couderc. 1995. Analysis and modeling of in situ boron-doped polysilicon deposition by LP CVD. *J Electrochem Soc* 142:538–47.

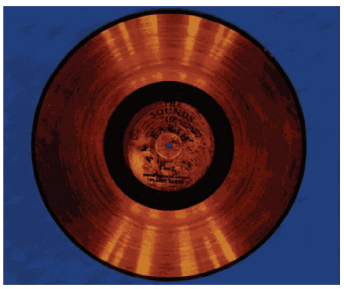
119. Core, T. A., W. K. Tsang, and S. J. Sherman. 1993. Fabrication technology for an integrated surface-micromachined sensor. *Solid State Technol* 36:39–47.
120. Monk, D. J., D. S. Soane, and R. T. Howe. 1992. LPCVD silicon dioxide sacrificial layer etching for surface micromachining. *Symposium on "smart" materials fabrication/materials for micro-electro-mechanical systems.* Materials Research Society Spring Meeting, San Francisco, April 27–May 1, 1992, MRS Symposium Proceedings, pp. 303–10.
121. Tenney, A. S., and M. Ghezzo. 1973. Etch rates of doped oxides in solutions of buffered HF. *J Electrochem Soc* 120:1091–95.
122. Levy, R. A., and K. Nassau. 1986. Viscous behavior of phosphosilicate and borophosphosilicate glasses in VLSI processing. *Solid State Technol* October:123–30.
123. North, J. C., T. E. McGahan, D. W. Rice, and A. C. Adams. 1978. Tapered windows in phosphorous-doped silicon dioxide by ion implantation. *IEEE Trans Electron Devices* ED–25:809–12.
124. Goetzlich, J., and H. Ryssel. 1981. Tapered windows in silicon dioxide, silicon nitride, and polysilicon layers by ion implantation. *J Electrochem Soc* 128:617–19.
125. White, L. K. 1980. Bilayer taper etching of field oxides and passivation layers. *J Electrochem Soc* 127:2687–93.
126. Monk, D. J., D. S. Soane, and R. T. Howe. 1994. Hydrofluoric acid etching of silicon dioxide sacrificial layers-part I: experimental observations. *J Electrochem Soc* 141:264–69.
127. Monk, D. J., D. S. Soane, and R. T. Howe. 1994. Hydrofluoric acid etching of silicon dioxide sacrificial layers-part II: modeling. *J Electrochem Soc* 141:270–74.
128. Eaton, W. P., and J. H. Smith. 1996. *SPIE proceedings: micromachining and microfabrication process technology*. Austin, TX: SPIE.
129. Liu, J., Y.-C. Tai, J. Lee, K.-C. Pong, Y. Zohar, and C.-H. Ho. 1993. In situ monitoring and universal modelling of sacrificial PSG etching using hydrofluoric acid. *Proceedings: micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL.
130. Goossen, K., J. Walker, and S. Arney. 1994. Silicon modulator based on mechanically-active antireflection layer with 1 Mbit/sec capability. *IEEE Photonics Tech Lett* 6:1119.
131. Lober, T. A., and R. T. Howe. 1988. Surface-micromachining processes for electrostatic microactuator fabrication. *Technical digest: 1988 solid state sensor and actuator workshop*. Hilton Head Island, SC.
132. Fan, L.-S., Y. C. Tai, and R. S. Muller. 1988. Integrated movable micromechanical structures for sensors and actuators. *IEEE Trans Electron Devices* 35:724–30.
133. Mehregany, M., K. J. Gabriel, and W. S. N. Trimmer. 1988. Integrated fabrication of polysilicon mechanisms. *IEEE Trans Electron Devices* 35:719–23.
134. Walker, J. A., K. J. Gabriel, and M. Mehregany. 1991. Mechanical integrity of polysilicon films exposed to hydrofluoric acid solutions. *J Electron Mater* 20:665–70.
135. Eaton, W. P., and J. H. Smith. 1995. *Smart structures and materials 1995: smart electronics (Proceedings of the SPIE)*. San Diego, CA.
136. Watanabe, H., S. Ohnishi, I. Honma, H. Kitajima, H. Ono, R. J. Wilhelm, and A. J. L. Sophie. 1995. Selective etching of phosphosilicate glass with low pressure vapor HF. *J Electrochem Soc* 142:237–43.
137. Nakamura, M., S. Hoshino, and H. Muro. 1985. Monolithic sensor device for detecting mechanical vibration. *Densi Tokyo* (*IEEE Tokyo* Section) 24:87–88.
138. Bassous, E., and C.-Y. Liu. 1978. Polycrystalline silicon etching with tetramethylammonium hydroxide.US Patent 4113551.
139. Gennissen, P. T. J., P. J. French, D. P. A. D. Munter, T. E. Bell, H. Kaneko, and P. M. Sarro. 1995. Porous silicon micromachining techniques for acceleration fabrication. *Proceedings of the European Solid-State Device Research Conference. ESSDERC '95*. The Hague, the Netherlands.
140. Chen, L.-Y., and N. C. MacDonald. 1991. Selective tungsten process. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
141. Sugiyama, S., T. Suzuki, K. Kawahata, K. Shimaoka, M. Takigawa, and I. Igarashi. 1986. Micro-diaphragm pressure sensor. *Technical digest: IEEE international electron devices meeting (IEDM '86)*.
142. Sugiyama, S., K. Kawakata, M. Abe, H. Funabashi, and I. Igarashi. 1987. *Fourth international conference on solid-state sensors and actuators (Transducers '87)*. Tokyo, Japan.
143. Kim, Y. W., and M. G. Allen. 1991. Surface micromachined platforms using electroplated sacrificial layers. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco: IEEE.
144. Suzuki, K., I. Shimoyama, and H. Miura. 1994. Insect–model based microrobot with elastic hinges. *J Microelectromech Syst* 3, 4–9.
145. Chang, S., W. Eaton, J. Fulmer, C. Gonzalez, B. Underwood, J. Wong, and R. L. Smith. 1991. Micromechanical structures in amorphous silicon. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
146. Scheeper, P. R., W. Olthuis, and P. Bergveld. 1991. Fabrication of a subminiature silicon condenser microphone using the sacrificial layer technique. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
147. Yamada, K., and T. Kuriyama. 1991. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
148. van Mullem, C. J., K. J. Gabriel, and H. Fujita. 1991. Large deflection performance of surface micromachined corrugated diaphragms. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco: IEEE.
149. Mastrangelo, C. H., and C. H. Hsu. 1993. Mechanical stability and adhesion of microstructures under capillary forces—part I: basic theory. *J Microelectromech Syst* 2:33–43.
150. Mastrangelo, C. H., and C. H. Hsu. 1993. Mechanical stability and adhesion of microstructures under capillary forces—part II: experiments. *J Microelectromech Syst* 2:44–55.
151. Fan, L.-S. 1989. *Integrated micromachinery: moving structures on silicon chips.* PhD thesis. University of California, Berkeley.
152. Abe, T., W. C. Messner, and M. L. Reed. 1995. *Proceedings: IEEE micro electro mechanical systems (MEMS '95)*. Amsterdam, the Netherlands.
153. Fedder, G. K., J. C. Chang, and R. T. Howe. 1992. Thermal assembly of polysilicon microstructures with narrow-gap electrostatic comb drive. *Technical digest: 1992 solid state sensor and actuator workshop.* Hilton Head Island, SC.
154. Mastrangelo, C. H., and G. S. Saloka. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL: IEEE.
155. Guckel, H., J. J. Sniegowski, T. R. Christenson, S. Mohney, and T. F. Kelly. 1989. Fabrication of micromechanical devices from polysilicon films with smooth surfaces. *Sensors Actuators* 20:117–21.

156. Guckel, H., J. J. Sniegowski, T. R. Christenson, and F. Raissi. 1990. The application of fine-grained, tensile polysilicon to mechanically resonant transducers. *Sensors Actuators A* A21:346–51.
157. Takeshima, N., K. J. Gabriel, M. Ozaki, J. Takahashi, H. Horiguchi, and H. Fujita. 1991. Electrostatic parallelogram actuators. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
158. Mulhern, G. T., D. S. Soane, and R. T. Howe. 1993. Supercritical carbon dioxide drying of microstructures. *Seventh international conference on solid-state sensors and actuators (Transducers '93)*. Yokohama, Japan.
159. Kozlowski, F., N. Lindmair, T. Scheiter, C. Hierold, and W. Lang. 1995. A novel method to avoid sticking of surface micromachine structures. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
160. Alley, R. L., P. Mai, K. Komvopoulos, and R. T. Howe. 1993. Surface roughness modification of interfacial contacts in polysilicon microstructures. *Seventh international conference on solid-state sensors and actuators (Transducers '93)*. Yokohama, Japan.
161. Alley, R. L., R. T. Howe, and K. Komvopoulos. 1992. The effect of release etch processing on the surface microstructure stiction. *Technical digest: 1992 solid state sensor and actuator workshop*. Hilton Head Island, SC.
162. Man, P. F., B. P. Gogoi, and C. H. Mastrangelo. 1996. Elimination of post-release adhesion in microstructures using thin conformal fluorocarbon films. *Ninth annual international workshop on micro electro mechanical systems (MEMS '96)*. San Diego, CA..
163. Houston, M. R., R. Maboudian, and R. Howe. 1995. Ammonium fluoride antistiction treatments for polysilicon microstructures. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
164. Gogoi, B. P., and C. H. Mastrangelo. 1995. Post-processing release of microstructures by electromagnetic pulses. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
165. Li, J., J. Xu, and C. P. Wong. 2004. Lotus effect coating and its application for microelectromechanical systems stiction prevention. *Proceedings of the 54th Electronic Components and Technology Conference (ECTC)*. Las Vegas, NV.
166. Nguyen, C. T. C., and R. T. Howe. 1993. *Electron devices meeting, 1993. IEDM '93. Technical Digest, International*. Washington DC: IEEE.
167. Guckel, H., T. Randazzo, and D. W. Burns. 1985. A simple technique for the determination of mechanical strain in thin films with applications to polysilicon. *J Appl Phys* 57:1671–75.
168. Howe, R. T., and R. S. Muller. 1983. Stress in polycrystalline and amorphous silicon thin films. *J Appl Phys* 54:4674–75.
169. Murarka, S. P., and T. F. J. Retajczyk. 1983. Effect of phosphorous doping on stress in silicon and polycrystalline silicon. *J Appl Phys* 54:2069–72.
170. Orpana, M., and A. O. Korhonen. 1991. Control of residual stress of polysilicon thin films by heavy doping in surface micromachining. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
171. Lin, L., R. T. Howe, and A. P. Pisano. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL: IEEE.
172. Choi, M. S., and E. W. Hearn. 1984. Stress effects in boron-implanted polysilicon films. *J Electrochem Soc* 131:2443–46.
173. Ding, X., and W. Ko. 1991. Buckling behavior of boron-doped P+ silicon diaphragms. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
174. Hendriks, M., R. Delhez, and S. Radelaar. 1983. *Studies in inorganic chemistry*. Amsterdam: Elsevier.
175. Tang, W. C., T.-C. H. Nguyen, and R. T. Howe. 1989. Proceedings: IEEE micro electro mechanical systems (MEMS '89). Salt Lake City, UT.
176. Sekimoto, M., H. Yoshihara, and T. Ohkubo. 1982. Silicon nitride single-layer x-ray mask. *J Vac Sci Technol* 21:1017–21.
177. Guckel, H., D. K. Showers, D. W. Burns, C. R. Rutigliano, and C. G. Nesler. 1986. Deposition techniques and properties of strain compensated LP CVD silicon nitride. *Technical digest: 1986 solid state sensor and actuator workshop*. Hilton Head Island, SC.
178. Tang, W. C., T. H. Nguyen, and R. T. Howe. 1989. Laterally driven polysilicon resonant microstructures. *Sensors Actuators* 20:25–32.
179. Jerman, J. H. 1990. The fabrication and use of micromachined corrugated silicon diaphragms. *Sensors Actuators A* A23:988–92.
180. Spiering, V. L., S. Bouwstra, R. M. E. J. Spiering, and M. Elwenspoek. 1991. On-chip decoupling zone for package-stress reduction. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
181. Offereins, H. L., H. Sandmaier, B. Folkmer, U. Steger, and W. Lang. Stress-free assembly technique for a silicon-based pressure sensor. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
182. Spiering, V. L., S. Bouwstra, J. Burger, and M. Elwenspoek. 1993. Membranes fabricated with a deep single corrugation for package stress reduction and residual stress relief. *Fourth European workshop on micromechanics (MME '93)*. Neuchatel, Switzerland.
183. Zhang, Y., and K. D. Wise. 1994. Performance of non-planar silicon diaphragms under large deflections. *J Microelectromech Syst* 3:59–68.
184. Scheeper, P. R., W. Olthuis, and P. Bergveld. 1994. The design, fabrication, and testing of corrugated silicon nitride diaphragms. *J Microelectromech Syst* 3:36–42.
185. Yun, W. 1992. *A surface micromachined accelerometer with integrated CMOS detection circuitry*. PhD thesis. University of California, Berkeley.
186. Guckel, H., D. W. Burns, H. A. C. Tilmans, D. W. DeRoo, and C. R. Rutigliano. 1988. Mechanical properties of fine grained polysilicon—the repeatability issue. *Technical digest: 1988 solid state sensor and actuator workshop*. Hilton Head Island, SC.
187. Nguyen, C. T.-C. 1998. *Aerospace conference*. Snowmass, CO: IEEE.
188. Smith, J., S. Montague, J. Sniegowski, and P. McWhorter. 1995. *Technical digest: IEEE international electron devices meeting (IEDM '95)*. Washington, DC.
189. Varadan, V. K., and P. J. McWhorter, eds. 1996. *Smart electronics and MEMS: proceedings of the smart structures and materials 1996 meeting*. San Diego: SPIE.
190. Dunn, P. N. 1993. SOI: ready to meet CMOS challenge. *Solid State Technol* October32–35.
191. Abe, T., and J. H. Matlock. 1990. Wafer bonding technique for silicon-on-insulator technology. *Solid State Technol* November:39–40.

192. Kanda, Y. 1991. What kind of SOI wafers are suitable for what type of machining purposes? *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
193. Kuhn, G. L., and C. J. Rhee. 1973. Thin silicon film on insulating substrate. *J Electrochem Soc* 120:1563–66.
194. Pourahmadi, F., L. Christel, and K. Petersen. 1992. Silicon accelerometer with new thermal self-test mechanism. *Technical digest: 1992 solid state sensor and actuator workshop*. Hilton Head Island, SC.
195. Gennissen, P. T. J., M. Bartke, P. J. French, P. M. Sarro, and R. F. Wolffenbuttel. 1995. Automatic etch stop on buried oxide using epitaxial lateral overgrowth. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
196. Bartek, M., P. T. J. Gennissen, P. J. French, and R. F. Wolffenbuttel. 1995. Confined selective epitaxial growth: potential for smart silicon sensor fabrication *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden: IEEE.
197. Maluf, N. 2000. *An introduction to microelectromechanical systems engineering*. Boston: Artech House.
198. Greiff, P. 1995. *Micromachining and microfabrication process technology. Proceedings of the SPIE*. Austin, TX: SPIE.
199. Neudeck, G. W., P. J. Schuberts, J. L. Glenn, Jr., J. A. Friedrich, W. A. Klaasan, R. P. Zingg, and J. P. Denton. 1990. Three dimensional devices fabricated by silicon epitaxial lateral overgrowth. *J Electron Mater* 19:1111–17.
200. Schubert, P. J., and G. W. Neudeck. 1990. Confined lateral selective epitaxial growth of silicon for device fabrication. *IEEE Electron Device Lett* 11:181–83.
201. Bashir, R., G. W. Neudeck, H. Yen, E. P. Kvam, and J. P. Denton. 1995. Characterization of sidewall defects in selective epitaxial growth of silicon. *J Vac Sci Technol* 11:923–28.
202. Luder, E. 1986. Polycrystalline silicon-based sensors. *Sensors Actuators* 10:9–23.
203. Obermeier, E., and P. Kopystynski. 1992. Polysilicon as a material for microsensor applications. *Sensors Actuators A* A30:149–55.
204. Voronin, V. A., A. A. Druzhinin, I. I. Marjamora, V. G. Kostur, and J. M. Pankov. 1992. Laser-recrystallized polysilicon layers in sensors. *Sensors Actuators A* A30:143–47.
205. Burggraaf, P. 1991. Epi's leading edge. *Semicond Int* June:67–71.
206. Lin, L. Y., S. S. Lee, M. C. Wu, and K. S. J. Pister. 1995. *Proceedings: IEEE micro electro mechanical systems (MEMS '95)*. Amsterdam, the Netherlands.
207. Burgett, S. R., K. S. Pister, and R. S. Fearing. 1992. *ASME 1992, micromechanical sensors, actuators, and systems.* Anaheim, CA: ASME.
208. Ross, M., and K. Pister. 1994. Micro-windmill for optical scanning and flow measurement. *Sensors Actuators A* 46–47:576–79.
209. Lin, G., K. S. J. Pister, and K. P. Roos. 1996. Standard CMOS piezoresistive sensor to quantify heart cell contractile forces. *Ninth annual international workshop on micro electro mechanical systems (MEMS '96)*. San Diego, CA.
210. Yeh, R., E. J. Kruglick, and K. S. J. Pister. 1994. *ASME 1994, micromechanical sensors, actuators, and systems.* Chicago: ASME.
211. Hoffman, E., B. Warneke, E. Kruglick, J. Weigold, and K. S. J. Pister. 1995. *Proceedings: IEEE micro electro mechanical systems (MEMS '95)*. Amsterdam, the Netherlands: IEEE.
212. Keller, C. G., and R. T. Howe. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*, 376–79. Stockholm, Sweden.
213. Keller, C. G., and R. T. Howe. 1995. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*, 99–102. Stockholm, Sweden.
214. Kahn, H., S. Stemmer, K. Nandakumar, A. H. Heuer, R. L. Mullen, R. Ballarini, and M. A. Huff. 1996. Mechanical properties of thick, surface micromachined polysilicon films. *Ninth annual international workshop on micro electro mechanical systems (MEMS '96)*. San Diego, CA.
215. Brodie, I., H. R. Gurnick, C. E. Holland, and H. A. Moessner. 1990. *Method for providing polyimide spacers in a field emission panel display. US Patent 4923421*.
216. Frazier, A. B., and M. G. Allen. 1993. Metallic microstructures fabricated using photosensitive polyimide electroplating molds. *J Microelectromech Syst* 2:87–94.
217. Ahn, C. H., Y. J. Kim, and M. G. Allen. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL: IEEE.
218. Acosta, R. E., C. Ahn, I. V. Babich, E. I. Cooper, J. M. Cotte, W. J. Horkans, C. Jahnes, S. Krongelb, K. T. Kwietniak, N. C. Labianca, et al. 1995. Integrated variable reluctance magnetic mini–motor. *J Electrochem Soc* 95-2:494–95.
219. LaBianca, N. C., J. D. Gelorme, E. Cooper, E. O'Sullivan, and J. Shaw. 1995. *JECS 188th Meeting*. Chicago: JECS.
220. Field, L. 1987. Electrical Engineering and Computer Sciences/Electronics Research Lab; Research summary. University of California, Berkeley.
221. Judy, M. W., and R. T. Howe. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL.
222. Judy, M. W., and R. T. Howe. 1993. Highly compliant lateral suspensions using sidewall beams. *Seventh international conference on solid-state sensors and actuators (Transducers '93)*. Yokohama, Japan.
223. Lebouitz, K. S., R. T. Howe, and A. P. Pisano. 1995. Permeable polysilicon etch-access windows for microshell fabrication. *Eighth international conference on solid-state sensors and actuators (Transducers '95)*. Stockholm, Sweden.
224. Mardilovich, P., D. Routkevitch, and A. Govyadinov. 2000. Eds. P. J. Hesketh, S. S. Ang, W. E. Bailey, J. L. Davidson, H. G. Hughes, and D. Misra. *Micro-fabricated systems and MEMS*. Electrochemical Society: Pennington, Vol. 2000–19, p. 33.
225. Lin, L. 1993. *Selective encapsulations of MEMS: micro channels, needles, resonators, and electromechanical filters.* PhD thesis. University of California, Berkeley.
226. Kamins, T. 1988. *Polycrystalline silicon for integrated circuits.* Boston: Kluwer.
227. Heuberger, A. 1989. *Mikromechanik*. Heidelberg: Springer Verlag.
228. Sharpe, W. N., B. Yuan, R. Vaidyanathan, and R. L. Edwards. 1997. *Tenth IEEE international workshop on microelectromechanical systems.* Nagoya, Japan: IEEE.
229. Petersen, K., D. Gee, F. Pourahmadi, R. Craddock, J. Brown, and L. Christel. 1991. Resonant beam pressure sensor fabricated with silicon fusion bonding. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
230. Chrisey, D. B., and G. K. Hubler. (Eds.) 1994. *Pulsed laser deposition of thin films*. New York: John Wiley and Sons.
231. 2006. *Atomic layer deposition*. Cambridge NanoTech Inc., April 24. http://www.cambridgenanotech.com

8

Chemical, Photochemical, and Electrochemical Forming Techniques



"The Sounds of Earth," a gold-plated copper record placed on Voyager. Two hours of sound and a movie plus some pictures and a message from President Jimmy Carter (From http://voyager.jpl.nasa.gov/spacecraft/goldenrec.html.)

Outline

Introduction

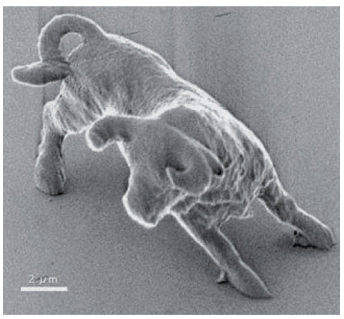
Chemical Forming

Photochemical Forming

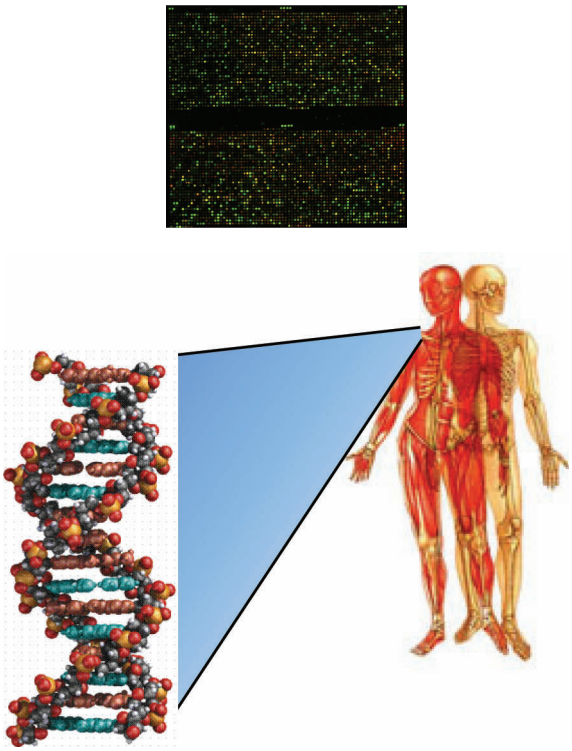
Electrochemical Forming Processes

Questions

References



Photochemical forming: Light causes a polymer to solidify. Structural details of 120 nm (as a result of two-photon photopolymerization). (From Kawata, S., H.-B. Sun, T. Tanaka, and K. Takada, 2001. Finer features for functional microdevices. *Nature* 412:697–698.[86])



DNA array and the human genome.

> Elderly: In physics, mathematics and astronautics it means over thirty; in other disciplines, senile decay is sometimes postponed to the forties. There are of course, glorious exceptions; but as every researcher just out of college knows, scientists of over fifty are good for nothing but board meetings, and should at all costs be kept out of the laboratory.
>
> **Arthur C. Clarke**
> *Profiles of the Future*

Introduction

In this chapter we review chemical, photochemical, and electrochemical forming.* In chemical forming, chemical reactions create features by forming new compounds. The energy source at the work piece in this additive process is chemical. The method is extremely broad and covers fabrication with metals, ceramics, and polymers. Reactions are induced by combining different reagents and are often aided by heat and/or catalysts and also may be accompanied by a shape-forming step using, for example, a spraying nozzle, a mold, or a template. Almost all manufacturing processes include one or more chemical forming steps and can be part of lithography- or nonlithography-based schemes. Chemical and thermal forming processes are difficult to distinguish because they are almost always intertwined. Some chemical forming processes involve an exothermic reaction, and such processes we classify as purely chemical (they do not require the application of external heat). This is the case, for example, in reaction injection molding (RIM; see Chapter 10) and for self-assembled monolayers (SAMs; see page 523 below). Other chemical forming processes are endothermic and require a significant increase in temperature for the reaction to initiate such as in the oxidation of Si (see Volume I, Chapter 4) or chemical vapor deposition of polycrystalline Si (CVD; see Chapter 7). The distinction between chemical and thermal forming is somewhat arbitrary, and in this chapter we cover as additional chemical forming methods those low temperature material deposition techniques that are used in chemical and biological sensor manufacture, often arranged in some type of an array configuration. The additional chemical forming methods covered are sol-gel deposition from alkoxide-based precursors, organic film spin coating, polymer dry film lamination, polymer dip coating, polymer spraying, polymer casting, organic film doctor's blade (also called *knife coating*), glow discharge polymerization, low temperature silk screening of organics, Langmuir-Blodgett deposition, and SAMs. Some of the chemical forming processes discussed here involve self-assembly such as in the case of SAMs, in which case one refers to bottom-up manufacturing. Whereas top-down nanofabrication is principally based on lithography and traditional mechanical machining, etching, and grinding, bottom-up manufacturing, which we also call nanochemistry in this book, is based in chemical synthesis, self-assembly, and positional assembly (see Volume III, Chapter 3). Arraying of low temperature materials (often organics) for sensor arrays may be accomplished by lithography, digital mirror deposition (a type of lithography as well), ink-jetting, mechanical microspotting, and

* In Part II on subtractive manufacturing techniques, we did make the same distinction between chemical, photochemical, and electrochemical etching. All three subtractive manufacturing techniques were covered in Chapter 4 on wet chemical etching and wet bulk micromachining-pools as tools.

microcontact printing. Some of the same additive methods discussed in this section will reappear in Chapter 9 on thermal forming where they involve the manufacture of ceramics in which heat application is more essential.

In photochemical forming, photoenergy solidifies a material into a three-dimensional (3D) shape as we illustrate with our description of rapid prototyping and microphotoforming (stereolithography) with, as newest embodiments, parallel stereolithography and two-photon polymerization.

In the case of electrochemical forming, the energy at the work piece is electrochemical. Electrochemical and electroless metal deposition are gaining renewed interest because of their emerging importance in making mold inserts for replication meth-odologies in both microelectromechanical sys-tems (MEMS) and nanoelectromechanical systems (NEMS) (see also Chapter 10). The foundations underlying electrochemical deposition were covered in detail in Volume I, Chapter 7; here we contrast electroless metal deposition with electrodeposition and introduce some additional electrochemical techniques. These new electrodeposition methods sort out into through-mask plating, such as in the fabrication of thin-film read-write heads, the damascene process, and in anodization; instant masking in EFAB (Electrochemical FABrication); and maskless metal deposition methods such as microjet plating, laser-enhanced jet plating, local electrochemical deposition using scanning electrochemical microscopes (SECMs), and techniques to electrodeposit slanted and curved surfaces.

Chemical Forming

Introduction

It should come as no surprise that chemical forming techniques are gaining importance because today bottom-up manufacturing or nanochemistry, a type of chemical forming, constitutes one of the most intensely researched topics. Bottom-up fabrication or nanochemistry constitutes a special class of chemical forming incorporating chemical synthesis of small particles, self-assembly, and positional assembly. In nanochemistry device structures are created via systematic assembly of atoms, molecules, or other basic units of matter. Instead of manipulating Si wafers or other solid substrates, we manipulate small amounts of chemicals, even molecules and atoms. Molecular engineers, manipulating and modifying natural polymers such as DNA and proteins, led the way into nanochemistry and practiced this type of nanotechnology long before electrical and mechanical engineers got involved. Now we are designing and manipulating at the molecular level, whereas before it was either evolution that did it for us (e.g., a virus as a nanomachine) or results happened that we never really understood and so could not optimize (e.g., stained glass window—small metal nanoparticle comparable in size with the wavelength of light).

Although there is some overlap, not all chemical forming techniques involve minute amounts of reacting chemicals, nor do they all entail self-assembly or positional assembly. The chemical forming methods covered in this chapter are sol-gel, organic film spin coating, polymer dry film lamination, polymer dip coating, polymer spraying, polymer casting, organic film doctor's blade (or knife coating), glow discharge polymerization, low temperature silk screening of organics, Langmuir-Blodgett deposition, and self-assembled monolayers (SAMs).

Arraying of organics for sensor arrays may be accomplished by lithography, digital mirror deposition, ink-jetting, mechanical microspotting, and microcontact printing.

Sol-Gel Technique

Introduction

Colloids, we saw in Volume I, Chapter 7, are mixtures in which one material is dispersed in another, with the dispersed material divided into particles ranging in size from 1 nm to 10 μm. They have many properties distinct from those of solutions of ordinary-sized molecules and from mixtures of macroscopic phases such as slurries of visible particles. The dispersed material may either be dissolved (i.e., in the case of solutions of macromolecules), giving a lyophilic colloid, or they may constitute a separate phase, giving a lyophobic colloid. We only deal with lyophobic colloids here,

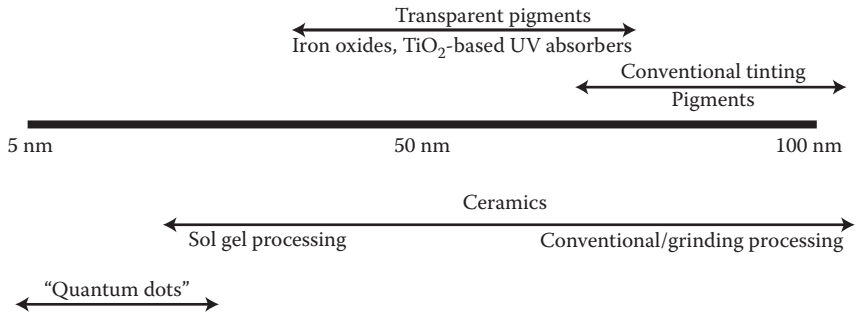


FIGURE 8.1 Nanoparticle products and their manufacture.

which may be further classified according to the types of phases that compose them, as shown in Volume I, Chapter 7. Some important definitions pertaining to the sol-gel field follow and might help in the understanding of the sol-gel process.

- Colloid: A colloid is a suspension in which the dispersed phase is so small (1 nm–10 μm) that gravitational force is negligible, and interactions are dominated by short-range forces, such as van der Waals attraction and surface charges. The suspended particles usually exhibit Brownian motion.
- Sol: A sol is a colloidal suspension of solid particle or clusters in a liquid.
- Gel: Semirigid mass of a lyophilic sol in which all the dispersion medium is penetrated into the sol particles.
- Emulsion: An emulsion is a suspension of liquid droplets in another liquid.
- Aerosol: An aerosol is a collection of very small particles suspended in a gas. The particles can be liquid (mist) or solid (dust or fume). The term *aerosol* is also commonly used for a pressurized container (aerosol can) designed to release a fine spray of a material such as paint.
- Precursors: The starting compound of the sol-gel process.
- Alkoxide: A basic salt derived from an alcohol by the replacement of the hydroxyl hydrogen with a metal. They are formed by the reaction of an alcohol and an alkali metal. For example, silicon tetraethoxide (TEOS), $Si(OC_2H_5)_4$.
- Hydrolysis: Reaction in which a hydroxyl ion is attached to the metal atom [$Si(OR)_4 + H_2O \rightarrow HO–Si(OR)_3 + ROH$, where R is an alkyl].
- Condensation reaction: Reagents combine splitting off water [$HO–Si(OR)_3 + HO–Si(OR)_3 \rightarrow (OR)_3\ SiO\text{-}Si(OR)_3 + H_2O$].

The sol-gel process is a wet-chemical additive technique for the fabrication of materials (typically metal oxides) starting from a chemical solution that reacts to produce colloidal particles (a sol). The Derjaguin, Landau, Verwey, and Overbeek theory (the DLVO theory) was introduced in Volume I, Chapter 7 to describe the stability of these colloids.

Sol-gel processing received increasing attention since the 1990s from both the research community and industry as nanoparticles became more important. Colloids form a most important gateway toward bottom-up nanotechnology: the sol-gel technique constitutes a chemical forming process for making very small particles, e.g., 20–40-nm diameter, a particle size that is virtually impossible to make by conventional mechanical ball milling or grinding* (Figure 8.1).

Sol-Gel Process

The first metal alkoxides were prepared in 1846 by Ebelmen using $SiCl_4$. Ebelmen noticed that on exposure to the atmosphere, these chemicals formed a gel. This process was then used in the 1930s by the Schott glass company in Germany to produce metal oxides (http://www.solgelchemist.com/web1.doc). In a sol-gel process, as illustrated in Figure 8.2, chemical precursors, mainly alkoxides (organic epoxides are an alternative for alkoxides) and metal chlorides, are dissolved in an aqueous or alcohol-based solution. Hydrolysis and condensation reactions nucleate

* High-energy ball milling is required to make particles less than 50 nm (see Volume III, Chapter 3).

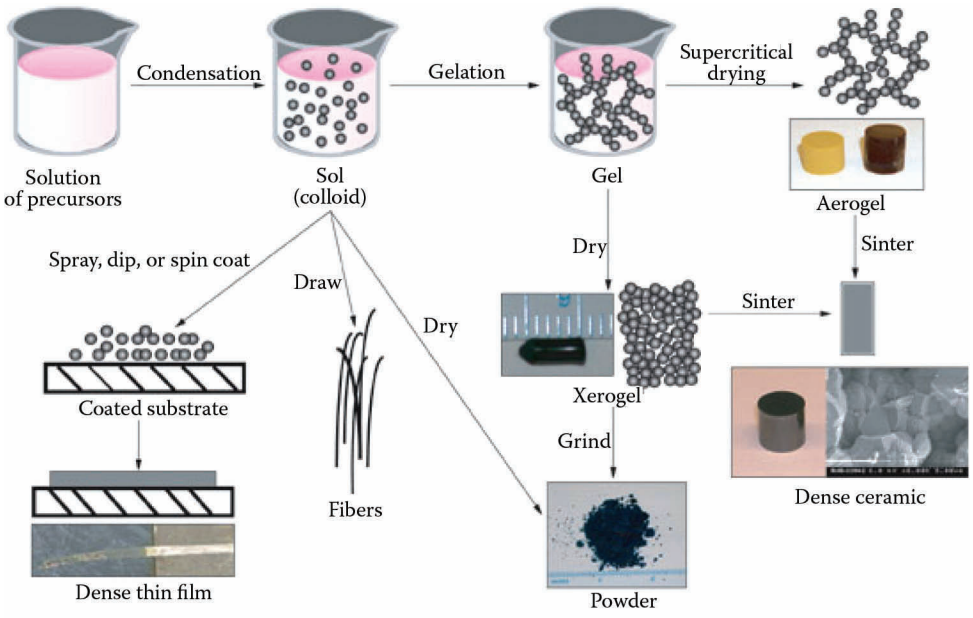


FIGURE 8.2 Sol-gel processes and sol-gel products. In sol-gel chemistry, molecular precursors are converted to nanometer-sized particles, to form a colloidal suspension, or sol. The gel can be processed by various drying methods (shown by the arrows) to develop materials with distinct properties.

and generate many small nucleation seeds (groups of atoms that build the basic crystal structure). The seeds grow by taking more atoms/molecules from the solution to form clusters. The growth pattern is random, with the atoms or molecules coming from every direction, and fractal-like cluster structures result. The clusters are nanosized, forming a colloid. As we saw in Volume I, Chapter 7, the stability of this colloid depends on the electrokinetic potential ζ (also the zeta potential), from which all electrokinetic effects can be derived. When the absolute value of the zeta potential is more than 50 mV, the dispersion is very stable, whereas with a ζ potential close to zero, the coagulation of particles leads to precipitation.

When growth continues, the nanoclusters start to build up a porous network. This is when gelation takes place. The gel point is the point in the phase diagram where the sol abruptly changes from a viscous liquid to a gelatinous, polymerized network and where the dispersion medium is fully penetrated into the sol network (Figure 8.3). Both sol and gel formation are low-temperature processes.

The transformation to a gel can be initiated in several ways, but the most convenient approach is to change the pH of the reaction solution. Formation of a metal oxide involves connecting the metal centers with oxo (M-O-M) or hydroxo (M-OH-M) bridges, therefore generating metal-oxo or metal-hydroxo polymers in solution. The drying process serves to remove the liquid phase from the gel, thus forming a porous material; then a thermal treatment (firing) may be performed to favor further polycondensation and enhance mechanical properties. The method used to remove the solvent from the gel strongly affects the end product's properties. For example, from Figure 8.2 it can be seen that if the solvent in a wet "gel" is removed under a supercritical condition, the gel's original 3D structure is preserved, and a highly porous and extremely low-density material called an *aerogel* is obtained. The supercritical processing condition is realized by exchanging the liquid solvent with CO_2. The latter is then removed at slightly elevated temperatures and under high pressure. This way aerogels can be

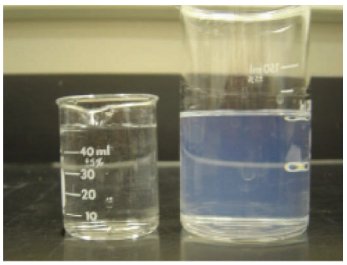


FIGURE 8.3 Gelation: TiO_2 from sol (left) to wet gel (right).

FIGURE 8.4 Aerogel. This photo illustrates the excellent insulating properties of aerogel. The crayons on top of the aerogel are protected from the flame underneath and are not melting.

produced where 99% of the bulk material consists of pores (Figure 8.4). A Lawrence Livermore group created an aerogel weighing only 1.0 mg/cm^3, which is listed in the *Guinness Book of World Records 2005* as the lightest material on Earth.[1]

If, in an alternative process route, the gel is dried slowly in a fluid-evaporation process, the gel's structural network collapses, which creates a high-density material known as a *xerogel*. When a xerogel is further densified by sintering, a dense ceramic material forms.

As the viscosity of a sol is adjusted into a proper viscosity range, ceramic fibers can be drawn; it can be applied as a film onto a substrate by spinning, dipping, or spraying and cast into a suitable container with the desired shape (e.g., to obtain a monolithic ceramics, glasses, fibers, membranes, and aerogels); or finally, ultrafine and uniform ceramic powders may be formed by grinding of the xerogel or by slowly drying the sol.

For a silica sol formation, an appropriate chemical precursor is dissolved in a liquid, for example, tetraethylsiloxane (TEOS) in water. In this case, a silica gel is formed by hydrolysis and condensation using hydrochloric acid as the catalyst. Drying and sintering at temperatures between 200 and 600°C transforms the gel into a glass and then densification into silicon dioxide.

Alkoxides can be very reactive and are commercially available for only a select number of elements. Livermore chemists have found that organic epoxides also initiate the sol-gel reaction. With this approach, precursors that are more widely available can be used, thus increasing the number of potential materials that can be developed. In addition, the starting materials, solvents, and gelling agents used with epoxides are less expensive than those used with alkoxides.

Summarizing, sol-gel is a cheap and low-temperature technique that enables mixing of precursors at the molecular level, leading to finer control and higher purity than most other ceramic synthesis methods (even small quantities of dopants, such as organic dyes and rare earth metals, can be introduced in the sol and end up finely dispersed in the final product). The method is particularly suited for the production of nanosized multicomponent ceramic powders densified through thermal annealing (sintering at relatively low temperature) to an inorganic product like a glass or polycrystal or a dry gel. Applying the sol-gel process, it is possible to fabricate ceramic or glass materials in a wide variety of forms: ultrafine or spherical-shaped powders, thin film coatings, ceramic fibers, microporous inorganic membranes, monolithic ceramics and glasses, or extremely porous aerogel materials. Different material compositions with different porosity may be fabricated depending on the process parameters (such as temperature) and the chemicals used. Its main traditional use is for optical coatings, where the finer particles give better optical clarity, and for the manufacture of fine ceramic fibers.

Sol-gel-derived materials have diverse applications in optics, electronics, energy, space, (bio)sensors, medicine (e.g., controlled drug release), and separation (e.g., chromatography) technology.

Sol-gel methods have been used in MEMS, for example, in the fabrication of piezoelectrics such as lead-zirconate-titanate (see Volume III, Chapter 8). Also, a commercially available, room temperature chemical gas sensor (a CO fire alarm) is fabricated with the sol-gel process (available from Quantum Group Inc.; http://www.qginc.com). The Sol-Gel Gateway (http://www.solgel.com/bookstore/bookstore.htm) is an excellent take-off point for further study of the sol-gel technique on the Internet (in particular, check http://www.solgel.com/Tutorials/tutorials.htm).

Deposition and Arraying Methods of Organic Layers in Bio-MEMS

Introduction

Today, miniaturization methods are often applied to biotechnology and biomedical problems. With the growth of the "bio-MEMS" field, techniques for depositing organic materials for chemical and biological sensors, often arranged in some type of an array configuration, are gaining importance. Organic gas-permeable membranes, ion selective membranes, hydrogels, and organic monolayers are needed for the manufacture of room temperature gas sensors, ion-selective electrodes (ISEs), enzyme-based sensors, immunosensors, and DNA and protein arrays. Their deposition presents challenges not typically encountered in the integrated circuit (IC) world. Membranes may be based on classical polymers, such as PVC (polyvinylchloride), PVA (polyvinylalcohol), PHEMA (polyhydroxyethylmethacrylate), or silicone rubber, and incorporate pH and temperature-sensitive biological materials such as enzymes, antigens, and antibodies. Single layers of molecules may be deposited by Langmuir-Blodgett (LB) deposition techniques or as self-assembled monolayers (SAMs). Those LB films or SAMs may then be further used as anchor points for proteins or DNA probes, and the latter natural polymer molecules may be synthesized in situ. Not only are the materials dealt with very different, the fact that most of the bio-MEMS sensors are disposable requires new manufacturing approaches as well. Although the number of new tools available for microfabrication has grown dramatically, few methods have been gainfully applied to disposable biosensor construction for clinical applications. The low cost requirements and the tremendous variety and fragmentation of biomedical sensor applications often necessitate a nonsilicon, modular approach. The Si approach is mostly based on thin film techniques with a limited number of materials, too expensive cost, all steps tightly integrated, and too many requirements on process and materials compatibility. Having addressed materials choice and modularity, various competing manufacturing processes must be compared.

Because of the importance of this emerging field, we review the different options available to coat and pattern organic materials on various substrates for bio-MEMS in the following section. For the organic film deposition techniques reviewed earlier in this book, we provide a short summary only.

Deposition Methods for Organic Materials

Organic Film Spin Coating As we know from Chapter 1 on lithography, spin coating technology has been optimized for the deposition of thin layers of photoresist, about 1–2 μm thick, on round and nearly ideally flat Si wafers. A typical spin coater schematic is shown in Figure 1.14. Resists are applied by dropping the resist solution, a polymer, a sensitizer (for two-component resists), and a solvent on the wafer. The wafer is then rotated on a spinning wheel at high speed so that centrifugal forces push the excess solution over the edge of the wafer, and a residue on the wafer remains as a result of surface tension. In this way, films down to 0.1 μm can be made. An empirical expression relating film thickness to solution viscosity and rotation speed was given in Chapter 1 (Equation 1.3). However, biosensor substrates rarely are round or flat, and many chemical sensor membranes require a thickness considerably thicker than 1 μm for proper functioning. For example, a typical ion-selective electrode (ISE) membrane is 50 μm thick (see Volume I, Chapter 7). Consequently, spin coating technology does not necessarily fit in with thick chemical membranes on a variety of substrates, and for monolayer deposition the method is not adequate. For building ISEs, dip coating or membrane casting is preferred, and for monolayer deposition one chooses Langmuir-Blodgett or self-assembled monolayer (SAM) methods (see below). For a tutorial on spin coating, visit http://www.mse.arizona.edu/faculty/birnie/Coatings/index.htm.

In Chapter 1, we compared several alternatives to spin coating for depositing photoresist (see Figure 1.15). Obviously all these techniques can be adapted to deposit other organic coatings. This includes spray coating, curtain coating, extrusion coating, roller coating, dip coating, plasma deposited, electrodeposited, silk screening, and meniscus coating. We provide some additional information here complementing the material presented in Chapter 1.

Polymer Dip Coating In the dip-coating method a substrate is slowly dipped into and withdrawn from a tank containing the dissolved polymer, with a uniform velocity, to obtain a uniform coating. The dip-coating process may be divided in five stages: immersion, startup, deposition, evaporation, and drainage. Dip coating of a substrate in a solution of a dissolved polymer typifies the simplest method to apply an organic layer to a substrate. It is especially suited for wire-type ISEs and enzyme-based biosensors where the membrane forms a droplet at the end of a wire (see Volume I, Chapter 7). A substrate, for example, a chloridized silver wire, is dipped into a solution containing the polymer and a solvent. After evaporation of the solvent, a thin membrane forms on the surface of the sensor. To obtain pinhole-free membranes, the dipping is repeated several times, interspersed with drying periods. Even though the eventual goal is more typically the production of a planar sensor structure, a Ag wire may be applied in the research phase to quickly evaluate a new membrane composition. The method is difficult to commercialize because of the variability in coating thickness and uniformity. For further reading on dip coating, visit http://www.solgel.com/articles/Nov00/mennig.htm.

Dipping and spinning are also used in the fabrication of sol-gel layers (see Figure 8.2).

Polymer Dry Film Lamination The use of dry film lamination, such as the dry photoresist films explored in Chapter 1, is introduced as a better approach for the mass manufacture of chemical and biological sensors than spin coating in Volume III, Example 1.1. Most resists in IC and MEMS fabrication are deposited as liquids, whereas resists used in printed wiring board manufacture are usually dry film resists that come in rolls (ranging from 2–60 in. wide and 125–1000 ft. long) and are laminated onto the substrate instead of being spin-coated on it. A typical dry film lamination setup is shown in Figure 1.33. The heat and pressure of the laminating rollers cause the dry film to soften and adapt to surface topologies. The resist is then exposed to a UV light source. The lamination approach can be extended to other polymer coatings as exemplified in the manufacture of the compact disc (CD)-based fluidic platform shown in Volume III, Chapter 5. In the manufacture of this CD, we use a layering approach, with different sheets of thin machined plastic CDs (polycarbonate-PC) laminated together using pressure-sensitive adhesive layers (PSA-FLEX mount DFM 200 clear V-95). The PSA layers are as thin as 100 μm and are cut using a computer numerical controlled (CNC) cutter-plotter (in Volume III, Figure 5.27, CNC plotter, Graphtec CE-2000-60). The use of PSA layers allows for channels as narrow as 250 μm. When sandwiched between two plastic sheets, these channels in the PSA connect the larger CNC-machined reservoirs in the PC disks.

Polymer Spraying Polymer or plastic spray-coating techniques are the most widely used methods for applying organic coatings from the liquid, gas, or the solid phase. Spray coating of a liquid involves pressurization by compressed air or, in an airless method, pushing liquid mechanically through tiny orifices. Vapors are carried in an inert dry vapor carrier. In the case of a solid, a powdered plastic resin is melted and blown through a flame-shrouded nozzle. Such thermal spray coating involves heating a material, in powder or wire form, to a molten or semimolten state. The material is propelled using a stream of gas or compressed air to deposit it, creating a surface structure on a substrate. The coating material may consist of a single element but is often a composite with unique physical properties that are only achievable through the thermal spray process.

There are two main classes of powder coatings: thermosetting and thermoplastic coatings. In a thermosetting film, cross-linking occurs between the molecules in the powder during baking. This cross-linking turns the baked film into a single giant molecule that cannot melt or flow. In a thermoplastic film, thermal energy makes the binder molecules mobile enough to become entangled; thus, a continuous film forms, and this film hardens on cooling. Although a thermoplastic film can still melt or flow, it can do that only at elevated temperatures. In electrostatic spraying, a negatively charged plastic powder is spray gunned onto grounded conductive parts (see Figure 8.5).[2] Electrostatic spray is the primary technique used for thermoset powders. The particles of powder are given their electrical charge in the powder coating gun, and the target part is

FIGURE 8.5 Electrostatic spray gun from ITW Ransburg. (From http://www.itwransburg.com/products.html.)

attached to a fixture that is grounded. The electrically charged powder particles are attracted to the grounded part and build up on the surface until the part is covered with charged particles. At this point, the oncoming particles are repelled by the part, and the coating process stops. This provides for an even film thickness.

Thermoplastic and thermosetting coatings may also be formed by dipping a heated part into a container of resin particles set in motion by a stream of low-pressure air (fluidized-solid bed). This method is only of use if an inert coating is desired, for example, for biocompatibility, but not for heat-sensitive functionalized coatings. The fluidized bed is the original powder-coating technique. It is still the primary technique used for the application of thermoplastic powders.

Thermal spraying is ideally suited for large structures that otherwise could not be dipped in a polymer suspension. Unlike electrostatic powder coatings, nonconductive components can be coated, and unlike fluidized bed coatings, heat-sensitive materials (aircraft skins) can be sprayed. Polymer coatings can be repaired by heating (for remelting or curing) and by applying additional material to the desired location. Polymer coatings can be applied in high humidity and at temperatures below freezing. Certain polymers have excellent adhesion to metallic surfaces as a result of interfacial bonding. Metals, ceramics, or other polymers can be incorporated into the polymer matrix during spraying to act as a fill.

FIGURE 8.6 Yoda, typical polymer casting application.

Polymer Casting Cast polymer operations are operations where a gel coat resin is sprayed to a mold, after which a casting resin is applied without spraying. These are used in the manufacture of pools, bathtubs, marble-look countertops, masks, and toys (Figure 8.6). Casting in bio-MEMS is often based on the application of a given amount of dissolved organic material on the surface of a sensor substrate and letting the solvent evaporate. A rim structure is fashioned around the substrate, providing a "flat beaker" for the solution. This method provides a more uniform and a more reproducible membrane than dip coating. Membranes in planar ion-selective electrodes (ISEs) are often made this way. Casting is also often used to obtain thicker photoresist layers than typically possible with spin coating (see also Chapter 1). Casting is the main technology used in soft lithography (see Chapter 2).

Polymer Doctor's Blade or Knife Coating Doctor's blade or knife coating relies on a coating being applied to a substrate that then passes through a slit between a knife and a support roller (Figure 8.7). The "doctor's blade" refers to the scraping knife blade for the removal of excess substance from the moving surface being coated. The doctor's blade is adjustable with a precisely controlled height, limiting the dispensed material to a known thickness. This process can be used for high-viscosity coatings, and there are innumerable variants of this relatively simple process, which is rugged but somewhat inaccurate. It is well known in many industries, including paper,

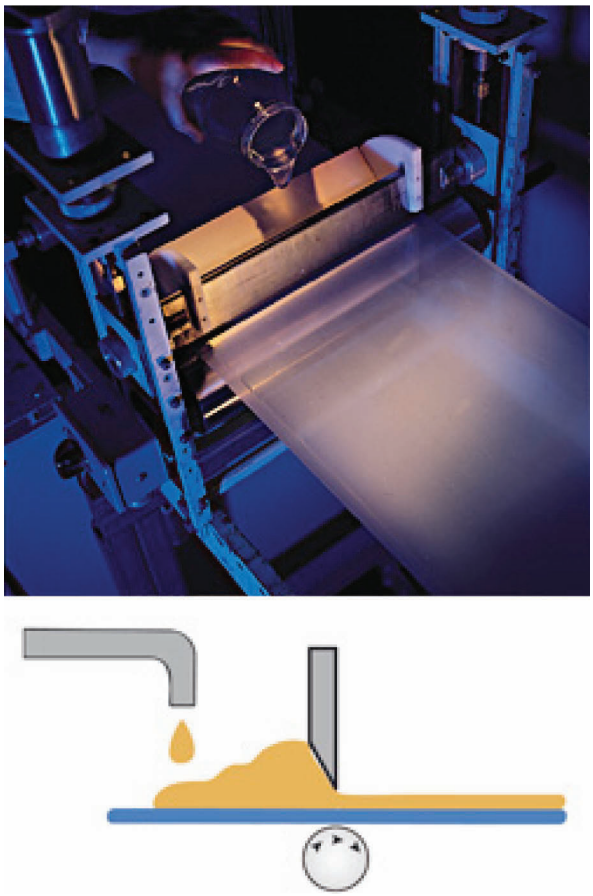


FIGURE 8.7 Doctor blading or knife coating of organic materials on a moving substrate.

plastic, and paint manufacturing. Doctor's blade technology is a continuous fabrication method, which can also be used to apply all types of organic cocktails, for example, to make enzyme-based biosensors or to deposit ceramics, in which case one also refers to tape casting and green tape technology. Howatt was the first to obtain a patent to use the technique for "forming ceramic materials into flat plates, especially useful in the electric and radio fields."[3] This is still the principal application of tape casting today, although its use extends far beyond the field envisioned in 1952 (http://www.drblade.com/history.htm). The ceramic application of doctor blading is a thermal forming process well suited for fabrication of mesoscale devices (layers with a thickness between several micrometers and millimeters) and is covered in Chapter 9 on thermal energy-based forming.

As a specific application of a doctor's blade process in bio-MEMS, consider the mass production of amperometric glucose sensors discussed in Volume III, Chapter 1. In that chapter, as an illustration of an informed choice of manufacturing options, we find out that using a silicon batch approach, it is almost impossible to make a glucose sensor for less than $1 (or any disposable biosensor for that matter!). The current industrial process, and our conclusion, is that for mass-producing glucose sensor strips, the optimal choice is doctor's blade on a continuous moving web, making a 10-cents cost per glucose sensor possible.

From this example of a glucose sensor strip, it appears that one of the major challenges in bio-MEMS is the fabrication of affordable disposable biosensors. In the case of a glucose sensor strip, this may be accomplished using doctor's blade in a continuous process. An example of a generic approach for biosensor manufacture on large plastic sheets and eventually on a moving web is shown in Volume III, Figure 1.19. The process illustrated allows for the fabrication of a sensor array composed of sensors that may otherwise have fabrication incompatibilities. We call this futuristic bio-MEMS approach "beyond batch."

Glow Discharge (Plasma) Polymerization Plasma polymerization is a type of plasma-enhanced chemical vapor deposition procedure (PECVD; see Chapter 7), in which gaseous precursor monomers, activated by the plasma, deposit on freely selectable substrates as highly cross-linked polymer layers. A polymerizable gas can, in principle, include any substance that can be brought into the gas phase at reduced pressure. The *sine qua non* for this process to take place is the presence in the gas phase of polymer-producing atoms, such as carbon, silicon, or sulfur. These react with one another and with any surface groups to form a polymer film, the chemical structure of which is dependent on the process conditions used. The monomer molecules in the plasma, for the most part, become fragmented into reactive segments, and remain only partially preserved in the chemical structure of the final product, which results in a highly cross-linked and disordered structure compared with a conventional polymer (Figure 8.8). The degree of structural preservation and cross-linking gradients in the deposit are controlled

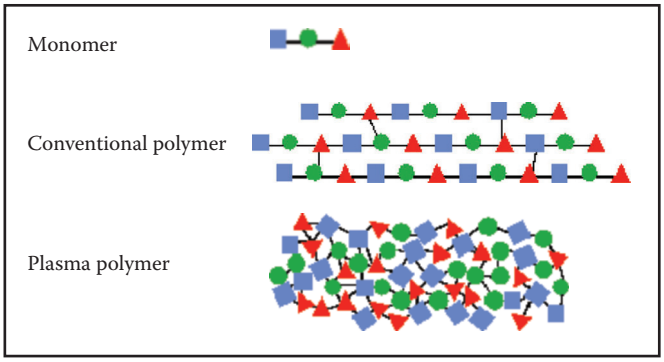


FIGURE 8.8 Polymers deposited with plasma polymerization. Comparison of the structures of plasma polymers and conventional polymers.

through process parameters, such as gas pressure, working gas flow, and the applied electrical field. For example, it has been found that pulsing the plasma in the millisecond regime constitutes a very simple and unique method to control the film chemistry.[4] Unlike the more conventional continuous wave plasma deposition processes, pulsed plasma deposition enables the polymerization of monomers containing labile groups such as amines, ethers, and anhydrides, to name but a few. By selecting the monomer type and the energy density per monomer, known as the *Yasuda parameter*, the chemical composition and structure of the resulting thin film can be varied over a wide range.[5]

The so-obtained polymer coats can be used to further attach specific molecules by conventional wet chemistry techniques. One can also construct so-called *gradient layers*, i.e., with increasing degree of cross-linking over the film thickness.

Figure 8.9 illustrates the setup we used in our own work on plasma polymerization of doped polymers. The apparatus consisted of a quartz tube reaction chamber; a gas-handling system to introduce the carrier gas, monomer, and dopant into the system, as well as to remove the unreacted material; and a power supply (operating at 27 MHz) to provide the radio-frequency (RF) energy necessary to create and maintain a plasma within the system. Despite the complex chemistry of this process, good conformal coatings often result. Guckel applied this technique to obtain deposition of poly(methylmethacrylate) (PMMA) in layers more than 100 µm thick.[6] In research at SRI International, we synthesized electroactive plasma polymers using I_2 or N_2O as dopants from the following monomers: thiophene, furan, aniline, benzaldehyde, benzene, indole, diphenylacetylene, and 1-methylpyrrole.

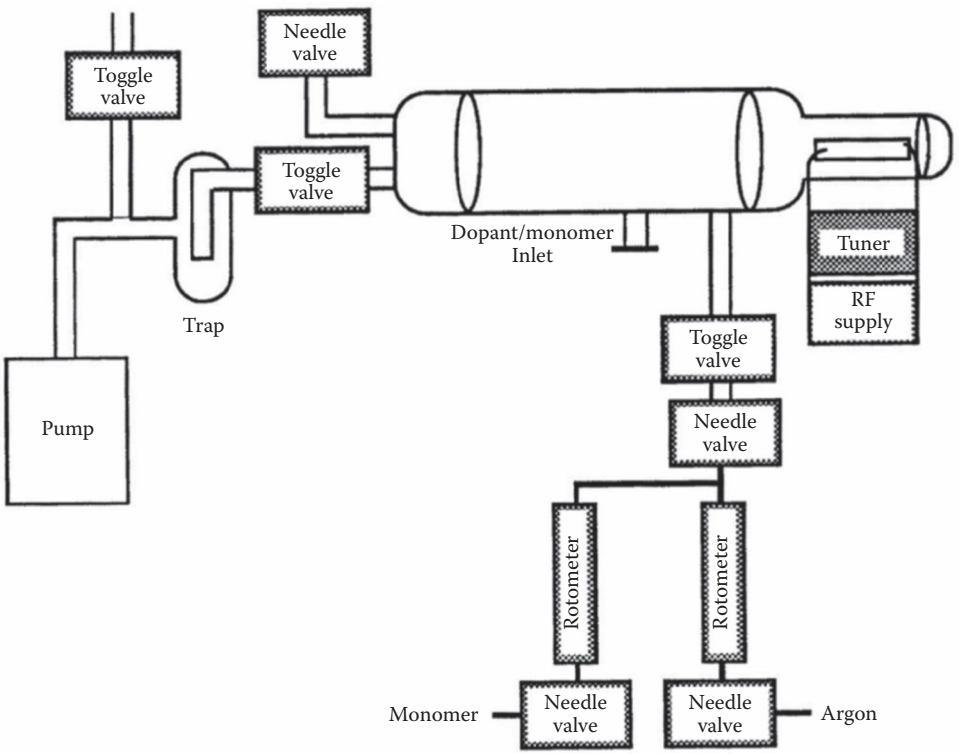


FIGURE 8.9 Schematic of plasma deposition system.

In general, plasma-polymerized materials offer the following advantages:

- The plasma-polymerized films are uniform, pinhole-free, chemically resistant, and mechanically strong.
- A thin to thick film (200 Å to >100 μm) can be formed in a flawless manner at ambient temperature onto any substrate.
- The organic films deposited by the plasma process adhere strongly to the substrate.
- The choice of monomers is unlimited; almost any organic compound convertible into vapor can be polymerized.
- The plasma process (one-step process from a vapor source) is compatible with conventional CMOS technology.
- Some functional groups can be introduced onto the surface of the organic film by subsequent glow discharge treatment in a reactive gas atmosphere.
- Highly irregular surfaces can be coated and patterned by depositing a light-sensitive polymer by plasma polymerization (e.g., polymerized PMMA).

Applications from plasma polymer coatings include scratch-resistant coatings, corrosion protection, antibonding, antisoiling coatings, barrier layers, and biocompatible coats for implants. For more details on plasma polymerization, see, for example, Yasuda[7] and *Plasma Polymer Films* edited by Biederman.[8]

Silk Screening or Screen Printing of Organics Silk screening or screen printing (also serigraphy) is an additive technique that can be used to create sharp-edged organic coatings/images using a stencil. The method is reviewed in more detail in Chapter 9, where we discuss its application for the deposition of higher temperate ceramic coats. Silk screening is popular both in fine arts and in commercial printing, where it is used to print images on T-shirts (Figure 8.10), hats, CDs, DVDs, ceramics, glass, polyethylene, polypropylene, paper, metals, and wood.

Screen printing presents a more cost-effective means of depositing a wide variety of films on planar substrates than does integrated circuit (IC) technology, especially when fabricating devices at relatively low production volumes. The technique constitutes one of several thick film or hybrid technologies used for selective coating of flat surfaces (e.g., a ceramic substrate). The technology was originally developed for the production of miniature, robust, and, above all, cheap electronic circuits. The up-front investment in a thick film facility is low compared with that of IC manufacturing. For disposable chemical sensors, recent industrial experience indicates that screen printing thick films is a viable alternative to Si thin film technologies.

FIGURE 8.10 Silk screening of T-shirts.

For biosensor applications, thick film technology based on pastes that can be deposited at room temperature is crucial. Special grades of polymer-based pastes (e.g., for carbon, Ag, and Ag/AgCl electrodes) are becoming commercially available for this purpose.[9] Polymer thick films with a thickness anywhere from 5–50 μm can be screen printed on cheap polymer substrates. In the commercial planar electrochemical glucose sensor from MediSense/Abbott sensor, the Precision QID, shown in Volume I, Chapter 7, all films are silk-screened onto a PVC substrate.

In the research phase of new chemical sensors, pastes must be developed from their pure components. Some examples follow. Pace et al.[10] screen printed a PVC/ionophore layer for a pH sensor. Belford et al.[11] investigated pH-sensitive glass mixtures and proceeded to screen print them on a multilayer metal conductor to make planar pH sensors. In Pace et al.,[12] a thick film, multilayered oxygen sensor with screen-printed chemical mem-

branes of PVA and silicone rubber is detailed (see also Karagounis et al.[13]). A thick film glucose sensor is presented by Lewandowski et al.[14] and Lambrechts et al.[15] The latter authors developed an enzyme-based thick film glucose sensor with RuO_2 electrodes. All the above thick film sensors were fabricated on Al_2O_3 substrates. Weetall et al.[16] present an extremely low-cost, silk-screened immunosensor on cardboard. Cha et al.[17] compare the performance of thick film Au and Pt electrodes with conventional bulk electrodes.

When faced with adapting a biosensor membrane to an IC process versus adaptation to a thick film process, it now seems that, once the specialty inks are available, a thick film approach is easier, less expensive, and more accessible. For in vitro chemical sensors where small size is not as important, we expect more research to result in a switch from the overly ambitious IC approach to the more realistic thick film approaches as sketched above. In Volume III, Chapter 1, we compare the pros and cons of thin and thick film technologies.

Important other driving forces for the development of new low-temperature inks for silk screening come from the exciting market opportunity of organic light-emitting diodes (OLEDs), OLED flexible displays, polymer transistors, and electronic paper. There are two types of OLED displays: the ones based on small molecule OLEDs and the polymer-based ones. The small molecule OLEDs use organic emissive materials that do not contain long polymer chains. "Small" molecules often used include organometallic chelates, flourescent and phosphorescent dyes, and conjugated dendrimers. The production of small molecule devices and dispalys usually involves thin film deposition such as a separation in a vacuum. Small molecule OLEDs are more mature, but the polymer based OLEDs have the potential of being lower cost because of the applicability of screen printing, which enables high throughput of large area devices, such as large area displays. Lifetime improvement and encapsulation are progressing fast; thus, polymer OLEDs now have the potential of being very bright, extremely thin, low power, and low cost. There is the eventual possibility of producing OLED displays using roll to roll (R2R) manufacturing (see Volume III, Chapter 1), which would dramatically further reduce their manufacturing cost. The aim is also to produce organic thin film transistors (TFTs) using printing techniques, but there are a number of issues that still need to be resolved in this case. However, reduction in cost can already be achieved today by using hybrid-processing technologies, i.e., using low-cost printing and dispensing techniques wherever it is possible, whereas other processing steps remain "conventional," such as photolithography or vapor deposition processes.

In Figure 8.11 some typical thin film organic devices are shown. Using thin sheets of plastic—similar to overhead transparencies—as the base, one can print the multiple layers of OLEDs or transistors with silk screening one layer at a time. The squeegee pushes a liquid plastic mixture over a stainless-steel mesh, and after the solvent evaporates, a new plastic feature remains.

With electronic paper, the ink is a liquid that can be printed onto nearly any surface. Within the liquid are suspended tiny microcapsules, each containing white and black particles (Figure 8.12). The relative movement of negatively charged black and positively charged white particles inside their microcapsules is controlled by the direction of an applied voltage. In an electric field, the white particles move to one end of the microcapsule and make the surface of the electronic paper appear white at that spot. An opposite electric field pulls the white particles to the other end of the microcapsules, where they are obscured by the black particles, making the surface appear dark at that spot.[18] To form an electronic ink display, the ink is screen printed onto a sheet of plastic film that is laminated to a layer of circuitry. The circuitry forms a pattern of pixels that can then be controlled by a display driver. The microcapsules are suspended in a liquid "carrier medium," allowing them to be printed using existing screen-printing processes onto virtually any surface, including glass, plastic, fabric, and even paper. Ultimately electronic ink will permit most any surface to become a display, bringing information out of the confines of traditional devices and into the world around us.

Applications of electronic ink include e-book readers capable of displaying digital versions of books and e-paper magazines, electronic pricing labels in retail shops, timetables at bus stations, and

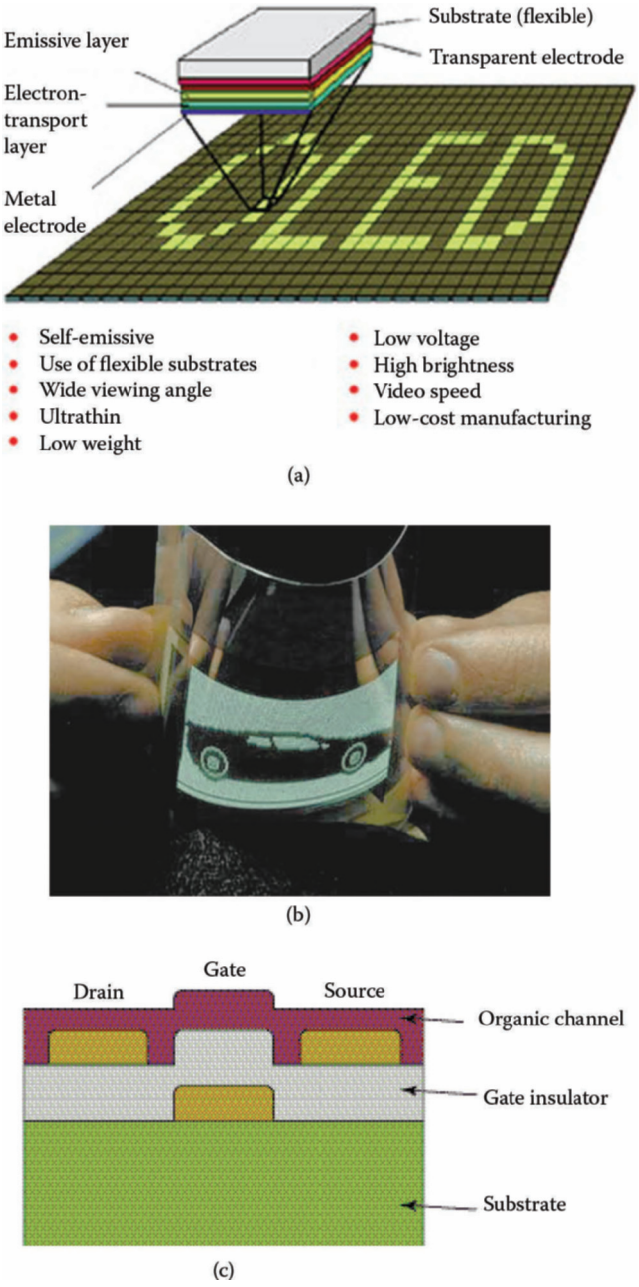


FIGURE 8.11 Organic light-emitting diode (OLED). (a) In a typical organic light-emitting device, luminescent molecular excited states are generated in the electron transport and luminescent layer near its contact region with the hole transport layer. (b) OLED display; an organic passive matrix display on a substrate of polyethylene terephthalate, a lightweight plastic, will bend around a diameter of less than 1 cm. The 1.8-mm-thick, 5 × 10-cm monochrome display consists of 128 × 64 pixels, each measuring 400 × 500 μm, and is operated at conventional video brightness of 100 cd/m^2. It was fabricated by Universal Display Corp. (Ewing, NJ), with a moisture barrier built into the plastic that prevents degradation of the pixels (http://www.universaldisplay.com). (c) Organic thin-film transistors (OTFTs) are constructed of an organic or inorganic gate insulator and an organic semiconducting channel linking the source and drain.

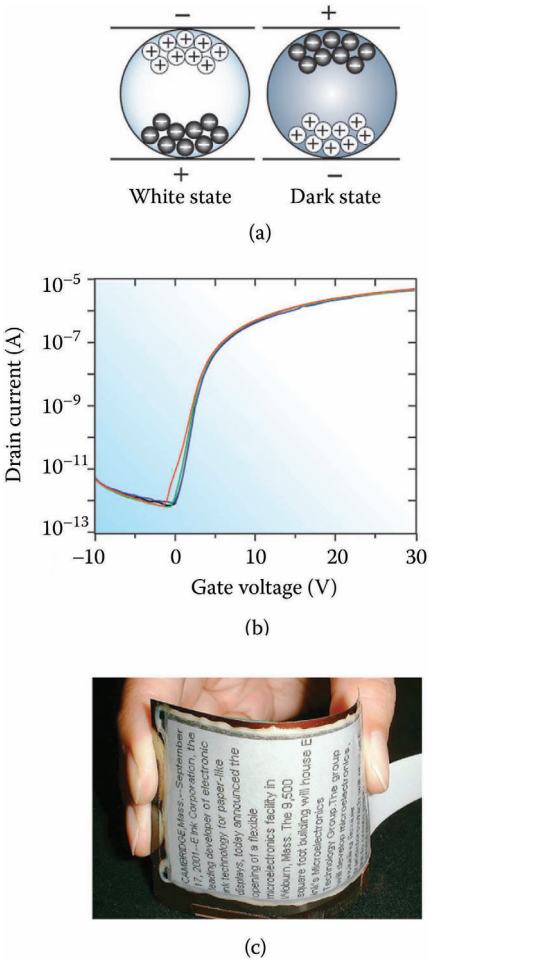


FIGURE 8.12 How electronic ink works. (a) Operating principle of electronic ink. The relative movement of negatively charged black and positively charged white particles inside their microcapsules is controlled by the direction of the applied voltage. (b) A backplane thin-film transistor measured in situ under compressive stress. The transistor is bent to three different radii of curvature: green, 2.0 cm (0.19% strain); blue, 1.3 cm (0.29% strain); and red, 1.0 cm (0.38% strain). The thin-film transistor has identical characteristics when measured without bending (black curve) and at a radius of curvature of 2.0 cm; degradation is minimal even at 1.0 cm. Results were similar under tensile stress. (c) Text image shown on a bent display whose resolution is 96 dpi and which has a white-state reflectance of 43% and a contrast ratio of 8.5:1. (From Chen, Y., J. Au, P. Kazlas, A. Ritenour, H. Gates, and M. McCreary. 2003. Electronic paper: Flexible active-matrix electronic ink display. *Nature* 423:136.[18])

electronic billboards. In February 2006, the Flemish daily *De Tijd* distributed an electronic version of the paper to select subscribers in a limited marketing study. This was the first recorded application of electronic ink to newspaper publishing.

Langmuir-Blodgett Deposition and Self-Assembled Monolayers The Langmuir-Blodgett technique, invented by Irving Langmuir and Katharine Blodgett in 1935, allows the controlled deposition of monomolecular layers on a wide variety of substrates. This ultrathin film deposition technique is limited to materials that consist of amphilic long chain molecules with a hydrophobic molecule at one end and a hydrophilic molecule at the other. In the Langmuir-Blodgett process, a monolayer of film-forming molecules (stearic acid is a model molecule) on an aqueous surface is compressed into a compact floating film and transferred to a solid substrate by passing a substrate through the water surface at a constant speed and film surface tension (Figure 8.13). Thus, layered films can be built up in thickness (up to 100 layers) by consecutive dipping in the Langmuir trough. For example, phthalocyanine thin films sensitive to oxidizing gases such as NO_2 and biological materials sensitive to odors resulted via this method. Most of the difficulties with Langmuir-Blodgett films stem from the need to make the material pinhole-free and to overcome the problem of their lack of mechanical, chemical, and thermal stability.

In the early 1980s at the Bell Labs, David L. Allara (now at Pennsylvania State University) and Ralph G. Nuzzo (now at University of Illinois, Urbana-Champaign) discovered the self-assembly of disulfide and, soon thereafter, of alkanethiol monolayers (SAMs) on metal surfaces.[19,20] SAM films, especially on Au, spontaneously assemble into stable and highly organized molecular layers, bonding with the sulfur atoms onto the gold and resulting in a new surface with properties determined by the alkane head group. The preferred crystal face for alkanethiolate SAM preparation on gold substrates is the (111) surface, which is obtained either by using single-crystal substrates or by evaporation of thin

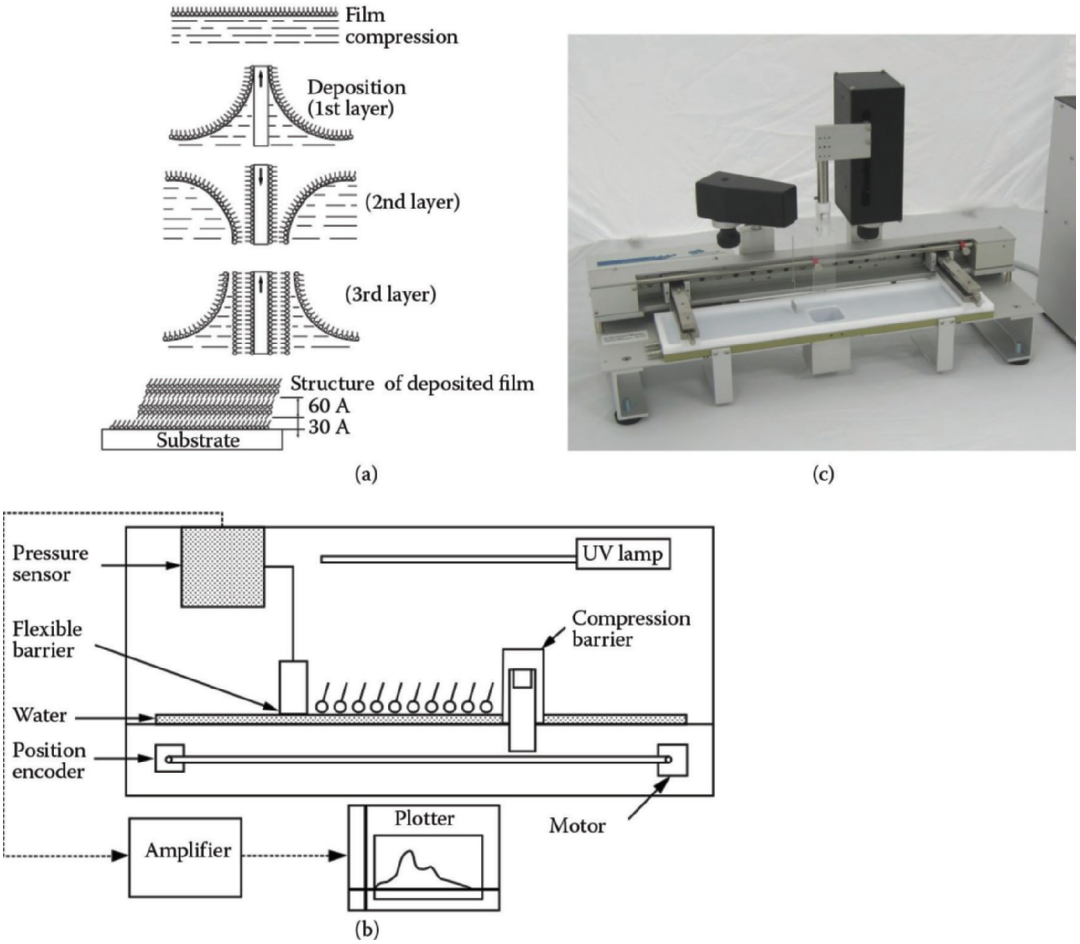


FIGURE 8.13 Langmuir-Blodgett film deposition: sequence of a deposition (a), schematic of apparatus (b), and photo of apparatus (c).

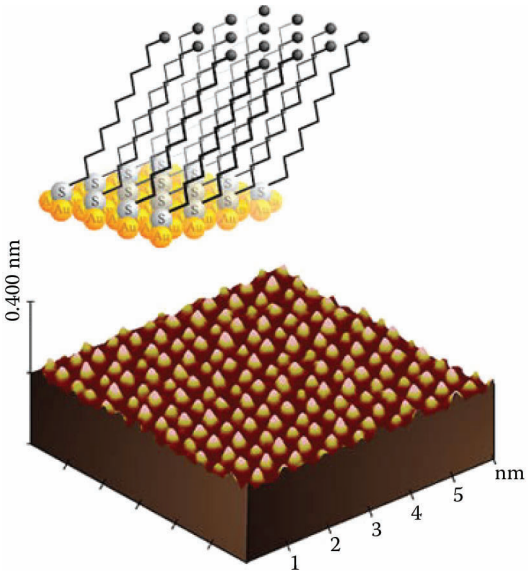


FIGURE 8.14 Preparation of SAMs. The substrate, Au on Si, is immersed into an ethanol solution of the desired thiol. Initial adsorption is fast (seconds); then an organization phase follows, which should be allowed to continue for >15 h for best results. An STM of a fully assembled SAM is also shown.

Au films on flat supports, typically glass or silicon (Figure 8.14). Several different solvents are usable at the low thiol concentrations (typically 1–2 mM) that are used in preparation of SAMs, but the most commonly used one is ethanol. It is important to minimize the water content in the solvent if the SAMs are to be used in ultrahigh vacuum (UHV) experiments. Drying the solvent limits incorporation of water into the SAM structure, which reduces outgassing and increases the repeatability of UHV experiments. Even though a SAM forms very rapidly on a Au substrate, it is necessary to use adsorption times of 15 h or more to obtain well-ordered, defect-free SAMs. Multilayers do not form, and adsorption times of 2–3 days are optimal in obtaining the highest-quality monolayers.

The tail group that provides the functionality of the SAM can be widely varied. CH_3-terminated SAMs are commercially available, and other functional groups can be easily synthesized by any well-equipped chemical laboratory, providing almost infinite possibilities. In addition, chemical modification of the tail group is also possible after formation of the SAM, expanding the available range of functionalities even further. Examples of functionalities are $-CH_3$, $-OH$, $-(C{=}O)OCH_3$, $-O(C{=}O)CH_3$, $-O(C{=}O)CF_3$, $-O(C{=}O)C_6H_5$, $-COOH$, $-OSO_3H$, and so on. SAMs on gold are generally stable and can withstand strong acids and bases; they are not destroyed by solvents and can withstand physiological environments.[21]

This method of building monolayers forms an important alternative to the Langmuir-Blodgett method. The use of monolayer electron beam resists, as discussed in Chapter 2, affords nanometer-scale lithography resolution. The chemical sensor industry exploits monolayers of organics, for example, in immunosensors and to provide anchor points for subsequent organic molecules or membranes.

Patterning of Organic Materials

Introduction

In the previous sections, we summarized some of the different methods available to deposit organic thin layers. In what follows, we will encounter some of the same techniques and introduce some new ones as we look into the process of depositing small amounts of material in a well-defined, small spot on a substrate, i.e., the patterning of organic films. Patterning of organic thin layers into discrete array elements has recently spawned a wide variety of applications. These range from sensor arrays for electronic noses and tongues to DNA and protein arrays.

Patterning through Photolithography

Patterning of Hydrogels, Gas-Permeable Membranes and Ion-Selective Electrodes Spinning, UV exposure, and development of resists are well-known, low-cost, mass-production patterning procedures in photolithography for IC fabrication. Naturally, this approach became one of the first methods to be applied to the patterning of other organic materials. However, photosensitized organic materials such as hydrogels, gas-permeable membranes, and ion-selective electrodes (ISEs), e.g., to fabricate a biosensor array, are usually not commercially available. To cope with this problem, photosensitive materials from high-purity materials must be prepared. For example, to pattern a hydrogel, e.g., a water-soluble polymer like polyvinyl alcohol (PVA), a photosensitizer such as $(NH_4)_2Cr_2O_7$ must be added.[15] After spin coating, the polymer film is

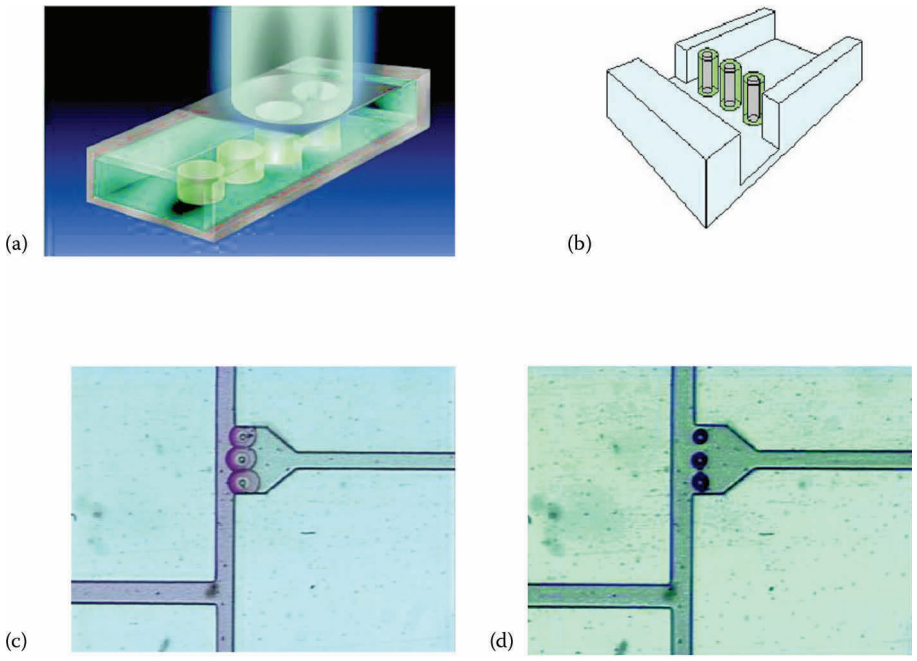


FIGURE 8.15 Patterning hydrogels. (a) Schematic showing the in-channel photo-polymerziation technique. (b) Schematic of a 2D shut-off microvalve consisting of hydrogel "jackets" (50 μm thick) around three prefabricated SU-8 posts. (c) Micrograph of the hydrogel jackets blocking the regulated channel (pH 7) in their expanded state in a pH 12 solution (dyed). (d) Micrograph showing the contracted hydrogels allowing fluid to flow down the side branch. (Courtesy of Dr. D. Beebe, University of Madison, Wisconsin.[22])

photochemically cross-linked by UV light. In the exposed regions, an insoluble hydrogel materializes. The development or removal of unexposed regions is carried out in warm water. An example of patterning a hydrogel actuator material inside a fluidic network through the use of photolithography is described by Liu et al.[22] The actuator hydrogel, after in situ photopolymerization, responds to changes in local pH by changing its volume and thus provides a means for valving action in the fluidic network.[23] The in situ polymerization of these polymeric valves and their function in a fluidic network are demonstrated in Figure 8.15.[22]

An alternative approach to patterning organic materials without modifying the material through the addition of a photosensitizer is through lift-off. Lift-off to lithographically pattern hard to etch materials (e.g., Pt), or in general to pattern materials that are not photosensitive, was introduced in Chapter 1. At one time, a problem with lift-off for patterning organic materials was the thickness required for the patterning photoresist layer. As discussed in Chapter 1, a typical resist layer used to measure only about 1 μm thick, whereas the thickness needed for the organic layers involved in chemical sensors frequently can reach up to 50–100 μm, demanding resist layers that are thicker. The many new thick resist technologies (e.g., SU-8) now available make this approach feasible today. However, a remaining problem with lift-off is that it can be used only with materials resistant to the solvent necessary for the resist removal; therefore, this technology probably will be limited to specific cases.

Hydrogels, gas-permeable membranes, and ISEs are all relatively thick, from several micrometers to 100 μm. Next we consider patterning of much thinner organic layers, such as monolayers of DNA and proteins.

Very Large-Scale Immobilized Polymer Patterning and Synthesis Affymetrix's *very large-scale immobilized polymer synthesis* (VLSIPS) is illustrated in Figure 8.16. The technique patterns and synthesizes biopolymers at the same time.[24] In light-directed oligonucleotide synthesis for DNA arrays (GeneChips), a solid support (e.g., a 1.28 cm × 1.28 cm Si chip) is derivatized with a covalently linked aminosilanated layer and terminated with a photolabile protecting

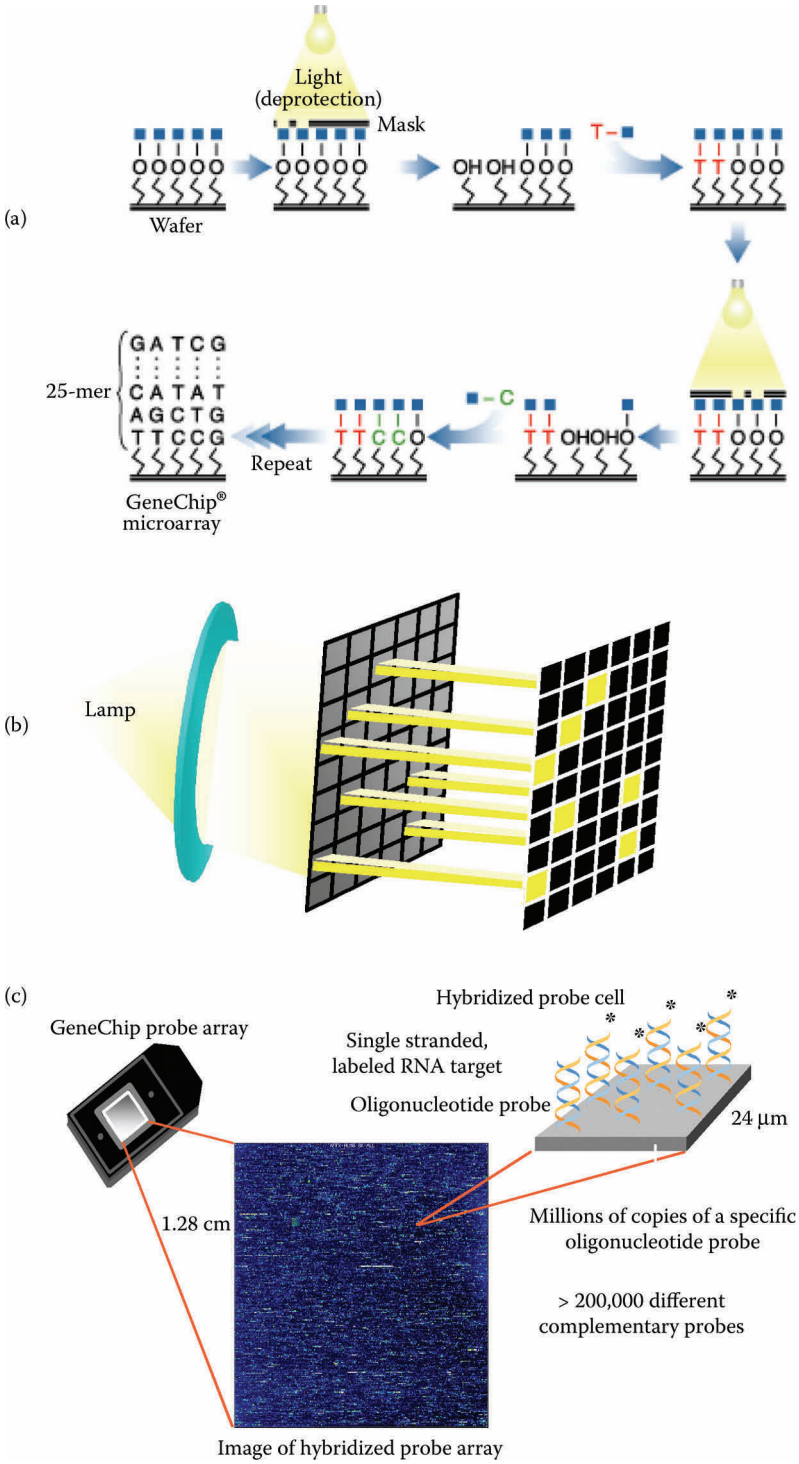


FIGURE 8.16 Very large-scale immobilized polymer synthesis (VLSIPS). Light-directed oligonucleotide synthesis. (a) A solid support is derivatized with a covalent linker molecule terminated with a photolabile protecting group. Light is directed through a mask to deprotect and activate selected sites, and protected nucleotides couple to the activated sites. The process is repeated, activating different sets of sites and coupling different bases, allowing arbitrary DNA probes to be constructed at each site. (b) Schematic representation of the lamp, mask, and array. (From Fodor, P. A. S., J. L. Read, M. C. Pirrung, L. Sryer, A. T. Lu, and D. Solas. 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251:767–73.[24]) (c) Finished GeneChip probe array with a fluorescent image of a hybridized probe array and details of an individual hybridized probe cell (24 µm).

group. The on-chip combinatorial synthesis proceeds by photolithographically deprotecting all array elements that are to receive a common nucleotide, coupling that nucleotide by exposing the entire array to the appropriate phosphoramidite, and then after the oxidation and washing steps, repeating the procedure for the next nucleotide. The in situ synthesis of oligonucleotides occurs in parallel, resulting in consecutive addition of A, C, G, and T nucleotides to the appropriate gene sequences on the array. To make an array of N-mers requires 4N cycles of deprotection and coupling, one for each of the four bases, times N base positions. This photolithographic method requires 4N masks, which are specific to the pattern of sequences on the array but can be used to make many copies of the array. The masks add considerable expense, and the procedure is best suited for generating large numbers of identical arrays. By going through 32 iterations of the oligonucleotide synthesis, 65,536 oligos containing eight units can be fabricated in about 1 day.

In Figure 8.17 we show how an Affymetrix oligo microarray can be used to establish the amount of messenger RNA (mRNA) in a sample. mRNA is extracted from a cell and is converted to complementary DNA (cDNA) using reverse transcriptase. After transcription, the cRNA is biotin-labeled and fragmented. This single-stranded and labeled RNA target is then put in contact with the GeneChip with millions of copies of different complementary probes. After hybridization, the chip is washed and stained, and the array is scanned and quantitated. Each gene is represented on the GeneChip using 16–20 (preferably nonoverlapping) 25-mers, and each oligonucleotide has a single-base mismatch partner for internal control of hybridization specificity.

The combinatorial approach described here to fabricate DNA arrays can also be used for peptide synthesis to create an assortment of peptides of almost any length. The approach is a generic and powerful method to create large numbers of compounds in a very small area in a reasonable amount of time.

Proteins Patterning with Lithography A wide range of proteins has been patterned with features in the micrometer range by making a surface locally hydrophobic or hydrophilic through lithography (see Chapter 2, "Very Thin Resist Layers"). Such patterns may be observed by a number of techniques including fluorescence microscopy, atomic force microscopy (AFM), and the growth response of cultured cells.[25] There are also efforts to produce nanopatterned protein arrays by using proteins that self-assemble into molecular lattices, such as the bacterial S protein. Douglas et al.[26] used metal-decorated crystals of the S proteins of *Sulfolobus acidocaldarius* as a

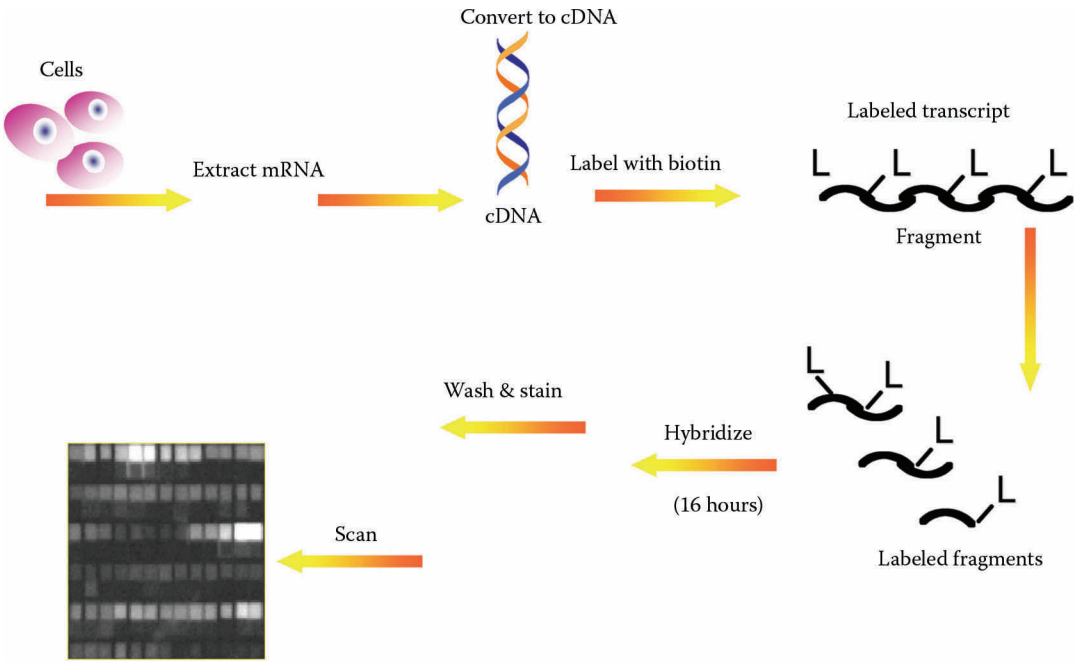


FIGURE 8.17 How an oligo microarray works.

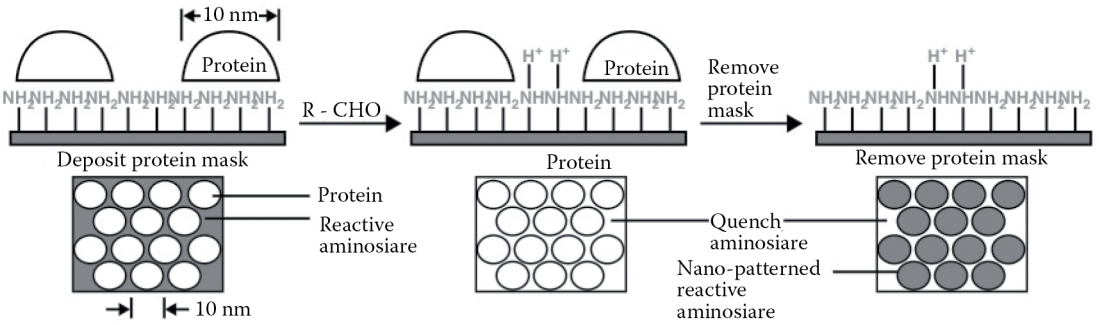


FIGURE 8.18 Nanopatterning of surface chemistry using a self-assembled protein mask. (From Douglas, K., G. Deavaud, and N. A. Clark. 1992. Transfer of biologically derived nanometer scale patterns to smooth substrates. *Science* 257:642–44.[26])

protein mask to pattern hexagonal arrays of 10-nm-diameter holes onto graphite surfaces. Figure 8.18 demonstrates how the self-assembled protein lattice may be used for nanopatterned protein surfaces.

UV lithography, in some special cases, may also work directly for patterning thin protein layers as demonstrated in Example 8.1. In this example a protein grating was created by shining UV, in the presence of oxygen, through a grating photomask. Besides UV lithography, soft and hard x-rays, e-beam, and ion projection lithography can all be used for direct protein film structuring.

Digital Mirror Array Patterning With VLSIPS (see above), a DNA chip of 25-mers may require as many as 100 different masks, leading to a very high cost and a too-long fabrication time.

Singh-Gasson et al.[27] came up with an elegant maskless fabrication alternative for light-directed oligonucleotide microarrays by using a digital micromirror array. Specifically, this group used the digital micromirror device (DMD) from Texas Instruments (http://www.dlp.com/dlp/default.asp), the fabrication of which is described in Chapter 7. This specific DMD consists of a 600 × 800 array of 16-µm-wide micromirrors (the same type used in projection systems). The tiny mirrors are individually addressable and can be used to create any given pattern or image in a broad range of wavelengths. With a 1:1 imaging system, the DMD can be exploited to address 480,000 pixels on a 10 × 14-mm area. This maskless array synthesizer (MAS) or virtual mask array synthesizer is shown in Figure 8.19. An added benefit of this clever approach is that the active surface of the glass substrate can be mounted in a flow cell reaction chamber connected to a DNA synthesizer. Chemical coupling cycles follow light exposure, and these steps are repeated with different virtual masks to grow desired oligonucleotides in any desired pattern. It is obvious that the MAS can also be used for other array patterning/synthesis experiments.

Virtual masks were discussed in Chapter 1 under "Inexpensive Masks and Maskless Optical Projection Lithography for Research and Development." A maskless imaging method was also used by Bertsch et al.[28] at the Swiss Federal Institute of Technology (EPLF) in microstereolithography. This research group uses a computer-controlled liquid crystal display (LCD) as a dynamic pattern generator in microstereolithography (see further below).

Ink Jetting and Microspotting

Introduction Drop delivery systems present an alternative mechanical approach to pattern organic materials on a planar substrate. Although resolution is lower than with lithography methods, these mechanical methods are fast and less expensive.

Ink-Jet Printing Drop delivery in bio-MEMS is based on the same principles as commercial ink-jet printing. The ink-jet nozzle is connected to a reservoir filled with the chemical solution and placed above a computer-controlled *x-y* stage. Depending on the ink expulsion method, even temperature-sensitive enzyme formulations can be delivered. The substrate to be coated is placed on the *x-y* stage, and under computer control, liquid drops (e.g., 50 µm in diameter) are expelled through the nozzle onto a well-defined place on the wafer (see Figure 8.20).

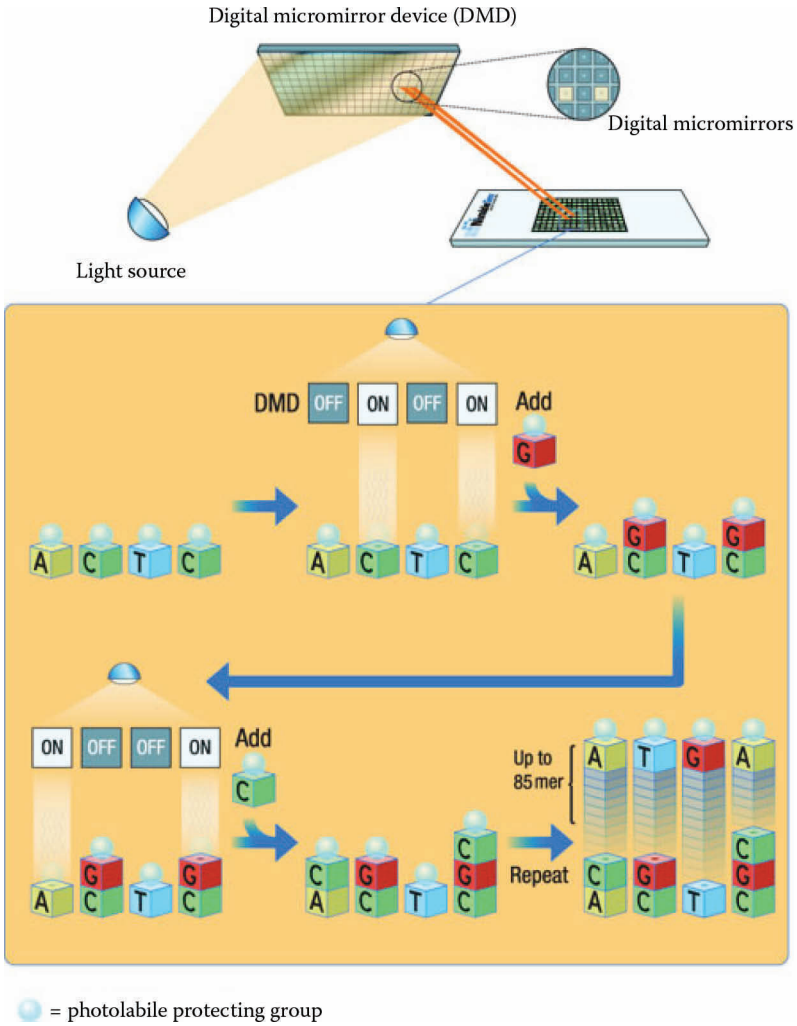


FIGURE 8.19 Schematic of the maskless array synthesizer (MAS). A UV light source is used to illuminate the digital micromirror array. A set of reflective OFF/ON mirrors and a 1:1 imaging system with a numerical aperture (NA) of 0.08 form an image of the pattern on the digital micromirror array on the active surface of the glass substrate. The glass substrate is enclosed in a flow cell connected to a DNA synthesizer. (Courtesy of Dr. R. Green, NimbleGen Systems.)

Different nozzles may print different spots in parallel. Although these drop delivery systems are serial, they can be very fast, as evidenced by epoxy delivery stations in IC manufacturing lines.

The ink-jet mechanism in Figure 8.20 is thermal and is often called a *bubble jet*, with droplets as small as a few picoliters (2–4 pL).[29] Some ink-jet printers can vary the drop size. Five years ago a typical drop size was 30–50 pL; today drop sizes of 2–4 pL are state of the art. This is comparable to the size difference between a softball and a ping pong ball. Smaller drops produce smoother-looking prints because they make smaller dots on the paper. Hewlett Packard's (HP) thermal ink jet (TIJ) represents the state of the art in this arena. In the TIJ, tiny resistors are used to rapidly heat (1,000,000°C/s) a thin (0.1 μm) layer of liquid ink to about 340°C. A superheated vapor explosion vaporizes a tiny fraction of the ink to form an expanding bubble that ejects a drop of ink (and any trapped air) from the ink cartridge onto the paper or other substrate. It is noteworthy that the ink does not boil. The bubble collapse and break off draws fresh ink over the resistor. The bubble formation, expansion, break off, and refill (all very fast) are illustrated in Figure 8.21a. The HP TIJ has all the active power electronics and orifice addressing integrated on the same Si chip with the drop generator as shown in Figure 8.21b. The orifices may be spaced at 300/in. in a single column for a 300-dpi printhead, or for a 600-dpi printhead,

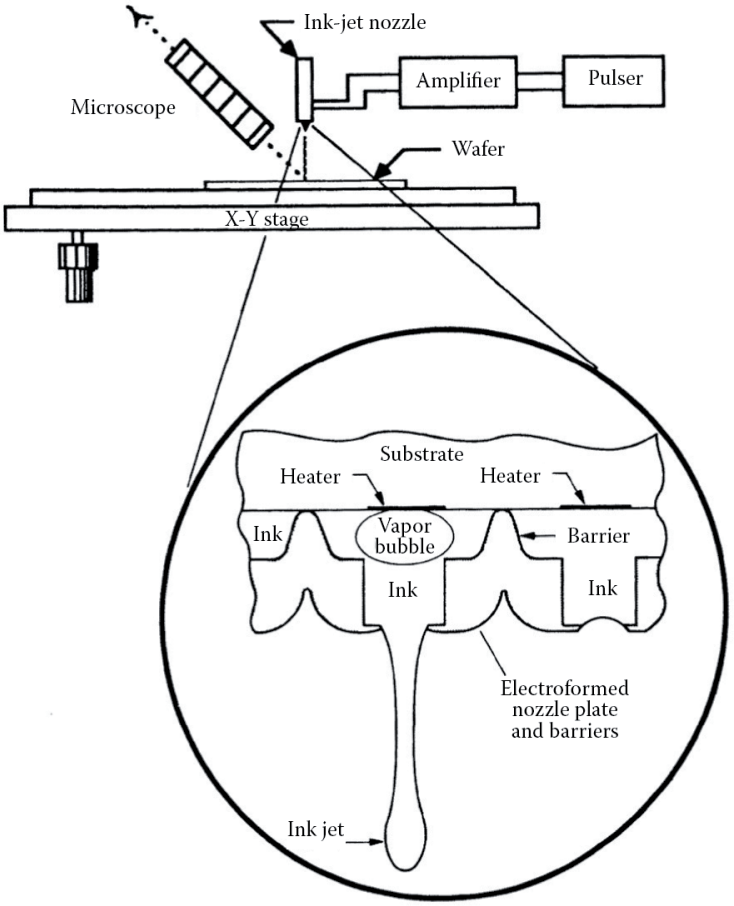


FIGURE 8.20 Ink-jet deposition system.

two offset columns of 300 orifices may be used. It is even possible to put 600 TIJ orifices/in. in a single column. Current piezo ink jets (see next paragraph) only have 90 orifices/in. The TIJ technology is a remarkable piece of engineering. To put this in perspective, engineers trying to develop a commercial ion-sensitive field effect transistor (ISFET) have struggled for more than 30 years to integrate liquids with electronics with little success or market acceptance until very recently (see Volume I, Chapter 7 and Volume III, Chapter 10). Not only did HP succeed in integrating liquid with electronics, they were also able to find an effective method of repeatedly heating the liquid (ink). This way HP has also carved out a very lucrative business in disposable ink cartridges. Perhaps their efforts will also pay off in other fluidic areas, and it may even be worth reconsidering the ISFET using HP technology.

A piezoelectric ink-jet head is used for another type of ink-jet printing and consists of a small reservoir with an inlet port and a nozzle at the other end. One wall of the reservoir consists of a thin diaphragm with an attached piezoelectric crystal. When voltage is applied to the crystal, it contracts laterally, thus deflecting the diaphragm and ejecting a small drop of fluid from the nozzle. The reservoir then refills via capillary action through the inlet. One, and only one, drop is ejected for each voltage pulse applied to the crystal, thus allowing complete control over when a drop is ejected. Such devices are inexpensive and can deliver drops with volumes of tens of picoliters at rates of thousands of drops per second. In conjunction with a computer-controlled x-y stepping stage to position the array with respect to the ink-jet nozzles, it is possible to deliver different reagents to different spots on the array. Arrays of 150,000 spots can be addressed in less than 1 min, with each spot receiving one drop of reagent. As pointed out above, thermal ink jets have a higher orifice density than piezoelectric printing. For more details on both types of ink-jet printing, visit http://www.hp.com/oeminkjet/tij/about.htm.

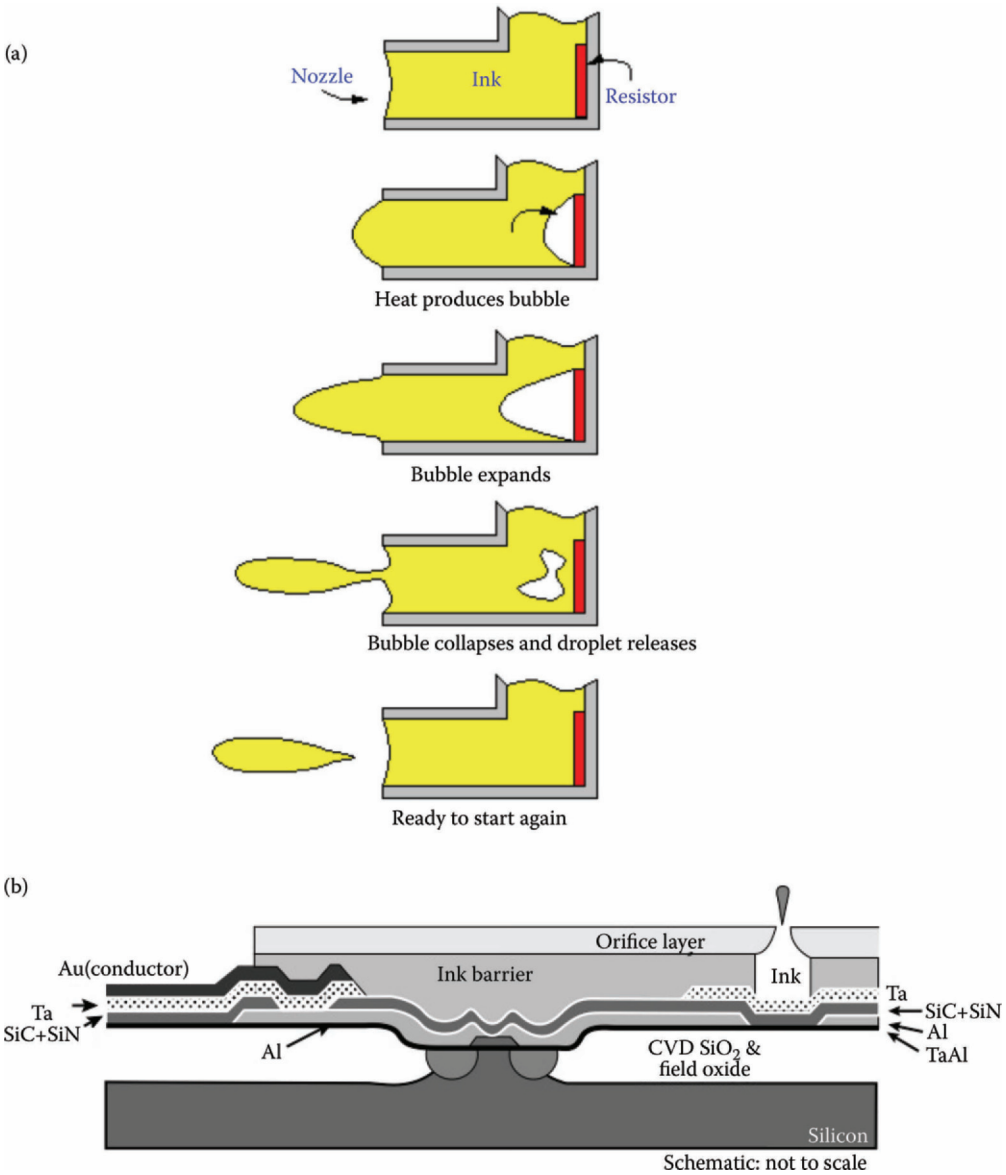


FIGURE 8.21 HP's thermal ink-jet (TIJ) technology. (a) Bubble formation, expansion, collapse, and refill. (b) Integrated power electronics and orifice addressing.

Because a computer controls the pattern of reagents as an array is being made with ink jetting, it is as easy to make 10 arrays with different sequences as it is to make 10 identical arrays. This flexibility is perhaps the main advantage of the ink-jet approach. Achieving high density with the ink-jet approach requires one more trick. Two drops of liquid applied too closely together on a surface will tend to spread into each other and mix. For 40-pL drops, the minimal center-to-center spacing is about 600 μm. This limits the array density achievable with the ink-jet method. One way around this is to engineer patterns in the surface chemistry of the array to produce spots of a relatively hydrophilic character surrounded by hydrophobic barriers. At the small length scales involved (approximately 100 μm), surface tension is the dominant force on a drop of liquid, and a hydrophobic surface will effectively prevent a drop from spreading out beyond the confines of the hydrophilic surface. There are several ways to engineer such a surface. Modern techniques use fluorinated alkyl silanes, which covalently couple to glass and present an extremely hydrophobic surface. They can be patterned by masking the areas to remain hydrophilic, derivatizing the exposed surface with the appropriate silane and then removing the mask by dissolving

it with various organic solvents. The mask itself can be formed by a photolithographic process wherein the array is covered with a thin, uniform layer of photoresist, which is then exposed to light through a photographic negative to define the array. The photoresist is developed, leaving behind a pattern of protective photoresist to act as a mask. Alternatively, a "Whitesides" rubber stamp can be used to apply either the protective mask or the hydrophobic silane itself. Any of these methods will easily produce 100-µm-diameter hydrophilic wells separated by 40-µm hydrophobic barriers, or 5,000 array sites/cm^2. Ink-jet technology has been used to prepare microarrays of single cDNA at a density of 10,000 spots/cm^2.

On a somewhat larger scale, epoxy-delivery systems used in the IC industry similarly deliver drops in a serial fashion on specific spots on a substrate. A typical commercial drop-dispensing system (e.g., the IVEK Digispense 2000) delivers 0.20–0.50 µl in a drop and has a cycle time per dispense of 1 s. A vision system verifies substrate position and accurate dispense location to within ±25 µm of a specified location.

Mechanical Microspotting Mechanically microspotting, as developed by Brown et al. (Stanford University), is simpler than ink-jet printing. In this approach, a popular and inexpensive method for making DNA arrays, a prepared DNA sample (e.g., cDNA or a PCR product) is loaded into a spotting pin by capillary action, and a small volume is transferred to a solid surface by physical contact between the pin and the solid substrate (see Figure 8.22). After a first spotting cycle, the pin is washed, and a second sample is loaded and deposited to an adjacent address. Robotic control systems and multiplexed printheads allow automated microarray fabrication. This very flexible mechanical drop delivery technique takes a small amount of liquid from a 96- (or 384-) well plate and places a tiny (1 nL) drop onto a microscope slide. Because the machine is fully automated, it requires little extra work to make additional slides. One can typically make 120 slides at a time. The microspotted microarrays currently manufactured contain as much as 10,000 groups of cDNA in an area of 3.6 cm^2.

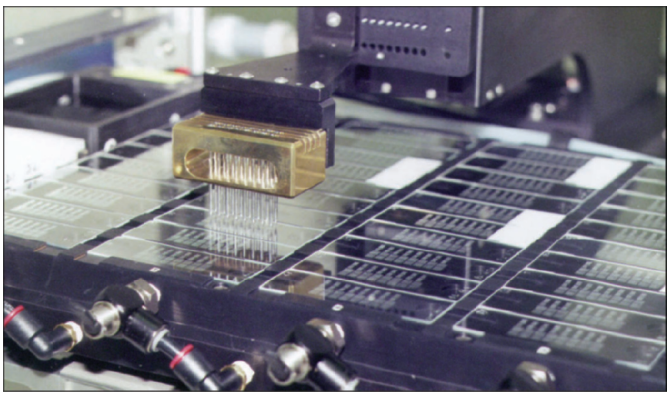


FIGURE 8.22 Spotted arrays of cDNA on 50 glass slides.

As sketched in Figure 8.23, a cDNA array as fabricated with the equipment shown in Figure 8.22 may be used to establish the relative amount of messenger DNA expression in a biological sample (see also Volume III, Chapter 2). For example, we might want to compare the gene expression levels for two cell populations (e.g., from tumor and normal cells) on a single microarray. For that application, both sample (tumor cell) and control DNA (normal cell) are used as targets. One first extracts mRNA from both sample and control and converts the mRNA to cDNA using reverse transcriptase. In a subsequent step, the cDNA is amplified using PCR. DNA probe molecules are then deposited onto the glass slide substrate. Finally, target DNA from tumor and healthy cells are fluorescently tagged (Cy3, Cy5; see Volume I, Chapter 7), and an equal amount of each is added to the cDNA probe array. Depending whether there is a match between target and probe DNA, hybridization does or does not take place. If no hybridization occurs, no color results. If the sample (red) DNA hybridizes and the control does not, a red color results, and if only control (green) DNA hybridizes, a green color results. If both hybridize equally, a yellow color results. The color intensity measured in a fluorescence reader reflects the abundance of the mRNA in the sample. The assumption in this experiment is that cellular mRNA levels directly reflect gene expression.

In Table 8.1 we compare the procedure for measuring mRNA as depicted in Figure 8.23, where we use a spotted cDNA array, with the one in Figure 8.17, where we use a lithographically synthesized oligo array.

In Figure 8.24, we compare three arraying methods: lithography, microspotting, and ink jetting.

TABLE 8.1 Spotted cDNA Array versus Synthesized Oligo Array

	Microarray: Spotted	DNA Chips: Synthesized
Probes	Probes (0.6–2.4 kb) are PCR-amplified full-length cDNA sequences	Probes are 20–25 deoxyoligonucleotides
Manufacture	Spotted by "robo-arms" on nonporous, solid support	Synthesized on glass by solid-phase DNA synthesis coupled with selectively masked light protection and deprotection (photolithography)
Number of spots	About 10,000 "spots" on a microscope glass slide. Maximum 24,000 features per array	Commercial GeneChips have about 300,000 probes on 1.28 × 1.28 cm surface; experimental versions exceed 1,000,000 probes per array
Quality of spots	Variability in spot quality from slide to slide	High quality with little variability between slides
Gene representation	Each gene represented by its purified PCR product	Each gene represented multiply using 16–20 (preferably nonoverlapping) 25-mers
Internal control	Simultaneous analysis of two samples (treated vs. untreated cells) provides internal control	Each oligonucleotide has single-base mismatch partner for internal control of hybridization specificity
Gene expression amount	Relative	Absolute
Cost/flexibility	More flexible and cheaper, homemade	More expensive yet less flexible

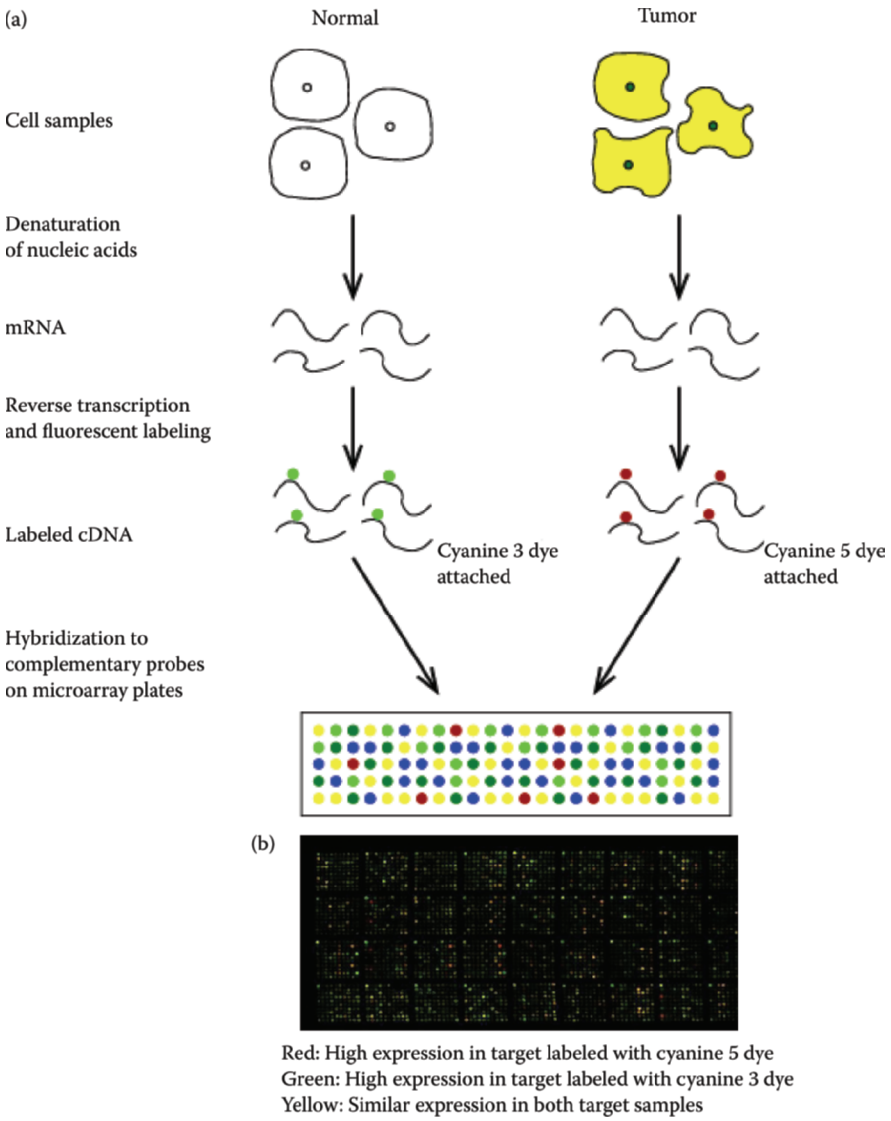


FIGURE 8.23 Example of spotted cDNA microarray analysis: experimental procedure (a) and microarray image (b).

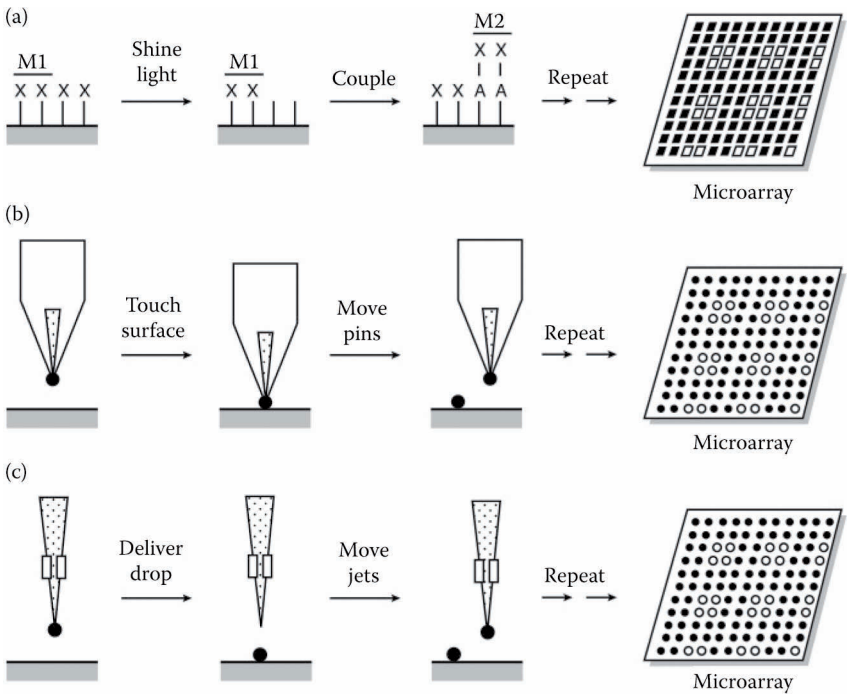


FIGURE 8.24 The three main arraying methods: lithography (a), microspotting (b), and ink-jetting (c).

One disadvantage of microspotting and ink jetting compared with VLSIPS is that each sample must be synthesized, purified, and stored before microarray fabrication.

Some good web sites for further information on arraying technology are Stanford University's *Brown's Lab Guide to Microarraying* at http://cmgm.stanford.edu/pbrown/mguide and BASF's Leming Shi's http://www.gene-chips.com.

Microcontact Printing Micro- or nanocontact printing with a polydimethylsiloxane stamp was reviewed in detail in Chapter 2. In this approach, a Si master is created using a high-resolution lithography technique (e.g., e-beam). Curing an elastomeric prepolymer on this master subsequently forms an elastomeric stamp. The resulting stamp can be inked, for example, with thiols that can be subsequently transferred to a solid support, forming highly localized structured monolayers (see Figure 2.53). Scanning tunneling microscope (STM) studies of the transferred SAMs reveal an achievable monolayer order indistinguishable from that found for SAMs prepared from solution. The technique has been used, for example, to build an antibody grating on a Si wafer by inking the rubber stamp with an antibody solution.[30] The antibody grating alone produces insignificant diffraction, but on immunocapture of cells, the optical phase change produces diffraction. This is an alternative method to make an immunograting than the one described in Example 8.1.

Conductive Polymer Patterning Polymers that can be electropolymerized may also be deposited and patterned by first patterning a thin film metal electrode in a desired pattern on the substrate. In the late 1980s, this author patterned conductive polymers such as polypyrrole and polyaniline with lithography techniques. In one example, arrays of conductive polymer posts were formed on a conductive substrate. Increased ion access to individual conductive polymer posts as compared with access to a uniform film of the same polymer leads to higher electrochemical reversibility, which might find applications in faster polymer battery electrodes, electrochromic devices, enzyme-based biosensors, and microelectronic or molecular electronic devices. Figure 8.25 depicts the procedure for fabricating 3D arrays of electronically conductive polymers and an SEM micrograph of the resulting patterned conductive polymer posts.[31,32]

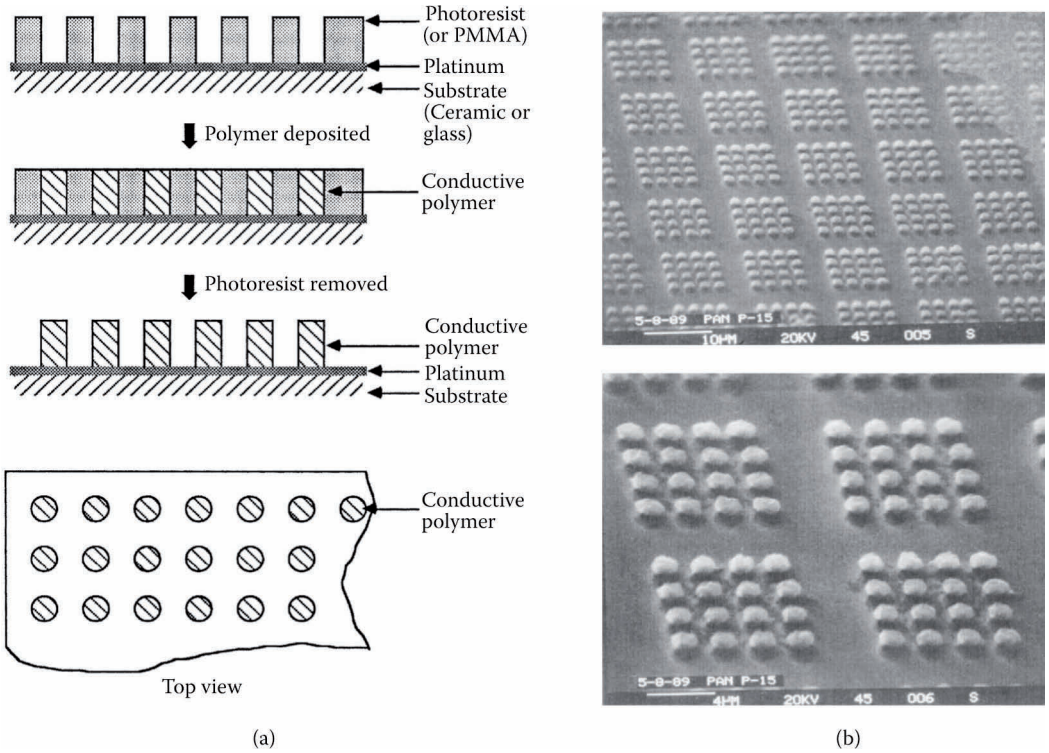


FIGURE 8.25 Micropatterning of conductive polymers. (a) Process sequence. (b) SEM micrograph of conductive polymer submicrometer electrode array (polyaniline doped with tosylate).

Table 8.2 summarizes some of the most popular membrane deposition and patterning techniques.[15] At first glance, photolithography seems very promising, but the chemistry is complex and few results have been obtained to date, the most important exceptions being the i-STAT Abbott Point of Care blood electrolyte and blood gas sensors (http://www.abbottpointofcare.com/) and the Affymetrix DNA arrays (http://www.affymetrix.com). More array developers today are opting for microspotting, ink-jet printing, and screen printing as the safest and least expensive approach.

Photochemical Forming

Introduction

In photochemical forming the energy at the work piece is provided by photons. A prominent example is photolithography, where negative-tone photoresists are cross-linked by photons of the right energy (Chapters 1 and 2). Traditional photolithography techniques result in projected shapes. A photoresist structure can take on any shape in the x-y plane and possess some limited height (in the z-direction), but

TABLE 8.2 Deposition and Patterning Techniques for Planar Chemical Membranes

	Typical Use	Thickness Range (μm)	Cost	Uniformity	Reproducibility	Patterning
Dip coating	Wire ISEs	0.1–50	Low	Poor	Poor	No
Casting	Planar ISEs	0.1–>100	Low	Moderate	Moderate	No
Photolithography	Planar sensors (PVA,PHEMA)	1–10 μm	Moderate	Good	Good	Yes
Lift-off	Immunosensors	0.1–3 μm	Moderate	Moderate	Good	Yes
Plasma etching	PVC, Teflon	1–10 μm	High	Good	Good	Yes
Ink-jet printing	Universal	1–5 μm	Moderate	Poor	Moderate	Yes
Screen printing	Universal	5–50 μm	Low	Moderate	Moderate	Yes

Source: Lambrechts, M., and W. Sansen. 1992. *Biosensors: microelectrical devices.* Philadelphia: Institute of Physics Publishing. With permission.[15]

they remain projected shapes that are not truly 3D because the method does not lend much control over the resist sidewalls along the *z*-axis; such structures are sometimes referred to as 2 1/2D. In miniaturization science, truly 3D lithography is the subject of increased research. Here we review methods of photochemical forming of truly micro- and even nano-3D shapes, including stereolithography or microphotoforming and its newest embodiments such as two-photon polymerization and parallel stereolithography. We start this section with a brief summary of rapid prototyping techniques. In rapid prototyping, additive layering or 3D polymerization techniques have long played an important role for quickly making a nonfunctional mock-up of a real component. Recently improved versions of those methods have also became promising candidates for micro- and nanomanufacturing. It is also important to notice that the term rapid prototyping is less and less the right word to use. Indeed quite often these additive or 3D polymerization manufacturing processes are leading more and more to final functional products rather than nonfunctional prototypes. We classify free-photoforming additive as either chemical or thermal, depending on the primacy of the chemical reaction or the heat application. Thus, low-temperature polymer photoforming is covered in this chapter, whereas the higher-temperature laser-assisted chemical vapor deposition process is covered in Chapter 9.

Rapid Prototyping

The past 20 years witnessed the emergence of a set of new manufacturing techniques with parts made on a layer-by-layer basis or by 3D polymerization. With these technologies, manufacturing time for even the most complex parts is measured in hours instead of days or weeks. Hence these technologies are called *rapid prototyping* (RP). RP, based on layered manufacturing (LM), and direct 3D polymerization in two-photon lithography where no layering is needed form a set of manufacturing techniques by which a solid physical model of a structure is made directly from 3D computer-aided design (CAD) model data, without any special tooling. The CAD data may be generated from 3D CAD modelers, computer tomography (CT),* and magnetic resonance imaging (MRI) scans or model data created by 3D digitizing systems. Although there is no limit to the prototype's complexity, its size is limited by the size of the manufacturing bed. The same data used for the rapid prototype creation can be used to go into production, eliminating all types of sources of human errors.

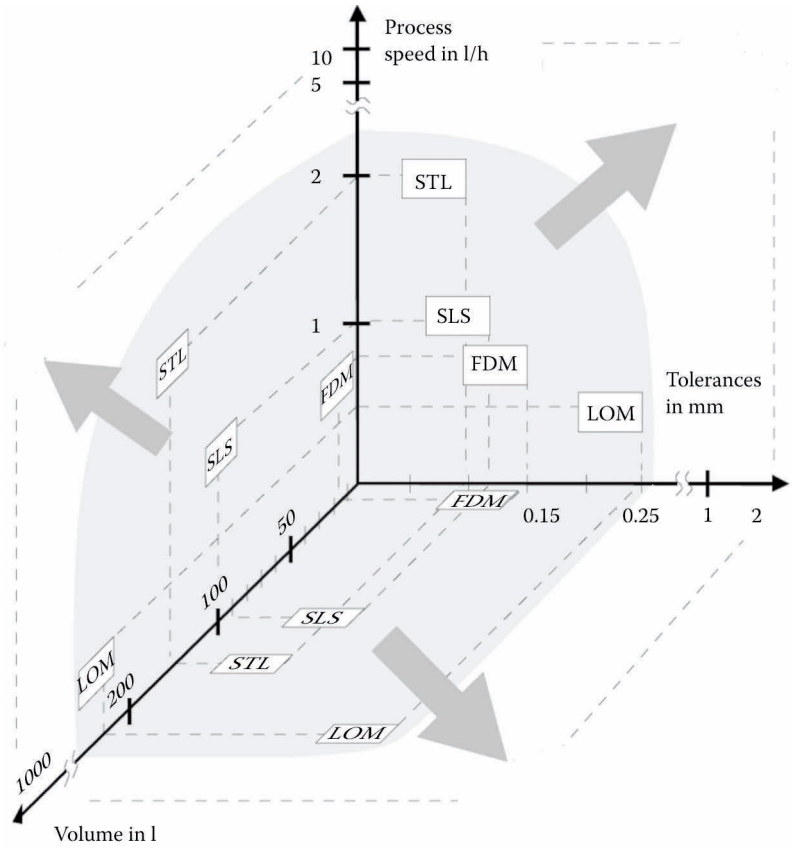
Besides 3D modeling, RP needs specific fluids, powders, wires, or laminates, as well as a machine tool that builds the desired 3D prototypes using different physical principles such as lasers, laminators, and 3D printers. This type of fabrication is referred to as additive manufacturing versus the subtractive manufacturing of milling and turning. Some very new RP methods combine additive and subtractive techniques. For example, the laser sintering/milling hybrid machine, the LUMEX 25C, developed by Matsuura Machinery Corporation, successfully combines freeform manufacturing and high-speed milling (see the photo in the heading of Chapter 9 on thermal energy-based forming).

The first commercial RP method appeared on the market in 1987 from 3D Systems (http://www.3dsystems.com) and was based on stereolithography (SL) where liquid polymers are solidified when exposed to UV light. Stereolithography remains the best known RP system and is widely used in automotive and aerospace industries to fabricate parts from basic design for "show and tell" before the parts are produced. As listed in Table 8.3, there are many different RP methods available that offer varying degree of accuracy, speed, product size (see Table 8.3, inset), as well as a large choice of materials. Two-photon lithography, the newest RP technique that does not rely on layering the part under construction, is dealt with separately below and is not listed in Table 8.3. In all RP methods listed in Table 8.3, a designer, using CAD software, exports a model of the intended product to a standardized portable file format for 3D objects, such as the stereolithography

* Computer tomography (CT) is a powerful nondestructive evaluation (NDE) technique for producing 2D and 3D cross-sectional images of an object from flat x-ray images. Characteristics of the internal structure of an object (such as dimensions, shape, internal defects, and density) are readily available from CT images.

Table 8.3 Rapid Prototyping Techniques

- SL: Stereolithography (see below for details).
- SLS: Selective laser sintering laser beam melts the layer of metal powder to create solid objects (see Chapter 9).
- 3DP: In 3D printing layers of ceramic (plaster) powder are glued to form objects. 3DP uses ink-jet printer technology (e.g., the Z402c from the Z Corporation) to print fine patterns of glue onto a smooth bed of plaster powder (so smooth it looks like paper). The bed is lowered, and a roller spreads a very fine layer (0.004″) over it to print a new layer.
- LOM: Sheets of material (paper, plastic), either precut or on a roll, are glued (laminated) together and laser cut in the laminated object manufacture process.
- FDM: In fused deposition modeling each cross-section is produced by melting a plastic filament (ABS) that solidifies on cooling. The FDM (e.g., the Stratasys FDM 1650) builds layers by extruding a hot, viscous strand of ABS plastic through a very fine nozzle,like squeezing toothpaste out of a tube. The fine strands (0.012″) stick together as they cool.



software (http://stl-software.qarchive.org) that interfaces with the RP machines, analyzes the surface vectors (boundary) of the model, and "slices" the part into many thin layers. Stereolithography files describe only the surface geometry of a 3D object without any representation of color, texture, or other common CAD model attributes. RP machines make parts by producing one of these layers at a time and eventually building up a complex, 3D object. As mentioned earlier, in two-photon lithography there is no need for this layering step.

Stereolithography (SL) is a photoforming-based RP process, and its application to MEMS and NEMS, as well as its newest embodiments such as two-photon polymerization and parallel stereolithography, is detailed below. Selective laser sintering (SLS)

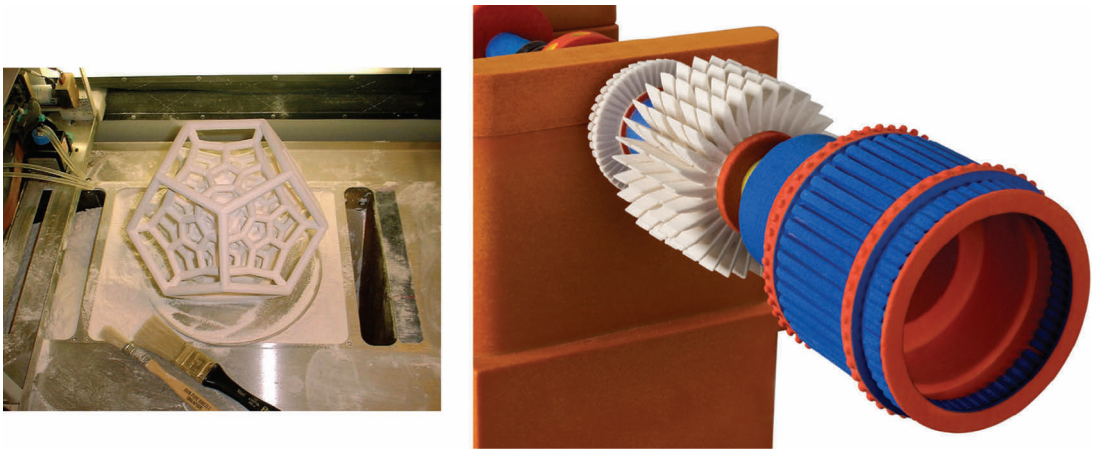


FIGURE 8.26 A 3D-printer by Z Corporation that creates models from plaster powder (http://www.zcorp.com).

is a thermal additive technique expanded on in Chapter 9. A brief explanation of how the other RP techniques listed in Table 8.3 work follows.

Ink-jet printing, discussed earlier, is essentially a 2D method; an important new direction in RP is the concept of 3D printing. In one approach to 3D printing (3DP), layers of ceramic powder (plaster) are glued/bound to form objects. 3DP uses ink-jet printer technology (e.g., from the Z Corporation; http://www.zcorp.com) to print fine patterns of glue onto a smooth bed of plaster powder (so smooth it looks like paper). In a first step the printer covers the bed with a 1/8th inch (3.18 mm) layer of a powder as support so that the parts when finished rest on this base for easy removal. The bed is then lowered, and a roller spreads a next very fine layer (0.004″) over it to print a first layer (Figure 8.26). Z Corporation first introduced high-resolution, 24-color, 3DP (HD3DP™) in 2005 (600 dpi). The printer accurately and precisely deposits colored binder in the desired areas. The binder solidifies the powder in that cross-section of the model, leaving the rest of the powder dry for recycling. The 3D printers control the printhead movement while positioning the head extremely close to the powder, reducing inaccuracies related to fanning of the binder spray.

Sheets of material (e.g., paper, plastic, ceramic, composite), either precut or on a roll, are glued (laminated) together and laser cut in the laminated object manufacture (LOM) process. The laminated part is constructed on a platform with vertical incremental movement. Above the platform, a heated roller heats and compresses each new sheet to the top of the construct. A laser scans the upper surface of the laminated stack to cut the very top layer. Scrap pieces of the laminate remain on the platform as the part is built. The scrap material is diced by the laser into cross-hatched squares that serve as support for the part under construction. The product then comes out as a rectangular block of laminated material containing the prototype and the scrap cubes. The scrap, support material—resulting from the cross-hatch cut by the laser—is easily separated from the prototype part. LOM is a relatively high-speed process (Figure 8.27), and parts can be used immediately after the process because no postcuring is required. Also, no additional support structure is required because the part is supported on its own.

The fused deposition modeling (FDM) technology was developed by S. Scott Crump in the late

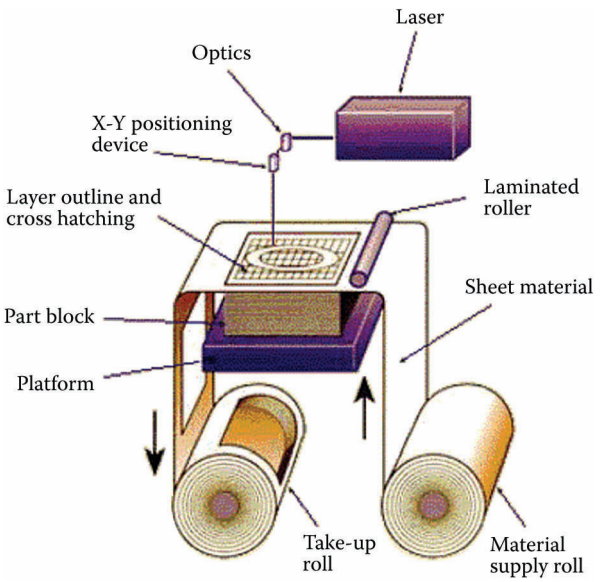


FIGURE 8.27 LOM process.

1980s and was commercialized in 1990. The FDM technology is marketed commercially by Stratasys Inc. (http://www.stratasys.com/selectapart.html). Stratasys-patented technology builds solid models by melting a continuously supplied plastic wire (see Figure 8.28). The supplied wire typically is acrylonitrile butadiene styrene (ABS, a thermoplastic) with a 0.012-in. diameter and is extruded as a hot, viscous strand through a fine nozzle (like squeezing toothpaste out of a tube). A computer controls the nozzle movement along the x- and y-axes, and each cross-section of the prototype is produced by melting a plastic filament (ABS) that solidifies on cooling. After one layer is done, the stage is moved downward, and another layer of deposition starts. The newest FDM models feature two nozzles; one carries the construction material and the other a support wax that can easily be removed afterward, allowing construction of more complex parts. Marketed under the name WaterWorks by Stratasys, this soluble support material is dissolved in a heated sodium hydroxide solution with the assistance of ultrasonic agitation.

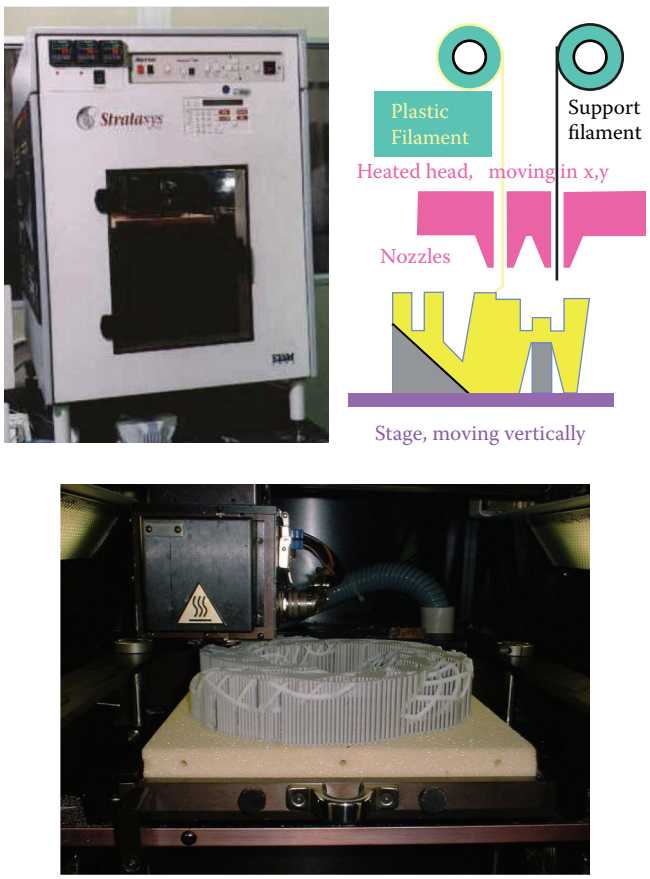


FIGURE 8.28 FDM machine; models are made from molten ABS plastic (http://www.stratasys.com).

The FDM method is environmentally friendly but has a somewhat restricted accuracy because of the nozzle size, which needs to accommodate a viscous material. Also, few types of materials can be incorporated in FDM.

Stereolithography/ Microphotoforming Process

Introduction

The application of rapid prototyping (RP) techniques to MEMS and NEMS requires higher accuracy than what is normally achievable with commercial RP equipment. From the RP techniques listed in Table 8.3, laminated object manufacturing (LOM), fused deposition modeling (FDM), and selective laser sintering (SLN) all must be excluded as microfabrication candidates on that basis. Only stereolithography has the potential to achieve the fabrication tolerances required to qualify as a MEMS or NEMS tool. In this section we explore the basic principles underlying stereolithography and detail the latest enhancements that have made it an attractive high-resolution micro- and nanofabrication method.

Basic Principles Underlying Stereolithography and Microstereolithography

In stereolithography, light exposure solidifies a special liquid resin into a desired 3D shape. Besides producing industrial 3D mockups (see above under RP), micromachinists are exploring this same technology for the production of functional micro- and nanomachines. Some basic concepts illustrating stereolithography are shown in Figure 8.29. A liquid resin is kept either in the free surface mode (see Figure 8.29b) or in the fixed surface mode (see Figure 8.29c). The latter has a resin container with a transparent window plate for exposure. The solidification happens at the stable window/resin interface. An elevator is pulled up over the thickness of one additional layer above the window for each new exposure (Figure 8.30). In the case of free surface, solidification occurs at the resin/air interface, and more care needs to be taken to avoid waves or a slant of the liquid surface. Microstereolithography, derived from conventional stereolithography, was

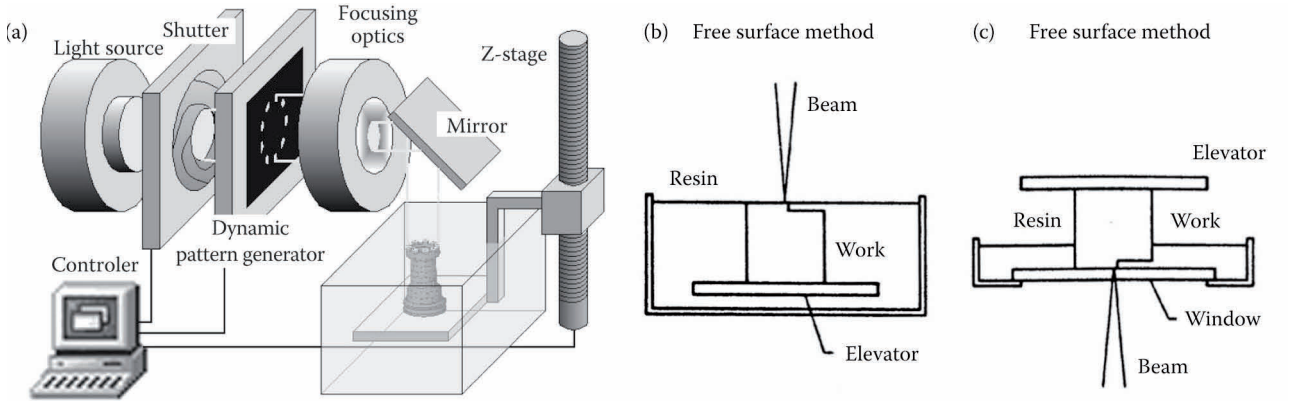


FIGURE 8.29 (a) Stereolithography or photoforming. Free surface method (b) and fixed surface method (c).

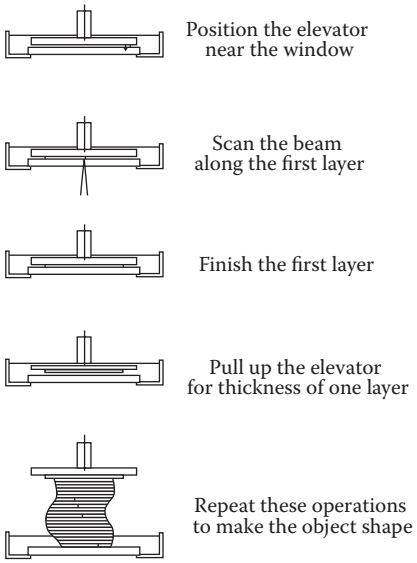


FIGURE 8.30 Stereolithography forming process with the fixed surface method.

introduced by Ikuta et al.[33] in 1993. Whereas in conventional stereolithography the laser spot size and layer thickness are both in the 100-µm range, in microstereolithography, a UV laser beam is focused to a 1–2-µm spot size to solidify material in a thin layer of 1–10 µm. The monomers used in RP and microstereolithography are both UV-curable systems, but the viscosity in the latter case is much lower (e.g., 6 cPs vs. 2000 cPs). In microstereolithography the viscosity should be as low as possible for optimal flat layer formation because high surface tension hinders both efficient crevice filling and flat surface formation in the microscale. Another difference resides in the fact that in microstereolithography the solidified polymer is light enough so that it does not require a support as is required for the heavier pieces made in RP. In Table 8.4 we summarize this comparison between RP and microstereolithography.

Ikuta's group named their scanning microstereolithography "integrated harden polymer stereolithography" or the "IH process." The original IH process used single-photon polymerization to develop a photo-curable polymer with an UV laser beam. A glass slide was placed above the layer being cured, and raising the glass slide small distances above the previously solidified layer generated an extremely thin layer (fixed surface method mode; see Figure 8.29c and Figure 8.30). The IH process resolution in Ikuta's lab was 5 µm in 1992, and, through the various enhancements discussed below, is better than 100 nm at present.[34]

The two major types of stereolithography are scanning stereolithography and projection stereolithography. The scanning stereolithography parts are constructed in a point-by-point and line-by-line

TABLE 8.4 Comparison of Rapid Prototyping and Microstereolithography

	Rapid Prototyping	Microstereolithography
Need for a support	Yes	No
Layer thickness	Hundreds of micrometers	Several micrometers to tens of micrometers
Spot size	Hundreds of micrometers	Micrometer or submicrometer
Type of monomers	UV curable	UV and visible light curable
Viscosity of monomer system	High	Low

fashion, with the sliced shapes written directly from a computerized design of the cross-sectional shapes by a beam in the liquid (Figure 8.30). Projection stereolithography is a parallel fabrication process that enables sets of truly 3D solid structures made of a UV polymer by exposing the polymer with a set of 2D cross-sectional shapes (masks) of the final structures. These 2D shapes are either a set of real photomasks used to subsequently expose the work (Figure 8.31), or they involve a dynamic mask projection system instead of a physical mask. We will compare two types of dynamic mask projection systems for their merit in microstereolithography: a liquid crystal display (LCD) and a digital micromirror display (TI's DMD chip). As we saw in Chapter 1, the DMD chip is also used in maskless lithography (see Chapter 1).

Scanning Microstereolithography, Parallel Microstereolithography and Two-Photon Polymerization

Most of the microstereolithography practiced today is based on the scanning laser beam approach introduced by Ikuta et al.[33] from Nagoya University (Japan) in 1992. The scanning method has the advantage of point-by-point controllability, thereby avoiding unevenness of solidification, which leads to nonuniform shrinking of the work piece. When applying the scanning technique, a laser beam (e.g., a He-Cd laser) is used to solidify one microscopic polymer area at a time to arrive at complicated 3D shapes by stacking thin films of hardened polymer layer upon layer. Process control in this case simply is directed from a CAD system containing the "slice data." The laser beam is focused down to a spot size of a few micrometers into the polymer through a glass window (Figure 8.30).

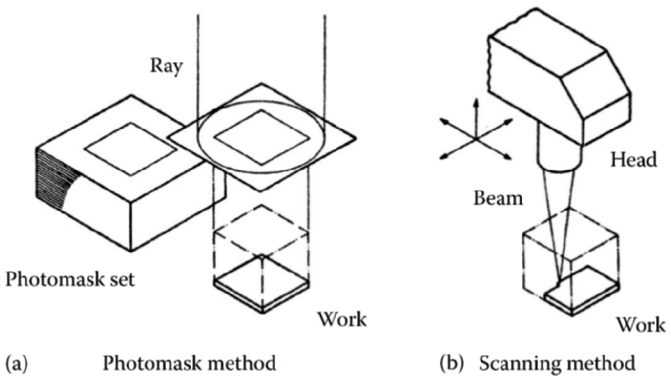


FIGURE 8.31 Exposure with photomask set (a) and exposure with a scanning beam (b). [From Ikuta, K., and K. Hirowatari. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93).* Fort Lauderdale, FL: IEEE;[33] and Ikuta, K., K. Hirowatari, and T. Ogata. 1994. *IEEE international workshop on micro electro mechanical systems (MEMS '94).* Oiso, Japan: IEEE.[35]]

Ikuta et al. used an x-y stage to move the work piece rather than galvanometric mirrors to deflect the laser beam, and this led to a smaller focus spot. They attached the glass window to the z-stage for precise layer thickness control. The position accuracy for the laser beam spot (5-μm diameter in the early days) reached 1 μm in the z-axis and 0.25 μm in the x- and y-directions. Takagi et al.[36] obtained an 8-μm resolution with their photoforming setup, and Ikuta et al.[33,35] reported a minimum solidification unit size of $5 \times 5 \times 3$ μm and a maximum size of fabricated structures of $10 \times 10 \times 10$ mm. As we will learn, this technique has undergone several improvements since these early days.

In scanning microstereolithography, no physical contact occurs between tools and works, and a very large number of layers can be achieved (e.g., 1000 layers) with typical machining times ranging from 30 min to 1 h. High-aspect ratio and very complex shapes, including curved surfaces, can be made with this type of desktop microfabrication method. The objects realized this way used to feature a medium-range accuracy of 3–5 μm.

Professor Ikuta's group and other researchers have introduced various IH process enhancements, including a "mass IH process," a "hybrid IH process," a "super IH process," and finally the "two-photon IH process."

Given the serial nature of the IH process, fabrication speed is a concern, and in 1996 Ikuta et al. introduced a mass IH process. In the mass IH process, multiple parts are produced in parallel, using multiple scanning fiberoptic beams.[37] An array of five single-mode optical fibers (4-μm core diameter) was used to demonstrate this concept. The mass IH process needs to be further developed using optical fiber arrays and by improving the resolution.

In a hybrid IH process other non-IH microcomponents are introduced in the IH structure. As an example of a hybrid IH process, consider Takagi et al.'s work[36,38] wherein a combination of IH with

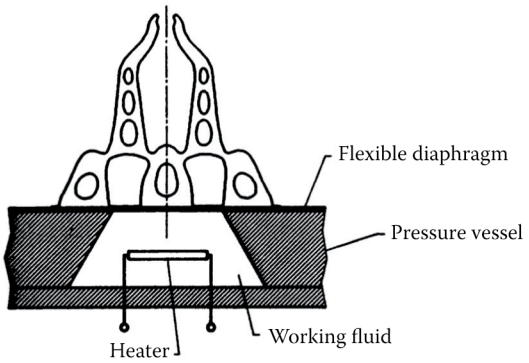


FIGURE 8.32 Thermally driven microclamping tool. The clamping tool is made by photoforming, and the silicon substrate is by silicon bulk micromachining (see text for dimensions). [From Takagi, T., and N. Nakajima. 1994. *IEEE international workshop on micro electro mechanical systems (MEMS '94).* Oiso, Japan: IEEE. With permission.[38]]

more traditional Si micromachined substrates is introduced. In Figure 8.32, a schematic for a photoformed plastic clamp anchored to a Si substrate is represented. The photoformed plastic clamp measures about 2 mm long and 2 mm high and is 250 μm thick. The Si substrate is bulk micromachined to hold a working fluid reservoir and a heater element to actuate the polymer clamp. On heating the working fluid in the micromachined chamber, the thin Si membrane cover bulges out and opens the clamp. Cooling down closes the clamp.

The original IH process and the mass IH process both involve laser scanning and layer preparation. Two major drawbacks associated with this layer-by-layer approach are the limited depth resolution (limited by the thickness of the stacked layers) and the viscosity of the liquid UV-curable monomers (limited form filling and deformation of solidified structures in the microdomain). In 1997, Ikuta and his team upgraded the IH and mass IH process to the "super IH process." [39] The enabling technology for this new process was "deep site pin-point solidification" of a UV polymer. Unlike the typical rapid prototyping (RP) stereolithography process, the polymer in this super IH process is cured below the polymer/air surface where the UV beam is focused. The entire structure is cured into a single body by moving the laser focus, and the curing occurs not only at the liquid surface but also at arbitrary locations inside the liquid resin. A shallow focus lens is used [low depth of focus (DOF) and large numerical aperture (NA) lens] such that the beam's energy is small enough not to cure polymer outside of its focal point. A 3D microstructure is made by scanning the focused laser beam in 3D space into liquid UV monomers, enabling 3D fabrication without any support or sacrificial layers. Because no layer preparation steps are involved, viscosity and surface tension influences are minimized. The properties of the UV-curable monomer system are precisely tuned to ensure that polymerization only happens in the focal point of the laser beam. The UV monomer system used in the super IH process is a mixture of urethane acrylate oligomers, monomers, and photoinitiators. Microparts with freely moving parts can be made this way—no assembly is necessary—allowing for a more advanced type of MEMS device. The resolution of the super IH process is less than 1 μm.

Summarizing, the super IH process has several advantages compared with the layer-by-layer method, such as 1) higher resolution, 2) movable 3D microstructures can be fabricated without supporting or sacrificial layer, and 3) the influence of viscosity or surface tension can be ignored during fabrication.

In 1997, "nanostereolithography" based on two-photon absorption was implemented.[40,41] As we learned in Chapter 2, in the section on quantum lithography, the use of linear absorption of laser light for cross-linking a monomer limits the resolution roughly to the wavelength of the laser. Moreover, deep penetration of the light into the resin is difficult because of the strong photon absorption; therefore, unless very tight focusing of the laser is used as in super IH (see above), the process must be carried out thin layer by thin layer (see Figure 8.30). Using two-photon excitation, one can penetrate deeper into the bulk of the resin and enhance the resolution even further than with super IH. It has been demonstrated that quantum lithography with entangled N-photon states beats the Rayleigh diffraction limit by a factor of N, and in two-photon lithography the resolution is thus improved by a factor of two (as if one used a classical source with wavelength $\lambda/2$).[42,43] When high-intensity light shines on a material, the probability for two-photon simultaneous absorption is proportional to the square of the field intensity, and thus is greatest at the center of a Gaussian laser spot. Subwavelength structures can

be produced because the absorption profile for the two-photon process is narrower than the beam profile. Using a high-NA (e.g., 0.85) lens, the laser spot can be made less than 1 μm in diameter, and significantly narrower structures can be obtained because a certain threshold of intensity must be reached to initiate polymerization. With a femtosecond laser beam, tightly focused into a resin, photo-induced reactions such as polymerization occur only in the close vicinity of the focal point, allowing the fabrication of subwavelength 3D polymer structures by directly writing 3D patterns. Importantly, because the energy required to initiate polymerization is twice that of the incoming radiation, it is possible to excite deep into the monomer as the material is transparent to one-photon excitation, whereas at the focus, where the intensity is high, two-photon excitation occurs. The 3D resolution has reached 140 nm in the two-photon IH process.

In two-photon stereolithography, one often works with epoxy resins and urethane acrylates. Several two-photon photoinitiators with very large two-photon absorption cross-sections to be added to the base polymers have been developed.[44] In these photoinitiator molecules, the photon energy that is absorbed is emitted as photons of a higher energy (up-converted fluorescence). These IH processes can be widely used for making polymeric microdevices. Some examples of 3D two-photon lithography were shown in Figure 2.37 (see also the inset in the heading of this chapter). Two-photon 3D lithography has also been used to make photonic crystals.

Microstereolithography is not limited to polymers: micro- or even nanoceramic and metallic powders can be mixed with the photosensitive resin in microstereolithography of ceramics[45] and metals.[46]

Projection Microstereolithography

In projection microstereolithography, a reduced-size image of complicated patterns is reproduced onto a photoresist. Similar to photolithography, an image is transferred to the liquid photopolymer by irradiating UV light through a patterned mask. The photomask approach, which solidifies a whole layer at the time, is significantly faster than the point-by-point technique but requires a large number of expensive photomasks. In dynamic mask projection microstereolithography, a dynamic mask generator is used instead of a real mask. We compare two types of dynamic mask projection systems: a liquid crystal display (LCD) and a digital micromirror display (DMD).

Bertsch et al.[28] at the Swiss Federal Institute of Technology (EPLF) pioneered an interesting variation on the photomask approach, which, given the many layers involved, is often impractical. This research group uses a computer-controlled LCD as a dynamic pattern generator. A light beam passes through the LCD, and a beam reducer focuses it on the surface of a polymerizable medium. Selective polymerization takes place in the irradiated areas corresponding to the transparent pixels of the LCD. Between the irradiation steps, a shutter blocks the beam, and a new layer of fresh resin (about 5 μm thick) spreads over the object under construction. Complex objects, with a resolution better than 5 μm in the *x*-, *y*-, and *z*-directions, as illustrated in Figure 8.33, have been made this way.[47] The same research group pioneered the combination of planar UV/LIGA lithography (e.g., using SU-8) with microstereolithography—to add nonvertical structures with curved and conical surfaces in a postprocessing step onto planar structures without the need of microassembly.[48] The surfaces of the parts made by microstereolithography are smooth enough to enable conformal electroplating to make a metal mold from the obtained 3D microobject. Polymer replication, of course, remains limited to convex and nonreentrant structures (see Chapter 10).

A comparison of the resolution between conventional stereolithography and small-spot and layer-by-layer microstereolithography was presented by Bertsch et al.[49] in *Rapid Prototyping Journal.* To further improve the lithography—3D photoforming in this case—it is necessary to better understand the shape of a "solidified cell," which depends on both the characteristics of the beam and the resin.

In 2005, Sun et al. introduced another dynamic projection microstereolithography approach using TI's DMD chip as the mask generator (Figure 8.34).[50] The authors point out that the LCD technique from Bertsch et al. has some serious disadvantages compared with their higher-resolution DMD approach.

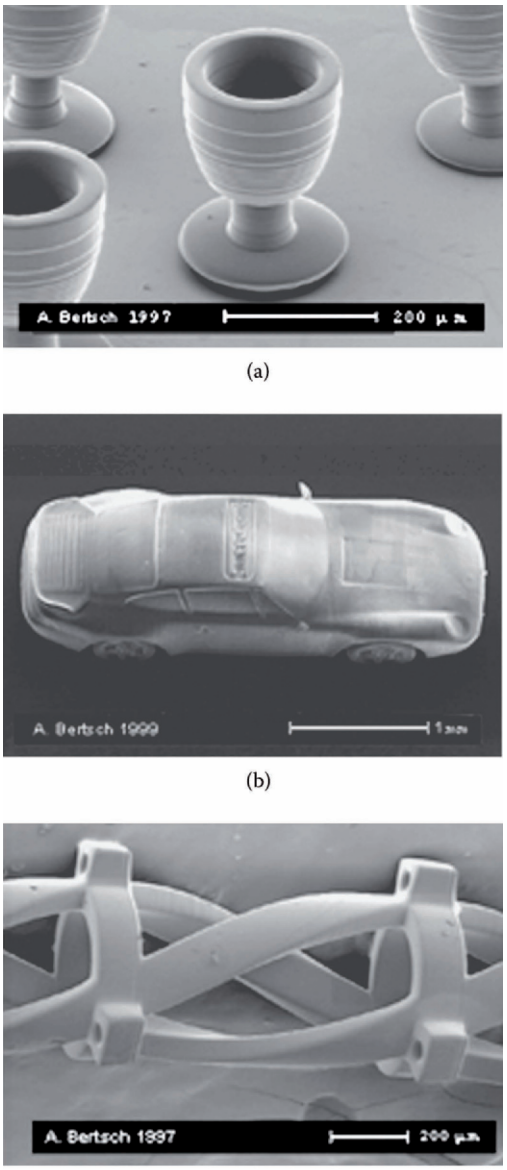


FIGURE 8.33 Examples of objects made by microstereolithography. SEM photos of microcups (80 layers × 5 µm, external diameter is 200 µm) (a); Porsche (673 layers × 5 µm, the diameter of a wheel is 400 µm) (b); microspring (1000 layers × 5 µm, external diameter is 500 µm) (c).

An LCD display not only features larger pixels and a lower filling factor (pixels vs. space between) than a DMD display, but it also has a slower switching speed (~20 ms), and the low optical density of the refractive elements in the off cycle makes for poor contrast. Moreover, the LCD is inserted between four glass plates that are opaque to UV light, restricting its use to visible light where fewer resins are available. Sun et al. made 3D structures with a smallest feature size of 0.6 µm.

Summarizing, microphotoforming offers several advantages over more classical photolithography-based micromachining processes:

1. The turnaround from CAD to prototype takes only 1 h or less.
2. Photoforming is an additive process accommodating virtually any shape.
3. It is a fully automatic process.
4. It requires a small capital investment (less than $30,000).
5. There is no need for a clean room.

This direct-write lithography technique might well represent an alternative to LIGA in cases where 3D shape versatility outweighs accuracy. As with the LIGA technique, the plastic shapes made by stereolithography may be used as a cast for electroplating metals or for other materials that can be molded into the polymer structures.

Electrochemical Forming Processes

Introduction

Both subtractive and additive electrochemical processes are of extraordinary importance in miniaturization science. Subtractive electrochemical techniques were reviewed in Chapter 4; here we review additive/forming electrochemical techniques. Electrochemical forming includes electroless and electrodeposition, and we analyze pros and cons of both. We add information on electrodeposition through masks, working with instant masks in electrochemical fabrication (EFAB) and maskless metal deposition methods. Today, there is a strong push in the IC industry and in micromachining toward dry processing, and deep dry-etching developments are particularly swift. Electrochemical machining, on the other hand, despite its many uses, is often still considered a "black art" and "dirty," with lack of cleanliness and creation of waste particles making it an unacceptable technique. In reality, besides its major technological advantages, electrochemistry offers several intrinsic advantages, such as low capital equipment, high deposition and etch rates, and relatively simple operation. In the sections below, we learn that there are many cases in which only wet

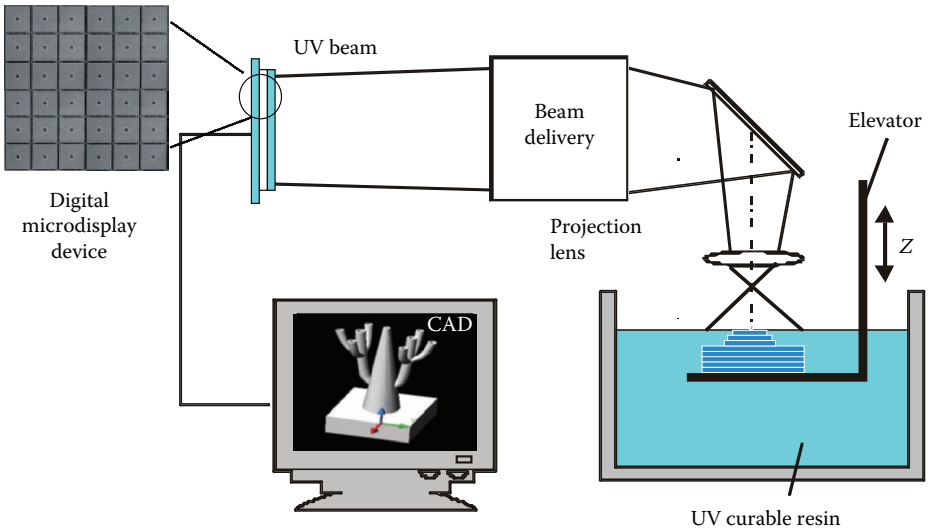


FIGURE 8.34 A dynamic projection microstereolithography approach using TI's DMD chip as the mask generator. (From Sun, C., N. Fang, D. M. Wu, and X. Zhang. 2005. Projection micro-stereolithography using digital micro-mirror dynamic mask. *Sensors Actuators A* 121:113–20.[50])

chemical and electrochemical processes are feasible; for example, line-of-sight deposition techniques cannot readily be used on curved or irregular surfaces or to plate into via holes or blind vias. CVD often leaves voids when metallizing (e.g., tungsten) in a high-aspect-ratio structure, whereas with electrodeposition such voids can be avoided completely. Multilayer PCBs with through holes connecting the individual metal layers would not be possible without electroless or electrochemical technology. As long as a solution can wet and fill a cavity to be plated, a metal pattern can be generated precisely, replicating the mold down to angstrom dimensions. In Table 8.5, many applications of electrochemistry are reviewed, and

TABLE 8.5 Situations for which an Aqueous Electrochemical or Chemical Technique Is the Only Means to Obtain the Desired Result

• Uniform plating on irregular surfaces
• Leveling of rough surfaces
• Plating of via holes and blind vias
• Formation of alloys of metastable phases
• Compositionally modulated structures
• Amorphous metal films
• Maskless, high-speed, selective deposition of metals and alloys
• Most faithful reproduction of features of polymeric masks
• Smallest dimension features with highest height-to-width aspect ratio (highest packing density of conductors)
• Metal deposits with incorporated particles (e.g., diamond, WC, TiC, Al_2O_3, Teflon, oil)
• Wear-resistant surfaces
• Corrosion-resistant surfaces
• Uniform thickness nonporous anodic films
• Anodic films with porous structure
• Very uniform, thick, pore-free, polymer films on irregular surfaces
• Very uniform thickness, glazed, and devitrified glass surfaces
• Shaping of almost any metal or semiconductor and electromachining without introducing stress
• Electropolishing of metal surfaces to mirror bright finishes

Source: Based on Romankiw, L. T., and T. A. Palumbo. 1987. *Proceedings: symposium on electrodeposition technology, theory, and practice.* San Diego, CA[51]; and Romankiw, L. T. 1984. Electrochemical technology in electronics today and its future: a review. *Oberflache-Surface* 25:238–47.[52]

Table 8.6 Electrochemical Applications in the IC Industry and Micromachining

Photocircuit boards
• Single and multilayer epoxy boards
• Flexible circuit boards
• Electrophoretically glazed steel boards
Contacts and connectors
• Beam leads
• Contacts (pins and sockets)
• Reed switches, etc.
Cabinets and enclosures
• Corrosion protective surfaces
• Electromagnetic shielding
• Decorative purposes (anodization, electrophoretic painting)
Auxiliary equipment used in device fabrication
• Paste-screening masks
• Metal evaporation masks
• X-ray lithography masks
• Diamond saws and cutting tools
Active elements
• IC chips
• Magnetic recording heads
• Recording surfaces
• Displays
• Wear-resistant surfaces
Chip carriers and packages
• Chip in tape packages
• Surface mount boards
• Dual in-line packages
• Pin-grid array packages
• Multilayer ceramic packages
• Hybrid packages

Source: Based on Romankiw, L. T., and T. A. Palumbo. 1987. *Proceedings: symposium on electrodeposition technology, theory, and practice.* San Diego, CA[51]; and Romankiw, L. T. 1984. Electrochemical technology in electronics today and its future: a review. *Oberflache-Surface* 25:238–47.[52]

in Table 8.6, we define several situations in which an aqueous chemical or electrochemical technique may be the only means by which a desired property or quality of a deposit can be achieved.

Electroless Metal Deposition

Underlying Principles

To continuously build thick deposits by chemical means without applying a voltage and without consuming the substrate, i.e., electroless plating, it is essential that a sustainable oxidation reaction be used instead of the dissolution of the substrate itself as in the case of immersion plating. In an immersion plating bath, the electrons for the metal deposition are supplied by the base metal, and effectively the base metal is the reducing agent. As the base metal donates electrons to the metal being plated, the base metal itself is oxidized and goes into solution, which is why this is also called a *replacement reaction*. The key characteristics of the immersion plating bath is that it is self-limiting, which means that once the base metal, e.g., copper, in the case of a printed circuit board (PCB), is covered with the new metal, the plating ceases. This results in a very thin, dense, nonporous coating, which is very economical, from a bath that is extremely stable. In an "electroless" plating bath, on the other hand, the electrons are donated by a reducing agent that is already in the bath. In Figure 8.35, the difference between immersion plating and electroless deposition is illustrated by comparing deposit thickness versus time; immersion plating is self-limiting, but electroless plating is not.[53] In an electroless copper bath, the formaldehyde is the reducing agent. In electroless Ni, it is either hypophosphite or dimethyl amine borane.

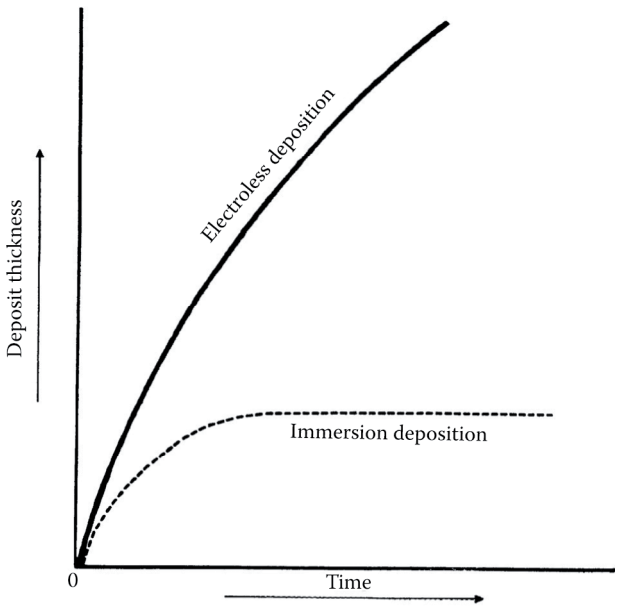


FIGURE 8.35 Thickness versus time comparison between electroless and immersion deposition. [Based Mallory, G. O., and J. B. Hadju, eds. 1990. *Electroless plating: fundamentals and applications.* Orlando, FL: American Electroplaters and Surface Finishers Society (AESF). With permission.[53]]

The reduction of Ni ions (Reaction 8.1) is fed electrons by an oxidation reaction such as that of hypophosphite (Reaction 8.2), resulting in an overall reaction for Ni deposition given by Reaction 8.3.

Reduction:

$$Ni^{2+} + 2e^- \rightarrow Ni \qquad \text{Reaction 8.1}$$

Oxidation:

$$H_2PO_2^- + H_2O + 2e^- \rightarrow H_2PO_3^- + 2H^+ + 2e^- \qquad \text{Reaction 8.2}$$

Overall reaction:

$$Ni^{2+} + H_2PO_2^- + H_2O \rightarrow Ni + H_2PO_3^- + 2H^+ \qquad \text{Reaction 8.3}$$

Reaction 8.1 continues (in principle) until all the hypophosphite in solution is consumed (Reaction 8.2). In immersion deposition, Reaction 8.1 stops as soon as the whole substrate surface is covered with Ni metal and there is no further dissolution of uncovered substrate possible. In parallel to Reaction 8.1, more or less severe hydrogen reduction goes on:

$$2H^+ + 2e^- \rightarrow H_2 \qquad \text{Reaction 8.4}$$

Copious hydrogen evolution can upset the quality of the depositing metal film and should be avoided. The hydrogen evolution rate is not directly related to that of the metal deposition and mainly originates from the reductant molecules. Stabilizers (i.e., catalytic poisons) are needed in electroless deposition baths because the solutions are thermodynamically unstable; deposition might start spontaneously onto the container walls. Substances that poison hydrogenation* catalysts such as thiourea, Pb^{2+}, and mercaptobenzothiazole function as stabilizers in such electroless baths. Besides stabilizers, the metal salt, and a reducing agent, electroless solutions may contain other additives such as complexing agents, buffers, and accelerators. Complexing agents exert a buffering action and prevent the pH from decreasing too fast. They also prevent the precipitation of metal salts and reduce the concentration of free metal ions. Buffers keep the deposition reaction in the desired pH range. Accelerators, also termed *exaltants*, increase the rate of deposition to an acceptable level without causing bath instability. These exaltants are anions, such as CN^-, thought to function by making the anodic oxidation process easier. In electroless copper, for example, compounds derived from imidazole, pyrimidine, and pyridine can increase the deposition rate to 40 μm/h. The electroless deposition must occur initially and exclusively on the surface of an active substrate and subsequently continue to deposit on the initial deposit through the catalytic action of the deposit itself. Because the deposit catalyzes the reduction reaction, the term *autocatalytic* is often used to describe the plating process. Electroless plating is an inexpensive technique enabling plating of conductors and nonconductors alike (plastics such as ABS, polypropylene, Teflon, and polycarbonate are plated in huge quantities). A catalyzing procedure is necessary for electroless deposition on nonactive surfaces such as plastics and ceramics. The most common method for sensitizing those surfaces is by dipping into $SnCl_2$/HCl or immersion in $PdCl_2$/HCl.[53] This chemical treatment produces sites that provide a chemical path for the initiation of the plating process.

Metal alloys such as nickel-phosphorus, nickel-boron, cobalt-phosphorus, cobalt-boron, nickel-tungsten, copper-tin-boron, and palladium-nickel can be produced by codeposition. In the case of Ni deposits, with or without phosphorus and boron incorporated, different electroless solutions are used to obtain optimum hardness, effective corrosion protection, or optimum magnetic properties. Recent experimental results show that composite material also can be produced by codeposition. Finely divided, solid particulate material is added and dispersed in the plating bath. Electroless Ni with alumina particles, diamond, silicon carbide, and PTFE have reached the commercial market. Table 8.7 presents a list of electroless plating baths.[53]

Methods to determine the electroless deposition rate can be split into two categories: electrochemical and nonelectrochemical. Electrochemical techniques include Tafel extrapolation, DC polarization, AC impedance, and anodic stripping. Nonelectrochemical methods include weight gain, optical absorption, resistance probe, and acoustic. For more details on each of these techniques, see Ohno[55] and references therein.

* *Hydrogenation* is a class of chemical reactions that result in an addition of hydrogen, usually to unsaturated organic compounds.

Table 8.7 Typical Electroless Plating Baths

Component	Concentration (per L)	Application/Remark	pH	Temp (°C)
Au	1.44 g $KAu(CN)_2$, 6.5 g KCN, 8 g NaOH, 10.4 g KBH_4	Plate beam leads on silicon ICs, ohmic contacts on n-GaAs	13.3	70
Co-P	30 g $CoSO_4 \cdot 7H_2O$, 20 g $NaH_2PO_2 \cdot H_2O$, 80 g Na_3citrate$\cdot 2H_2O$, 60 g NH_4Cl, 60 g NH_4OH	Magnetic properties	9.0	80
Cu	10 g $CuSO_4 \cdot 5H_2O$, 50 g Rochelle salt, 10 g NaOH, 25 mL conc. HCHO (37%)	Printed circuit boards	13.4	25
Ni-Co	3 g $NiSO_4 \cdot 6H_2O$, 30 g $CoSO_4 \cdot 7H_2O$, 30 g Na_2malate$\cdot 1/2H_2O$, 180 g Na_3citrate$\cdot 2H_2O$, 50 g $NaH_2PO_2 \cdot H_2O$		10	30
Ni-P	30 g $NiCl_2 \cdot 6\ H_2O$, 10 g $NaH_2PO_2 \cdot H_2O$, 30 g glycine	Corrosion and wear resistance on steel	3.8	95
Pd	5 g $PdCl_2$, 20 g Na_2EDTA, 30 g Na_2CO_3, 100 mL NH_4OH (28% NH_3), 0.0006 g thiourea, 0.3 g hydrazine	Plating rate is 0.26 μm/min		80
Pt	10 g $Na_2Pt(OH)_6$, 5 g NaOH, 10 g ethylamine, 1 g hydrazine hydrate (added now and then to maintain this concentration)	Plating rate is 12.7 μm/h		35

Source: Based on Mallory, G. O., and J. B. Hadju, eds. 1990. *Electroless plating: fundamentals and applications.* Orlando, FL: American Electroplaters and Surface Finishers Society (AESF)[53]; and Romankiw, L. T. 1976. *Etching for pattern definition.* Pennington, NJ: The Electrochemical Society.[54]

Electroless Plating in the IC Industry

The IC industry applies electroless metal deposition for a wide variety of applications, some of which are incorporated in Tables 8.5 and 8.6. In Figure 8.36, we single out the schematic for the electroless Cu deposition of a buried conductor as an illustrative example of an IC application. The fabrication of this buried conductor illustrates many of the processes we have studied so far; lithographic patterning of a dielectric material (Figure 8.36a), pattern transfer by anisotropic dry etching of a groove in the dielectric substrate (Figure 8.36b), evaporative deposition of an Al seed layer (Figure 8.36c), seed layer patterning with lift-off (Figure 8.36d), trench filling with electroless copper (Figure 8.36e), and spin-on glass dielectric layer deposition for planarization (Figure 8.36f).

Electroless Plating in Microfabrication

In micromachining applications, electroless deposition is used for the same purposes as in the IC

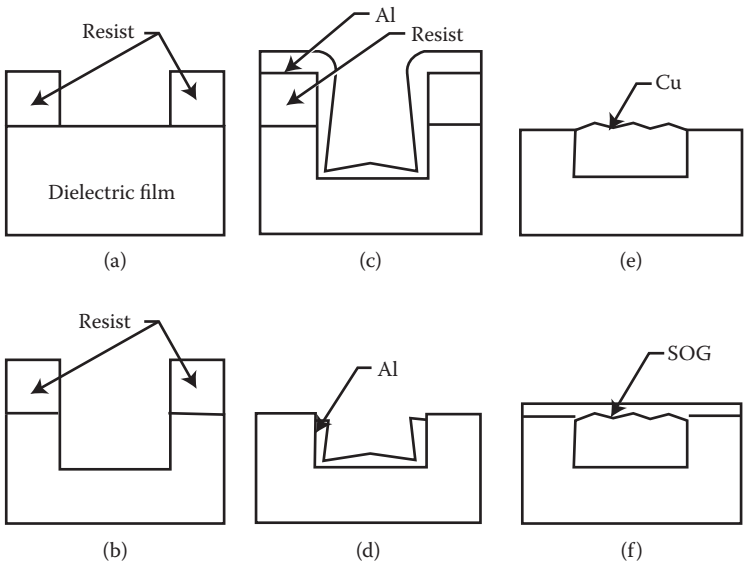


FIGURE 8.36 Schematic of a buried Cu conductor process. (a) Photoresist pattern over a dielectric. (b) Pattern transfer by anisotropic dry etching. (c) Deposition of a thin aluminum seed layer. (d) Seed metal patterning with lift-off. (e) Trench filling with electroless copper. (f) SOG dielectric layer for planarization.

industry, as well as to make higher-aspect-ratio structural microelements from a wide variety of metals, metal alloys, and even from composite materials. These applications often include polymer molds and introduce new challenges in terms of uniformity of plating rates and removal of the metal form from the master photoresist mold. Electroless plating may be elected above electrodeposition due to the simplicity of the process because no special plating base (electrode) is needed. This represents a major simplification for combining LIGA and pseudo-LIGA structures with active CMOS electronics, where a plating base could short out the active electronics. In addition, electroless Ni exhibits less stress than electrodeposited Ni, a fact of considerable importance in most mechanical structures. The major concern is the temperature of the electroless plating processes, which is often considerably higher than for electrochemical processes. Few studies of electroless plating in LIGA molds have been reported.[56]

We add two interesting MEMS applications of electroless plating: electroless plating of beveled structures and electroless plating using microfluidic structures. Usually, electroless plating is an isotropic process with no means to control the profile of plated features. However, van der Putten et al.[57,58] were able to induce anisotropy in the plating process by taking advantage of the increased poisoning effect a stabilizer has at the edges of any given pattern (see Figure 8.37a). Because the edges of a given pattern experience additional deposition from nonlinear diffusion, mass transport to these edges is enhanced compared with material supply to the bulk of a substrate. As a result, the surface concentration of the adsorbed stabilizer (Pb^{2+}) is higher at these edges. By choosing the proper Pb^{2+} concentration, the edges can be selectively poisoned: bevels can be grown with an angle, α, which is a function of the stabilizer concentration in the solution (see Figure 8.37b). More important, van der Putten's bevel-plating technique eliminates the commonly observed lateral overgrowth of resist patterns, and it might have broader applications in the area of micromachining. The technique is less controllable than Maciossek's maskless electrodeposition, discussed further below, but does not require a patterned plating base.[59]

Whitesides et al. introduced a means of depositing and patterning metal electrodes inside prefabricated channels by flowing plating and etching solutions as shown in Figure 8.38. Under conditions of laminar flow, reactions, including electroless plating, can take place at the interface between two or more liquids flowing side by side down a channel. In the example in Figure 8.38, the fabrication of a three-electrode system is demonstrated. Two gold electrodes are formed by etching a previously deposited gold strip. The silver reference electrode is formed using a two-phase laminar flow arrangement.[60]

Electrodeposition

Underlying Principles

Electroplating, we saw in Volume I, Chapter 7, takes place in an electrolytic cell. The reactions involve current flow under an imposed bias. As an example, consider the deposition of Ni from $NiCl_2$ in a KCl solution with a graphite anode (not readily attacked by Cl_2) and a Au cathode (inert surface for Ni deposition). With the cathode sufficiently negative and the anode sufficiently positive with respect to the solution, Ni deposits on the cathode and Cl_2 evolves at the anode. The process differs from electroless Ni deposition in that the anodic and cathodic processes occur on separate electrodes and that the reduction is affected by the imposed bias rather than a chemical reductant. Important process parameters are pH, current density, temperature, agitation, and solution composition. The amount of hydrogen evolving and competing with metal deposition depends on the pH, the temperature, and the current density. Important causes of defects in electrodeposited metallic microstructures are the appearance of hydrogen bubbles, and these three parameters need very precise control. Besides typical impurities, such as airborne dust or dissolved anode material, the main impurities are metal hydroxide formed at increased pH values in the cathode vicinity and organic decomposition products from the wetting agent. The latter two can be avoided to some degree by monitoring and controlling the pH and by adsorption of the organic decomposition products on activated carbon.

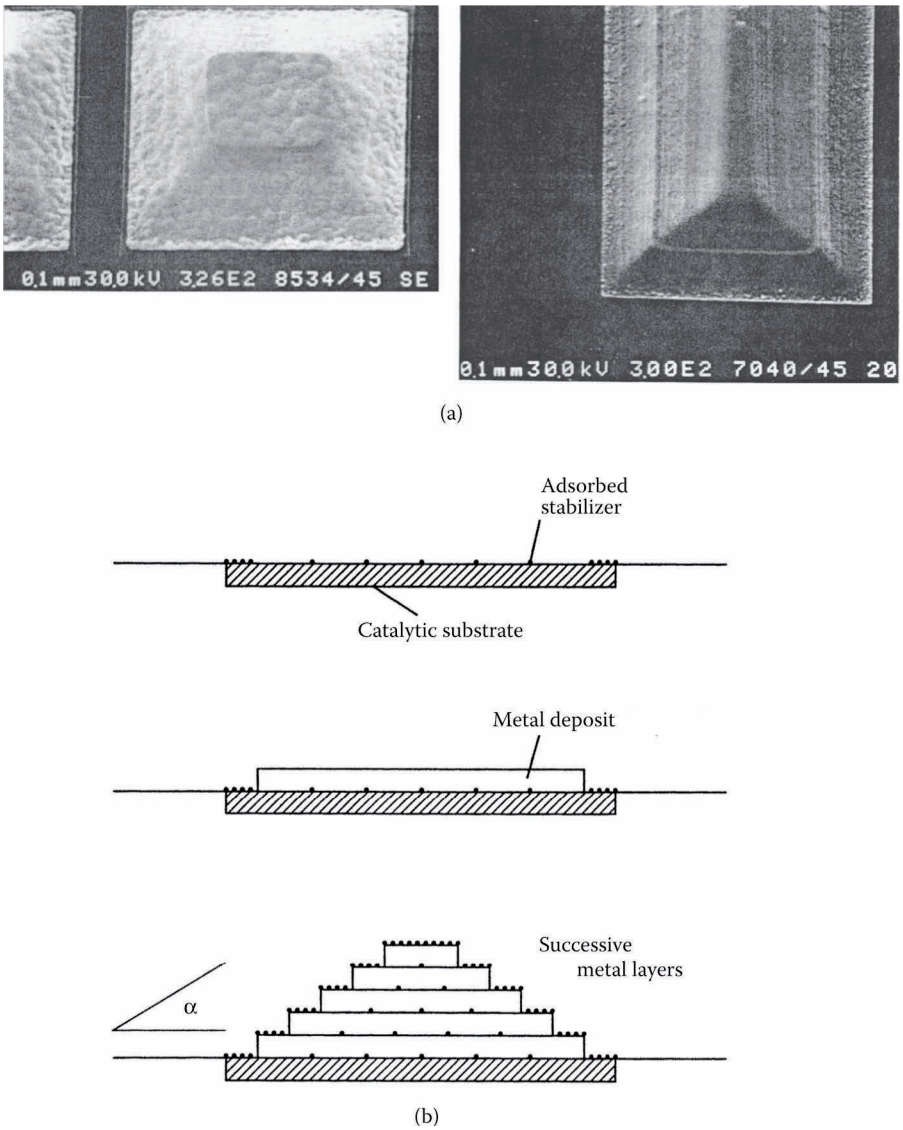


FIGURE 8.37 Electroless micromachining: bevel plating. (a) Examples of bevel-plated nickel structures. (b) Method to make bevel-plated devices. (From van der Putten, A. M. T., and J. W. G. de Bakker. 1993. Anisotropic deposition of electroless nickel-bevel plating. *J Electrochem Soc* 140:2229–35. With permission.[57])

Electrodeposition in ICs, MEMS, and NEMS

Through-Mask Electroplating

Overview In a typical electrodeposition through a polymer mask in IC applications, metal conductors must be deposited on a dielectric substrate. The dielectric is usually made conductive first by sputtering a thin adhesion metal layer (e.g., Ti, Cr) and then a conductive seed layer (e.g., Au, Pt, Cu, Ni, and NiFe). The thickness of the thin refractory metal adhesion layer may be as small as 50–100 Å, whereas the thickness of the conducting seed layer can range from 150–300 Å.[61] The key requirement for the seed layer is that it is electrically continuous and offers low sheet resistance. After forming a pattern in a spin-coated polymer by UV exposure, e-beam, or x-ray radiation and developing away the exposed resist, electrical contact is made to the seed layer, and electrodeposition in the resist mold is carried out. The use of a solvent-containing development agent ensures a substrate surface free of grease and ready for plating. The metal layer growing on the substrate fills the gaps in the resist configuration, thus forming a complementary metal structure. Pollutants may cause hydrogen bubbles to cling to the resist structures, resulting in pores in the metal deposit; thus, the bath must be kept clean, for example, by circulating through a membrane filter with 0.3-μm pore openings.[56]

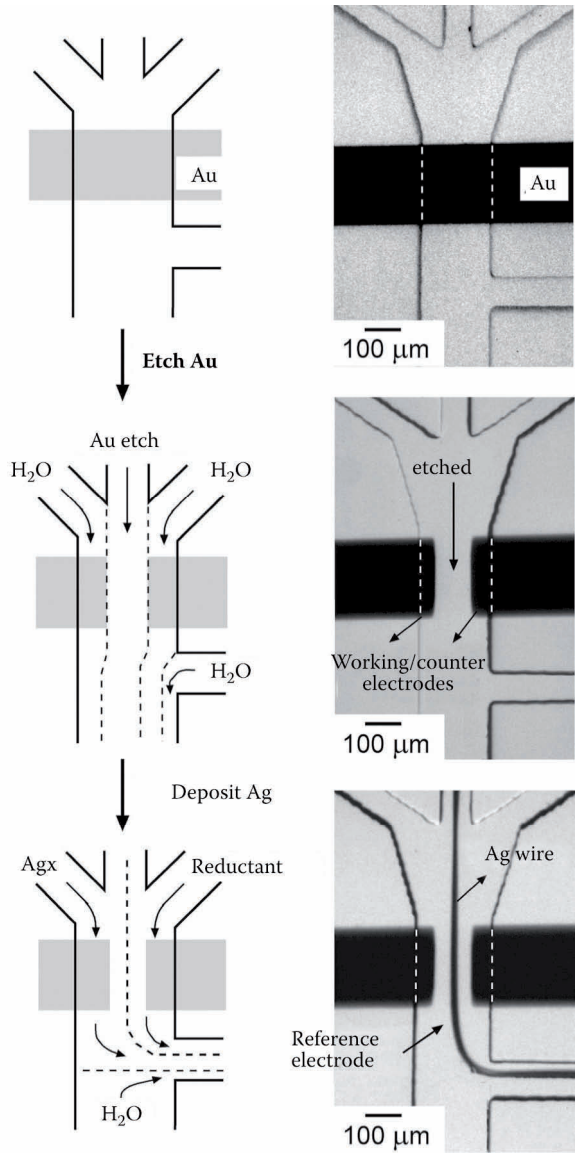


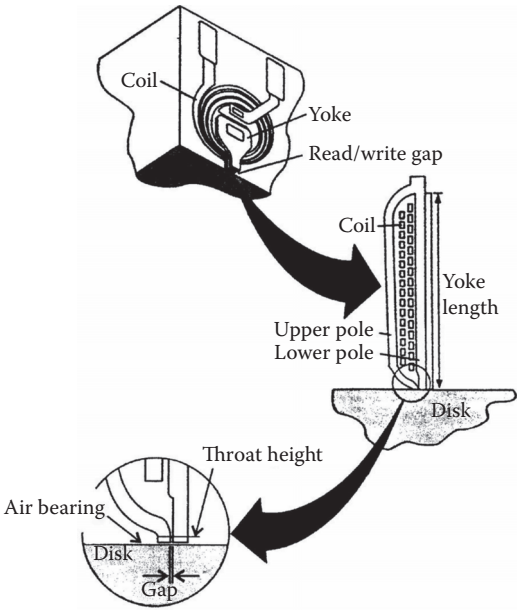
FIGURE 8.38 The formation of a three-electrode system within a channel as shown by optical micrographs. Using a three-phase laminar flow system, two gold electrodes are formed by etching a previously deposited gold strip, and using a two-phase laminar flow system, a silver reference electrode is fabricated. (Courtesy of Dr. G. Whitesides, Harvard University.)

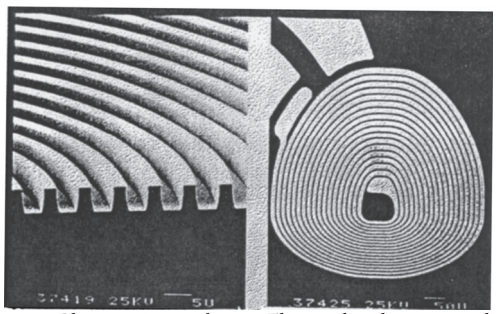
Plating through-mask technology is successfully used in volume production of beam leads and bumps on IC wafers, fabrication of thin film magnetic heads, fabrication of PCBs, x-ray lithography mask gratings, and diffraction gratings. Electrodeposition in general has been used extensively in the electronics industry in many stages of the manufacturing process, from the device stage, chip carriers, and PCB to corrosion protection and electromagnetic shielding of the electronic enclosures (see Tables 8.5 and 8.6). Processes include electrodeposition and electroless deposition of copper, nickel, tin, tin-lead alloys, and precious metals such as gold, gold alloys, palladium, and palladium alloys, as well as NiFe, CoP, NiCoP, and other magnetic alloys.[51]

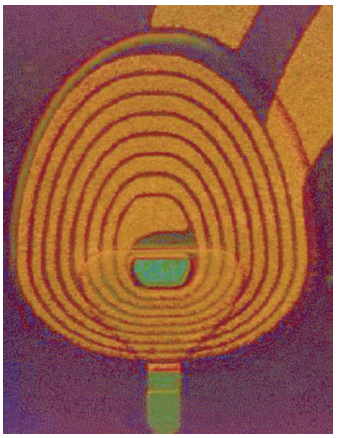
Whereas PCBs use electroless or electroplated Cu, contacts and connectors use electroplated Au and Au alloys or alternatively low-cost precious metal substitutes such as Pd, PdNi, PdAg, NiP, NiAs, and NiB with a thin gold overcoat. Separable, low-force, low-voltage contact applications require a contact finish whose resistance is stable for the projected contact life and which has sufficient wear resistance to withstand the projected number of insertions. The surface of one of the two mating parts is usually pure soft gold, whereas the other is a hard gold alloy. In less-critical applications, tin and tin-lead alloys are used as contact materials. Connecting chips with chip carriers or packages also often involves Au contacts. Gold does not adhere directly to silicon dioxide or silicon nitride, and Ti is typically used as an adhesion layer. Because Au and Ti form intermetallics, a Pt barrier is usually deposited between titanium and gold.

Copper Electroplating for Thin Film Read-Write Heads In one of many pioneering IBM MEMS efforts, Romankiw used plating through-mask technology to make thin film read-write heads as shown in Figure 8.39.[62] The development of batch-produced, thin film heads is an excellent example of how micromachining has been gainfully practiced in industry—with a definite, practical, well-specified application in mind and using the machining tools best suited for the problem.

Read-write heads, which initially were horseshoe magnets hand wound with insulated copper wire, are of utmost importance in magnetic storage. The fabrication of traditional ferrite heads reached its limit more than 20 years ago, and its further extension was hard to imagine. IBM's objective was to build the next-generation read/write heads using batch fabrication and lithography techniques as much as possible. To develop a multiturn head as shown in Figure 8.39a, it was necessary to develop a technology that would handle dimensions between those of PCBs and semiconductor devices. Plating of copper conductors through thick resist masks (Figure







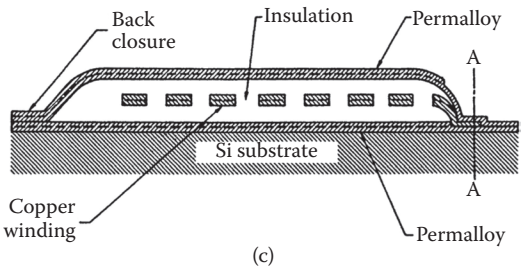


FIGURE 8.39 Thin film read-write head. (a) Schematic of a multiturn (32) thin film read-write head with inset of pole tip structure on the air-bearing surface. (b) Resist mold and electrodeposited copper coil after resist removal. (c) Schematic cross-section of an eight-turn thin film head. (Romankiw, L. T., I. M. Croll, and M. Hatzakis. 1970. Batch-fabricated thin-film magnetic recording heads. *IEEE Trans Magn* 6:597–601.[62])

8.39b) and of thin films of nickel-iron alloys of 80:20 nominal composition (Permalloy) (Figure 8.39c) made these thin film heads possible. The copper coils in Figure 8.39b carry considerable current and have to be at least 2–3 μm thick and nearly square in cross-section. The more turns in a given length, the higher the writing field at the pole tip gap and the higher the read back signal. The head in Figure 8.39c has eight turns; the 3380-K IBM head has 32 turns (Figure 8.39a). The Permalloy plating in the

thin film head is particularly challenging. Because of the very high sensitivity of composition to agitation of the plating solution, a plating cell had to be developed that assured uniform agitation over the entire part. This was achieved by a specially shaped paddle, which moves at a predetermined frequency with a precisely defined separation from the surface being plated. With the right choice of conditions, it was eventually possible to achieve both thickness and composition uniformity better than ±5%.[63] As a dielectric, hard-baked AZ photoresist was used, greatly simplifying the fabrication process. As a substrate, Si was used in the early days; today, Al_2O_3TiC is used.

Without these new micromachined thin film heads, the large increases in the areal densities* of magnetic media in the past few years would not have been possible. According to the data storage market research firm Peripheral Research Inc., market demand for thin film heads was projected to grow nearly 75% from 913 million units in 1998 to more than 1.5 billion units by 2001. This feat makes it probably the most successful micromachined device to date.

As the need for still more memory capacity increased, thin film inductive heads were replaced by newer technologies such as magnetoresistive (MR) heads (introduced in 1991 by IBM) and giant magnetoresistive (GMR) heads (introduced in 1997; see Volume III, Chapter 1). In MR heads, the read-back function is now performed by an MR sensor. The inductive portion of the head is used only for the write portion of the data-recording operation.[64]

In Volume III, Chapter 10, we learn that it is projected that read-write heads still will represent around 51% of the total MEMS market in 2009. Whereas traditional application of read-write heads in PCs will grow only moderately, the read-write head market is experiencing a renaissance in consumer electronics as hard disks are entering music players (e.g., in every iPod), smart phones (Samsung introduced the first cell phone with HDD in 2004), as well as digital video cameras, set top players, and DVD recorders. Of course, even further down the road we can envision all HDDs may be replaced by flash memory that can support today already up to 8-GB disk space, which is 5600 times more than a 1.44-MB floppy disk (see Volume III, Chapter 10)!

Copper Electroplating in Printed Circuit Boards

A PCB is made of layers, typically 2–10, that interconnect components via copper pathways. The main PCB in a system is called a *system board* or *motherboard*, whereas smaller ones that plug into the slots in the main board are called *boards* or *cards*. A typical multilayer PCB (which constitutes most of the PCBs used to implement complex circuitry) consists of a sandwich of conducting and insulating layers. Today's PCBs may consist of as many as five to six layers of metals and dielectrics. In Figure 8.40, a typical six-layer PCB is illustrated. By having the signals running on the inside of the board, more components can be more tightly integrated on the board to give a more compact design. PCB fabrication uses either an all electroless copper process or a combination of electroless and electroplating processes. The continuing effort to make active devices smaller and faster, to minimize the length of connecting wire, and to make them narrower and thicker (with smaller spaces between) necessitates more layers per board and smaller diameter holes with a larger ratio of hole length to diameter. A hole length-to-diameter ratio of 10:1 is routine and 20:1 is feasible. Without electrochemical copper plating technology, such a degree of integration would not have been possible.

Single and Dual Damascene Processes Copper wiring on IC chips was introduced by IBM in September 1997 with the "damascene process."[65]

* Areal density is used as a primary technology growth rate indicator for the memory storage industry. Areal density is defined as the product of the linear bits per inch (BPI) measured along the length of the tracks around a disk, multiplied by the number of tracks per inch (TPI) measured radially on the disk. The results are expressed in units of Mbit per square inch (Mbit/in.^{2}). Current high-end 2.5-in. drives record at areal densities of about 1.5 Gbit per square inch (Gbit/in.^{2}). Prototype drives with densities as high as 10 Gbit/in.^{2} have been constructed, allowing for capacities of more than 20 GB on a single 2.5-in. platter for notebook drives. Areal density (and, therefore, drive capacity) has been doubling approximately every 2–3 years, and production disk drives are likely to reach areal densities of 10 Gbit/in.^{2} before the end of 2001. A drive built with this technology will be capable of storing more than 10 GB of data on a single 2.5-in. platter, allowing 20- or 30-GB drives to be constructed that fit in the palm of your hand. The primary challenge in achieving higher densities is manufacturing drive heads and disks to operate at closer tolerances, which clearly is an area in which MEMS could have an impact.

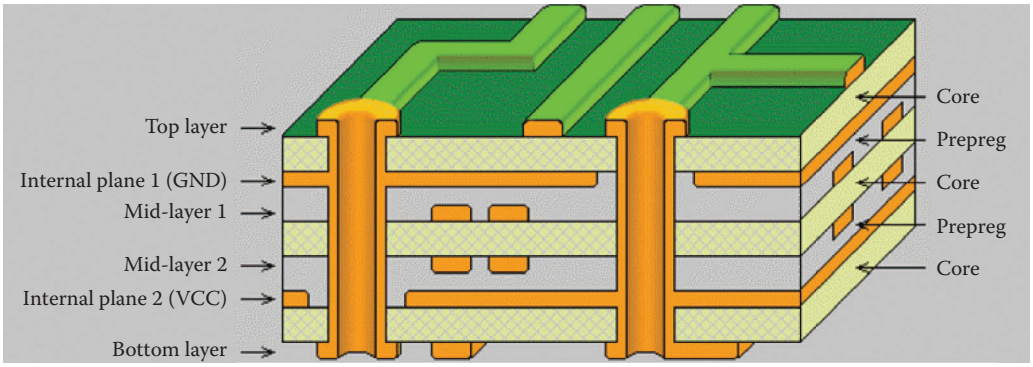


FIGURE 8.40 A typical six-layer printed circuit board. VCC stands for the positive supply voltage applied to the collectors of a transistor circuit. A "prepreg" is an epoxy-coated glass fabric.

The expertise with copper plating at IBM dates back to the late 1960s at Watson, when Romankiw succeeded in electroplating narrow wires of copper onto thin film read-write heads for memory, using a masking method that deposited the copper only in circuit patterns where it was needed (see above). Copper wires conduct electricity with about 40% less resistance than traditional aluminum wires, leading to faster microprocessors. Copper wires are also less prone to electromigration that ultimately induces wiring failure. Called *damascene* in reference to the metallurgists in Damascus who produced the finest polished swords in medieval times, the technique was initially used to form "vias" linking separate layers of wiring in chips. In the original damascene patterning, the pattern of vias was first formed by etching a silicon dioxide or other insulator layer. The metal was deposited second, and the excess metal was removed by polishing. Electrodeposition was selected as the metal deposition method of choice despite the fact that some people thought the copper patterns would be filled with bubbles, or that the process was too "dirty." It turned out to have a faster rate of deposition, and the evenness of the copper film was better than in the case of electroless deposition. Also, copper CVD, an early favorite option, ran into other severe problems. To polish the copper at an acceptable rate while controlling corrosion, erosion, and other defects in the patterns, IBM pioneered a special chemical-mechanical process (CMP) that proved critical to the copper technology. The latter process is detailed in Chapter 1. The damascene process can also be applied to other good via and interconnect materials such as Al.

In Figure 8.41 we compare through-mask copper plating with the damascene process. Through-mask plating (Figure 8.41a) uses a masking material on top of the seed layer. Electroplating occurs only on those areas of the seed layer that are not covered by the mask. The masking material and the surrounding seed layer are subsequently removed. Damascene plating (Figure 8.41b), in contrast, involves deposition of the seed layer over a patterned material, which, in the case of interconnect structures, is the insulator, a functional part of the device that remains in place. The plated metal covers the entire surface, and excess metal must be removed by CMP.

Damascene electroplating it turned out is ideally suited for the fabrication of complex interconnect structures because it enables inlaying of metal in via holes and overlying lines at the same time in a process called *dual damascene*, as illustrated in Figure 8.42. In a dual-damascene structure as shown here, only a single metal deposition step is used to simultaneously form the main metal lines and the metal in the vias. That is, both trenches and vias are formed in a single dielectric layer. The vias and trenches are defined using two lithography steps. Trenches are typically etched to a depth of 4000–5000 Å, and the vias are typically 5000–7000 Å deep. After the via and trench recesses are etched, the via is filled in the same metal-deposition step that fills the trench. After filling, the excess metal that is deposited outside the trench is removed by a CMP process, and a planar structure with metal inlays is achieved. The process starts with depositing a thin layer (~250 Å) of silicon nitride (Si_3N_4) on the silicon surface and between layers of a low

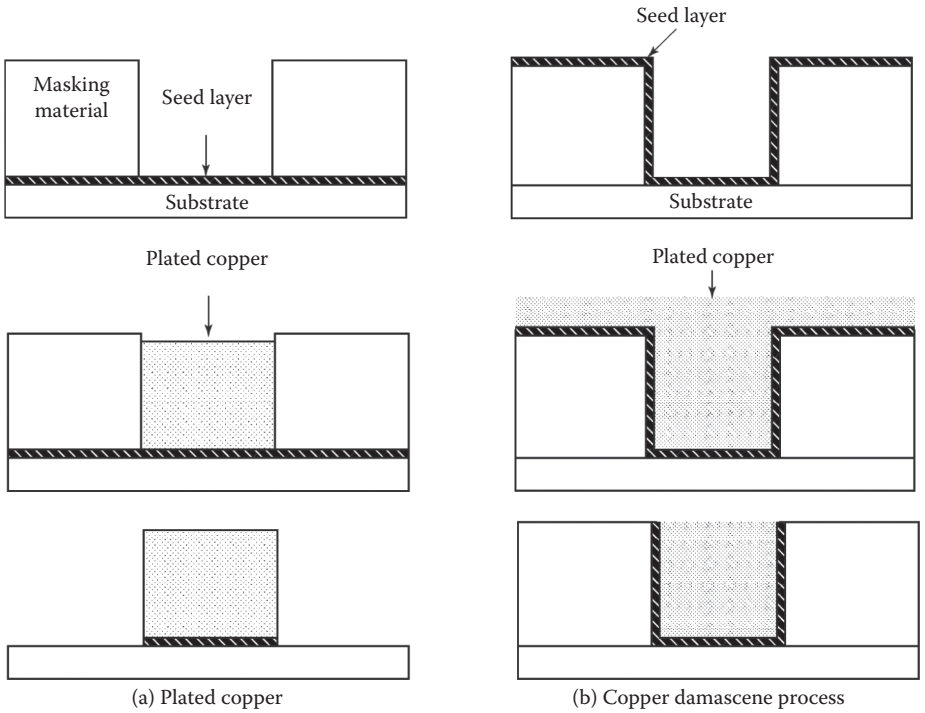


FIGURE 8.41 (a) Standard copper plating: seed layer deposited, mask layer deposited and patterned, copper plated up, mask layer removed, and seed layer etched away. (b) Damascene process used to obtain highly planar surfaces: dielectric layer (insulator) deposited and patterned, seed layer deposited, copper plated, and surface polished with CMP.

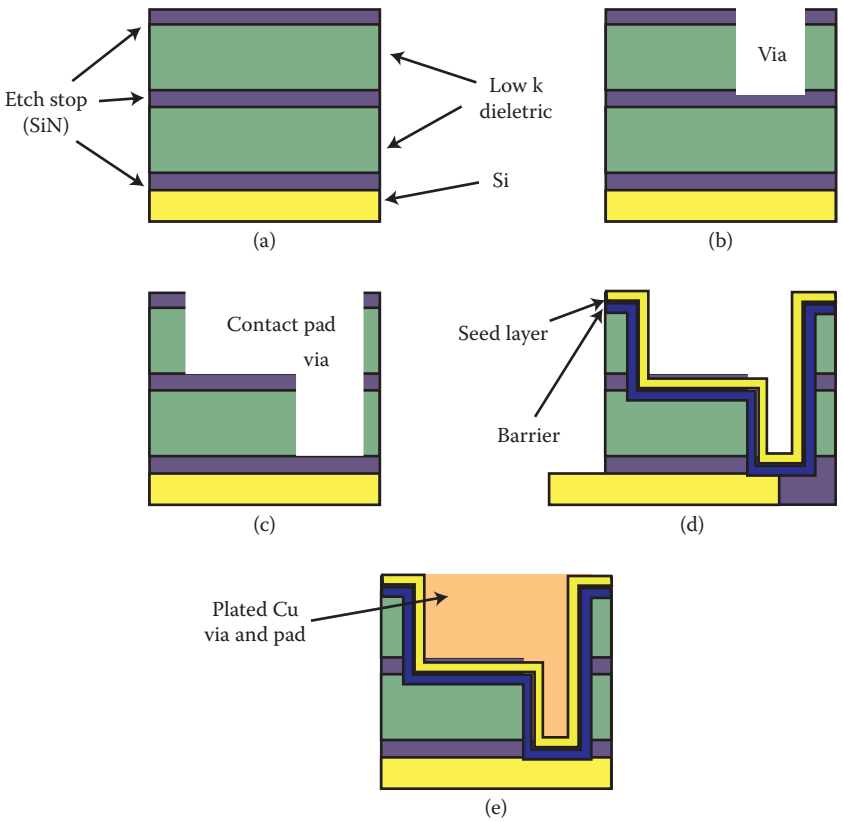


FIGURE 8.42 Dual-damascene process: process steps for a via and a line. (a) Insulator, low *k* dielectric deposition. (b) Via definition. (c) Line/contact pad definition. (d) Barrier and seed layer deposition. (e) Plating and CMP.

k dielectric, using CVD. The Si_3N_4 serves as an etch stop layer during via and trench formation. To overcome copper's tendency to diffuse in silicon, a metal diffusion barrier (e.g., Ta or TaN), preventing atoms from migrating out of a copper wire into surrounding chip material, is deposited next. Then follows a thin layer of copper deposited via CVD as a precursor to the subsequent electrodeposition of copper in both vias and trenches. The term "dual" refers to the formation of second channels, vias, within trenches. Further, it is compatible with the requirement for a barrier layer between the seed layer and the insulator.

As in conventional patterning, the damascene process is repeated many times to form the alternating layers of wires and vias that form the complete wiring system of a silicon chip. In Figure 8.43, an example copper dual-damascene structure is shown.

Anodization Anodization is an oxidation process performed in an electrolytic cell. The material to be anodized becomes the anode (+), whereas a noble metal is the cathode (–). Depending on the solubility of the anodic reaction products, an insoluble layer (e.g., an oxide, perhaps TiO_2) results, or in the case of a soluble reaction product, the electrode etches. The process can be either global or local through a masked area. If the primary oxidizing agent is water, the resulting oxides generally are porous, whereas organic electrolytes may lead to very dense oxides, providing excellent passivation.[66] Good oxides have been produced this way on W, Al, and Ta. In Volume I, Chapter 4, in the section on thermal oxidation, we learned that anodic oxidation of Si has not led to a commercially acceptable process, mainly because the interface state density at the SiO_2/Si interface is prohibitively high for IC applications. As a sacrificial layer, anodic SiO_2 plays a role in micromachining and leads to uniquely structured films that cannot be obtained by chemical means. For example, anodization of Si in a highly concentrated HF solution (excellent etchant for the anodic oxidation product SiO_2) leads to porous Si and very high-aspect-ratio pores, with diameters ranging from 20 Å to several micrometers. The growth rate and degree of porosity of the Si can be controlled by the current density (see Chapter 4). Similarly porous aluminum oxide films can be prepared by anodization of aluminum foils.[67]

Through an appropriate choice of electrolyte, films other than oxides can be produced by anodization. For example, sulfide and selenide layers can be produced by anodization of Cd and Zn electrodes, and polymerization of redox polymers can be induced from solutions containing the monomers.

FIGURE 8.43 Dual damascene copper. Note planarity of structure. (Courtesy of Motorola Inc.)

Conclusions on Electroless and Electroplating through Polymer Masks Electrochemistry through polymer masks represents one of the most powerful techniques available for formation of very high-density patterns and circuits with extremely large height-to-width ratios. Most other methods, such as chemical isotropic etching, sputter etching, ion milling, and reactive ion etching, cannot be used to produce metal patterns with better than 1:1 height-to-width aspect ratios. To understand this contrasting behavior, we must look into the fundamental nature of the different approaches. In nonelectrochemical technologies, a polymeric or inorganic mask is used to cover the existing metal, whereas either a chemical solution or an active gas phase removes the metal from the open areas. Hence the limitation in the height-to-width aspect ratio, the maximum circuit density, and the fidelity of reproduction of the features are caused by shadowing, which is the mechanism by which the patterning takes place. In contrast, with electrochemical techniques, one

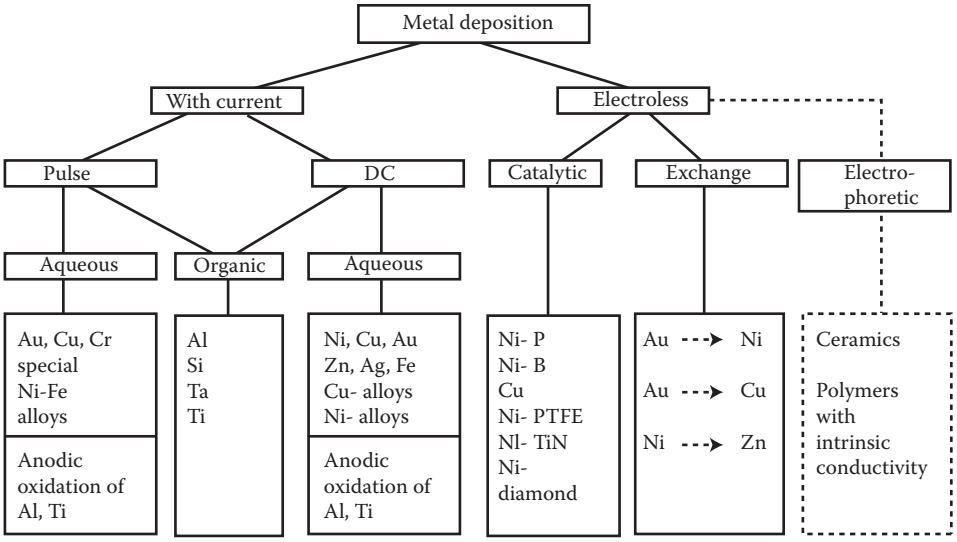


FIGURE 8.44 Processes and materials for through-mask metal deposition. (From Ehrfeld, W. 1994. Notes from handouts. Banff, Canada. With permission.[68])

obtains the exact replication of the recesses in the mask.[52] Thus, electrodeposition reproduces the finest features of the mask with the greatest fidelity. This is not surprising, considering that, in plating, metal ions discharge from solution present inside the mold. In so doing, the metal displaces the solution from the mold atom by atom, conforming to the smallest features that exist in the mold.[51] A comparison of conventional subtractive etching with electrochemical additive processes was presented as early as 1973 by Romankiw et al.[61] For the fabrication of highly miniaturized magnetic bubble memory devices, this IBM group showed clearly that the additive plating process is superior to dry etching-based techniques.

Figure 8.44 features a list of choices for material deposition using electrochemical and electroless techniques.[68] The first metal LIGA structures (see Chapter 10) consisted of nickel, copper, or gold electrodeposited from suitable electrolytes.[69] Nickel-cobalt and nickel-iron alloys were also experimented with. A nickel-cobalt electrolyte for deposition of the corresponding alloys has been developed especially for the generation of microstructures with increased hardness (400 Vickers at 30% cobalt) and elastic limit.[70] The nickel-iron alloys permit tuning of magnetic and thermal properties of the crafted structures.[70,71] From Figure 8.44 it is obvious that many more materials could be combined with through-mask metal deposition.

Instant Masking in Electrochemical Fabrication

Instant masking is a selective electroplating technique similar in some ways to through-mask plating. However, through-mask plating involves up to 10 separate steps, multiple pieces of equipment, multiple liquids, and is time consuming and difficult to automate. Obviously it is not a technique that is very suitable for automatically building complex 3D microdevices consisting of hundreds of layers or more. On the other hand, instant masking operates much more like printing, in which prefabricated plates transfer ink onto a substrate. In instant masking, instead of inks, one deposits materials electrochemically.

Instant masking is the enabling technology of electrochemical fabrication (EFAB), a solid, free-form fabrication technology that creates complex, miniature 3D metal structures impossible or impractical to manufacture using other MEMS technologies, such as electrical discharge machining (EDM), laser machining, or silicon micromachining.[72] The automated EFAB process creates metal structures by electroplating multiple, independently patterned layers. The process is similar in concept to rapid prototyping (RP) techniques, such as stereolithography (SL), in that multiple patterned layers are stacked to build structures. However, unlike stereolithography, EFAB is a batch process suitable for volume production of fully functional devices, not just models and prototypes.

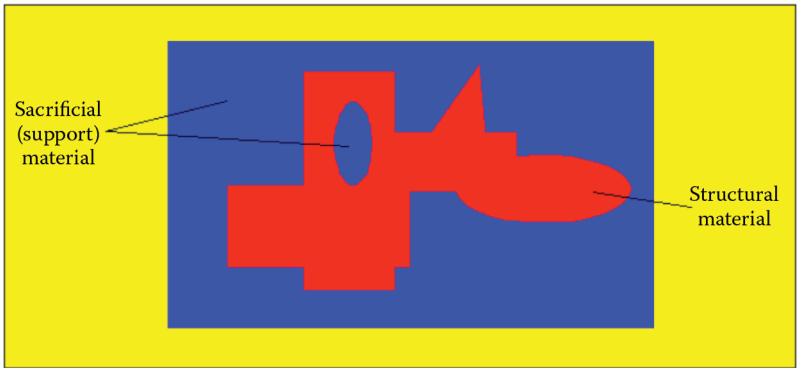


FIGURE 8.45 EFAB layer example. This 5-μm-thick layer consists of a structural material (e.g., Ni) and a sacrificial (support) material (e.g., Cu). (http://www.isi.edu/efab.)

Furthermore, EFAB provides greater accuracy than stereolithography. The EFAB process was invented at the University of Southern California with funding from the Defense Advanced Research Projects Agency and is commercialized by MEMGen Corp., a company founded in August 1999 (MEMGen Corp. changed its name to Microfabrica in 2003; http://www.microfabrica.com).

The three major fabrication steps in EFAB that are repeated are 1) the instant masking and selective electroplating, 2) blanket deposition, and 3) planarization. The manufacturing process starts with a substrate and grows devices layer by layer. Each thin layer, perhaps 5 μm thick, consists of a structural material (e.g., Ni) and a sacrificial material (e.g., Cu) as shown in the example in Figure 8.45. Microdevices are built by stacking many of these layers (as many as desired). To fabricate a multilayer device, the geometries of the layer cross-sections are automatically determined based on the desired 3D geometry, and a set of instant masks is generated that includes all the unique cross-sections of the device. The device under construction, composed of structural material, is imbedded within the sacrificial material, which, like a scaffold, supports it during fabrication and is later removed. Additional materials can be deposited over entire layers without constraint. Such geometrical freedom makes possible monolithically fabricated assemblies of discrete, interconnected, and rotating parts and eliminates the need for subsequent bonding or assembly steps.

The instant mask consists of an insulator pattern, made, for example, by patterning a resist layer on an anode plate using photolithography. A series of such masks, each representing a thin cross-section through the part being built, is fabricated in a separate process before the formation of the part. Instant masking then patterns a substrate by pressing this insulator pattern on the anode against the cathode substrate (e.g., Ni), and the assembly is immersed in an electrochemical bath (e.g., a Ni bath). Electroplating from the nickel salt solution between anode and cathode only deposits metal (Ni) between the insulator parts on the anode. The result is a layer that has been rapidly deposited and patterned in a single step. After the nickel plating, the mask is removed, and the substrate with its nickel pattern is copper plated, filling in areas left by the insulator. The two materials are then polished flat, providing a smooth layer for the next mask. The process can be repeated indefinitely, which creates the potential for defining arbitrarily complex shapes. Once the final plating has taken place, the sacrificial copper is removed by chemical etching, leaving a 3D nickel structure.

The complete process flow is shown in Figure 8.46. In Figure 8.46a, the first material is patterned onto a substrate, producing a patterned layer. In Figure 8.46b, the second material has been blanket deposited over the first material so that it contacts the substrate in those regions not covered with the first material. Then, as shown in Figure 8.46c, the entire two-material layer is planarized to achieve precise thickness and flatness. After repetition of this process for all layers, the embedded multilayer structure shown in Figure 8.46d is etched to yield the desired device as shown in Figure 8.46e.

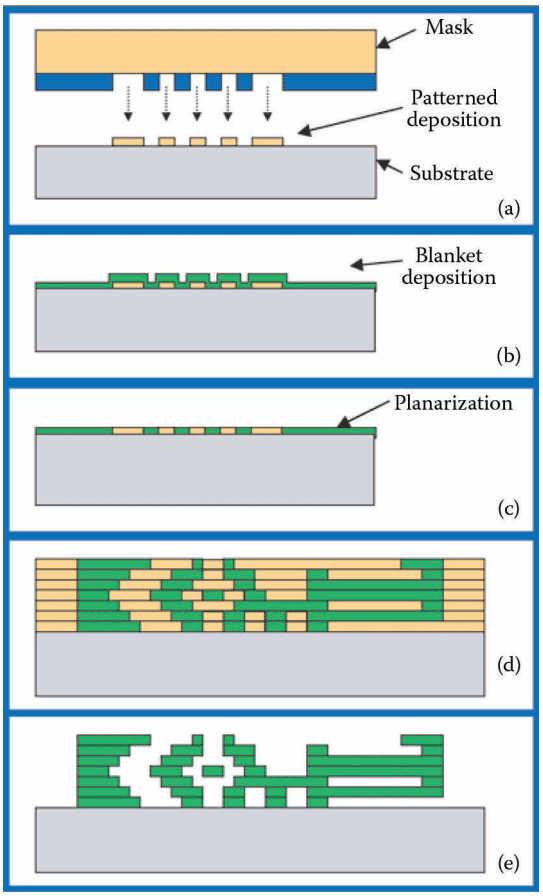


FIGURE 8.46 The EFAB process. (a) First material is patterned onto a substrate, producing a patterned layer. (b) Second material is blanket deposited over the first material. (c) The entire two-material layer is planarized. (d) Repetition of the process steps (a)–(c). (d) Embedded multilayer structure. (e) Sacrificial material is etched to yield the desired device as shown.

The instant masking process is fast, which makes it possible to fabricate devices with a dozen or more completed layers in 1 day.

The precision EFAB technology allows miniature and microdevices to be generated from 3D computer-aided design (CAD) data. The system uses data from any standard CAD package to determine device cross-sections. These cross-sections are then used to fabricate the instant masks that are then used to rapidly create the devices. EFAB is a precision manufacturing technology that can be used to form structures from any metal or alloy that can be electrodeposited. The only constraint is that the accompanying sacrificial metal can be selectively etched after the layers are formed. EFAB, depending on the lithography used to make

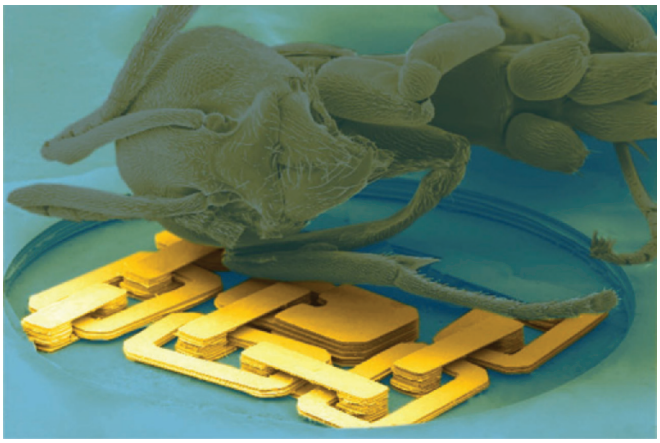


FIGURE 8.47 A small household ant lies over a 12-layer microchain with independently movable links, fabricated in nickel in a single process without the need for assembly. The chain is about ~100 μm high (about the thickness of a sheet of paper; http://www.isi.edu/efab/Resources).

the instant masks, can produce parts with feature sizes less than 0.001 in. and tolerances better than 0.0001 in. Such precision is difficult to obtain with mechanical precision machining technology. Like a machine tool, the EFAB process is suitable for a wide range of applications (see example microchain in Figure 8.47). Applications include radiofrequency (RF) electronic components such as high-Q inductors and transmission lines, optical components, microfluidic networks, and mold tooling for plastic devices.

Maskless Plating

Microjet Plating and Laser-Enhanced Jet Plating Through-mask electrochemical deposition, although a very important MEMS technique, is costly and requires extra steps, and several electrochemical deposition techniques have been proposed that obviate the need for masks.

One of the simpler approaches to avoid using masks in electrodeposition uses impinging microjets of electrolyte. This increases the mass transport greatly, and plating selectivity is achieved by virtue of the fact that the jet serves as the current path. The smallest spot size of the area being plated is usually limited to twice the diameter of the jet, which can be operated, if filtering the solution carefully, without too frequent plugging. The nozzle orifice diameter may be anywhere from 1 mm to as little as ~0.01 mm in diameter. The lower limit depends strongly on the

degree of filtration of the solution. Micromachining of nozzles will further impact obtainable microjet-plating resolution in the future.

The setup in Figure 4.15 for laser-enhanced jet etching can also be used for faster plating. The local heating of the substrate by the laser (up to 150°C) results in highly increased deposition. Using this technique, plating enhancement of up to 1000 times has been obtained with gold. Plating spots and lines as small as 2 μm wide have been observed. Laser enhancement also works with electroless deposition. Current densities of up to 15 A/cm^2 and copper plating rates of 50 μm/s have been obtained. For more background on jet plating and laser-enhanced jet plating, refer to the review by Romankiw et al. and references therein.[51]

Pulse plating, introduced in Chapter 10 in the context of plating in LIGA molds, represents only one of many emerging electrochemical plating techniques considered for microfabrication. For more background on techniques such as laser-enhanced plating, jet plating, laser-enhanced jet plating, and ultrasonically enhanced plating, refer to the review of Romankiw et al.[51] and references therein.

Localized Electrochemical Deposition Truly 3D microstructures can be formed using localized electrochemical deposition. The tip of a sharp electrode placed in a plating solution is brought near the substrate where deposition occurs when applying a potential between the tip and the substrate. The electric field is confined to the area beneath the tip, and the spatial resolution is determined by the size of the electrode. Deposition rates of Ni are as high as 6 μm/s, two orders of magnitude higher than in conventional electroplating.[73] The tip of a scanning electrochemical microscope (SECM) can be used for both local etching and deposition with high resolution in the x-, y-, and z-dimensions, basically forming a high-resolution electrochemical machining setup (Figure 4.16).

Slanted and Curved Plated Metal Shapes As a final example of maskless electrodeposition, we consider the work by Maciossek, who relies on overplating of a patterned plating base to create a variety of novel slanted and curved metallic microstructures.[59] The angle of the deposited wedge shapes is adjustable in a range from 0–45°. Besides linearly increasing structure profiles, a sinusoidal or parabolic surface can also be generated. The angles of the deposited wedges depend on the distance between and the width of a set of parallel metallic strips on the plating substrate as illustrated in Figure 8.48A. At the start of the plating process, only a central metal strip is biased, and during the isotropic growing process, the outlying metal strips are contacted one at a time by the electrodepositing metal. When the contact has been made, electrodeposition will occur also in this area, whereas further deposition in the original contact area (the central strip) results in a higher deposit on that first electrode. This innovative 3D fabrication process enlarges the MEMS arsenal of tools. The Maciossek method may be used to fabricate slanted and curved absorber patterns for x-ray masks (see Chapters 2 and 10). Two example Maciossek structures are shown in Figure 8.48B.

Three-dimensional metal structures on a substrate may also be produced by electrodeposition through a suitably prepared gelatin layer in a process similar to photolithography. Angus et al.[74] exposed a layer of gelatin on a Cu surface through a gray mask. Cross-linking of the gelatin occurs in the exposed areas, and unexposed material is washed away in a water/isopropanol solution. Thickness of the hardened gelatin layer, and hence the rate of diffusion of depositing metal ions through it, depends on the exposure dose. Metal electrodeposition through the developed gelatin layer film occurs more rapidly at the less exposed sites, thus generating a 3D metal structure on the substrate with a thickness and configuration governed by the exposure time, the mask thickness, and the deposition time.[75] As the mask is permeable to the plating solution, we treat this technique here as a "maskless" technology.

Example 8.1: Protein Patterning

Microlithographic techniques can be applied successfully to the field of protein patterning. In this first example, we show how clever use of lithography may help solve the problem

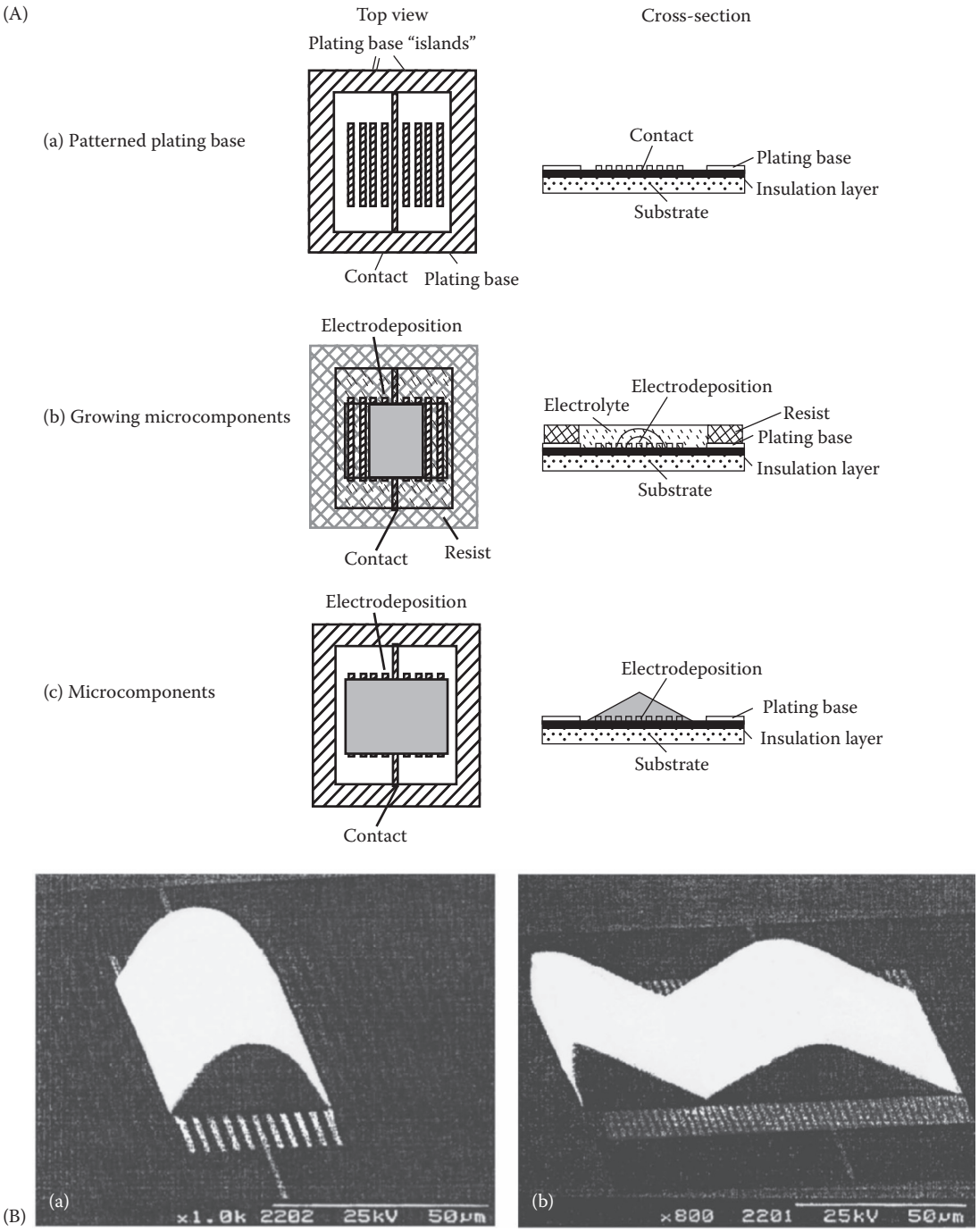


FIGURE 8.48 (A) Maciossek plating schematic: exposing the patterned plating base (Aa) to a plating solution, the isotropic growing process starts at the contact area and extends until the neighboring plating base islands are contacted. (Ab) Once contacted, electrodeposition also occurs in this area, whereas further deposition in the former contact area results in a higher structure there. On continuing the electroplating process, consecutive outlying "island" electrodes are reached, and electroplating starts later so that wedge-shaped microcomponents are formed (Ac). (B) Electroplated profiles show a very smooth and bright surface. The perpendicular sidewalls are made by a thick surrounding resist. (Ba) Gold microstructure, 18 µm height; (Bb) electroplated gold with two electrical contacts.

of nonspecific protein adsorption competing for detection sites in immunosensors. Proteins bind with considerable avidity to a wide range of surfaces, and a reference surface in an immunosensor should adsorb all the same proteins except for the protein of interest. The best way to avoid nonspecific binding effects in immunosensors is to make the reference surface as similar to the sensing surface as possible, i.e., a reference substrate, subject to all the same

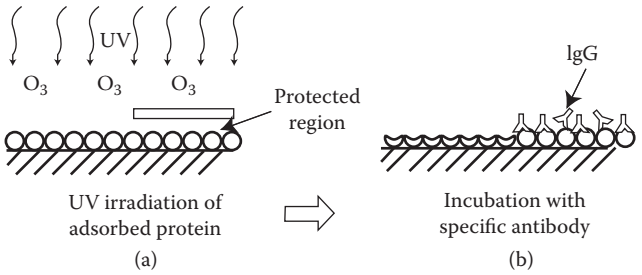


FIGURE 8.49 UV radiation of antigens leads to bands of "live/dead" protein. (a) Preadsorbed layer of antigen is briefly exposed, in air, to an intense UV source, resulting in the production of ozone (O_3). The combination of UV damage and ozone results in partial oxidation of the adsorbed protein. The area of protein shielded from the irradiation is not oxidized. (b) Following this treatment the protein-coated substrate is incubated with antiserum specific to the adsorbed antigen. The partially oxidized protein layer is no longer antigenic, and IgG molecules are able to bind only to the previously shielded portion of the antigen layer. (Clementi, E., G. Corongiu, M. H. Sarma, and R. H. Sarma, eds. 1985. *Structure and motion: membranes, nucleic acids, and proteins.* Guilderland, NY: Adenine Press.[76])

nonspecific protein binding phenomena as the sensor surface itself, except for the antigen-antibody of interest. This way, measuring differentially enables the best possible correction for nonspecific binding. An interesting way of implementing this idea is to create an optical grating pattern, with the antigen-antibody coupling providing the grating structure. This can be realized, for example, based on the loss of antigenicity on UV radiation of thin layers of antigen as observed by Panitz and Giaver.[76] Panitz and Giaver found that the antigenic sites on proteins display sensitivity to UV radiation in air and that a small dose of UV radiation can destroy the antigenicity of antigens (Figure 8.49). One presumes that the UV light breaks up chemical bonds in the adsorbed protein layer and simultaneously produces ozone. The broken bonds are targets for the highly reactive ozone, resulting in the progressive oxidation of the protein layer. The protein can be completely removed from the surface by too long an irradiation (e.g., 10 min). With a short radiation time the optical density does not change, but antibodies will no longer bind to the antigenetic sites. This scheme was put to use in the fabrication of an elegant biograting for immunosensing as shown in Figure 8.50.[77] By using a photomask grating and inactivating alternating bands of antibodies, a biological diffraction grating is created. In this device, the antigen-antibody coupling itself constitutes the grating structure, causing diffraction only if the target antigen is present. A CD-type He-Ne laser beam diffracts from the grating with an intensity related to the antigen concentration. The above grating device comes close to embodying an ideal relative reference because the radiated protein bands are almost identical to the active surface except for being incapable of reacting with the target complementary protein.[78]

It is easy to imagine arrays with several different enzymes, antibodies, or DNA probes immobilized precisely onto a transducer surface as a diagnostic panel for clinical diagnosis, high-throughput drug screening, or for environmental monitoring.[79]

Example 8.2: The Nanogen Chip

A schematic representation of an open electrophoretic DNA hybridization chip from Nanogen is shown in Figure 8.51a. In this chip DNA, which is negatively charged at most pHs, is collected at biased electrodes through electrophoresis. An array of 25 metal electrodes with a diameter of 80 μm is pictured. Arrays of

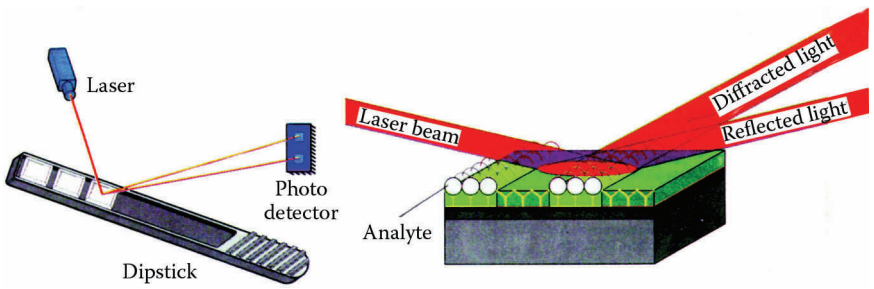


FIGURE 8.50 Idetek, Inc. immunosensor based on biological grating (Editorial. 1993. Biograting. *Res Dev* 51; Courtesy of Mark Platshon, Idetek, Inc., Sunnyvale, CA.[77])

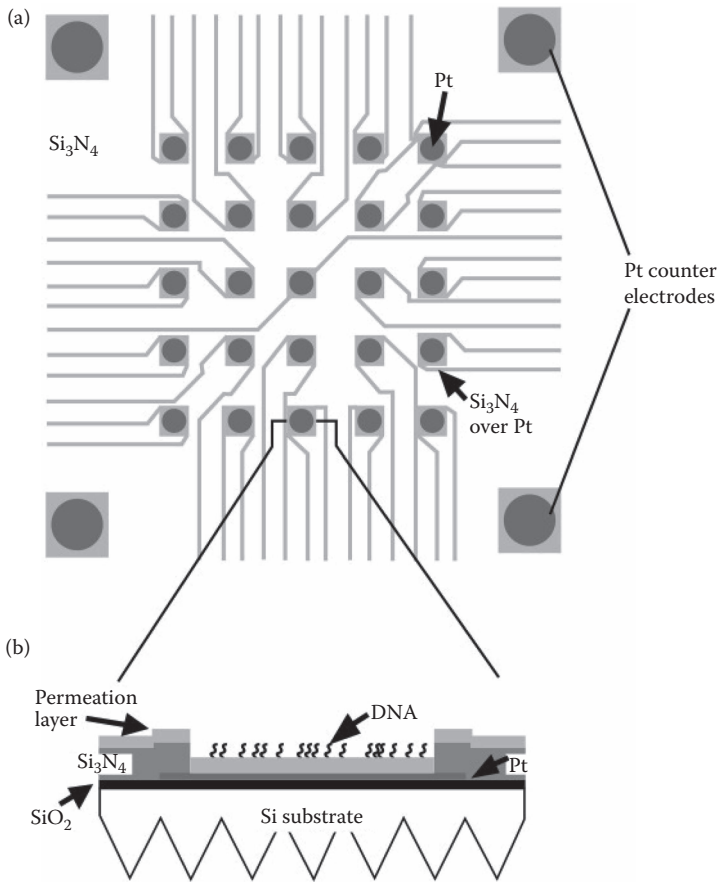


FIGURE 8.51 Nanogen DNA chip. (a) Electrode array region of the chip. The central 1 × 1-mm test site array region consists of four large (160-μm diameter) corner electrodes and 25 central 80-μm electrodes. Pt, exposed Pt test sites; Si_3N_4, dielectric; Si_3N_4 over Pt, and Pt insulated by Si_3N_4. (b) Cross-section of an electrode test site. Location of section is indicated by lines extending from (a). Pt, Si_3N_4, Pt insulated by Si_3N_4; SiO_2, dielectric layer; Si substrate, wafer material; permeation layer, agarose layer containing streptavidin; DNA, biotinylated oligonucleotides bound to streptavidin.

100 (80-μm diameter), 400 (50-μm diameter), and 10,000 (30-μm diameter) elements have been made as well.[80] Each of the metal electrode sites (microlocations) is a discrete, insulated, and addressable working electrode forming part of an electrochemical cell. An outer group of counter electrodes provides encompassing electric fields for concentrating DNA from the bulk sample solution to specific working electrodes. In Figure 8.51a, there are four such outer electrodes with a diameter of 160 μm. The microlocations are all electrically connected via the overlying electrolyte solution and protected with a permeation layer hydrogel (e.g., agarose or polyacrylamide) (Figure 8.51b). The permeation layer is a 1–10-μm hydrogel layer deposited by spin coating and permits water and ion flow, provides a matrix to attach molecules such as DNA, RNA, or proteins, and separates those sensitive biomacromolecules from the potentially damaging electrochemical reactions that occur on the active Pt electrodes (current levels >100 nA and voltages >1.2 V are possible with such hydrogel layers without damaging the biomacromolecules). Attachment of molecules to the permeation layer such as DNA probes can be achieved, for example, by impregnating the permeation layer with affinity binding substances such as avidin or streptavidin for subsequent attachment of biotinylated DNA or RNA probes.

Net negatively charged molecules such as DNA and RNA are moved to positively biased microlocations, whereas they are repelled from negatively biased ones by electrophoresis. For the electric field to be high enough throughout the solution for electrophoresis to take place, the solution must be desalted. Only low conductivity buffers may be used. The positive microlocations may concentrate target DNA sequences in a

very short time (because they reach those sites through electrophoresis rather than by diffusion), and if this location has complementary DNA capture probes, hybridization of the target DNA can occur. The concentrating effect on the positive microlocations facilitates the hybridization as a result of the law of mass action (hybridization in seconds rather than hours). Details of DC current and voltage level, solution conductivity, and buffer species for hybridization can be found in Edman et al.[81] This directed electrophoretic transport and addressing process can be carried out simultaneously at test sites that have different capture sites. Detection of successful hybridization can be accomplished using fluorescent probes and a color, charge-coupled detector (CCD). By reversing the polarity at a microlocation after hybridization, nonspecific sample DNA and unhybridized probes may be selectively removed. The electric field can be adjusted just so to affect selective dehybridization of the DNA sequences from the attached complementary probe. This is called *electronic stringency control.* Discrimination between single base pair mismatches in DNA has been demonstrated this way;[82] for example, Gilles et al. developed a rapid assay for single-nucleotide polymorphisms (SNPs) in a gene implicated in increased susceptibility to infection in pediatric patients.[83]

The power of this open-chip approach was further demonstrated by separating *Escherichia coli* from a mixture containing blood cells by means of dielectrophoresis, lyzing the isolated bacteria by a series of high-voltage pulses, and proteolytic digestion with proteinase K—all on the same chip. The microlocations on the Nanogen chip can be powered with DC and AC.[84] Even more exciting is the demonstration of amplification and detection of multiple targets in this same open format [multiplex strand displacement amplification (SDA)]. SDA is an isothermal (60°C) DNA amplification method. In a regular multiplex amplification in solution, nonspecific interactions between different primer sets reduce the amplification efficiency. Electronic anchoring of sets of amplification primers in distinct areas reduces their interaction, and sets of distinct zones of amplification that only share reagents and enzymes are generated, increasing the efficiency of the multiplex amplification reactions. Isothermal SDA, which uses the combined effects of a restriction endonuclease and DNA polymerase, asynchronously amplifies DNA exponentially and was shown to be ideal for use on the open electrophoresis chip.[85]

Questions

8.1: Describe two methods to make DNA arrays using a virtual mask approach.

8.2: Briefly sketch the sol-gel technique. Why is this technique so important for nanotechnology?

8.3: Compare Langmuir-Blodgett deposition with self-assembled monolayers (SAMs).

8.4: Sketch the dual damascene process. Where is it used?

8.5: Outline the EFAB process. Where is it used?

8.6: Why is DNA hybridization so much faster in an electronic DNA (Nanogen) than in a passive hybrydization chip?

8.7: Describe how you could make beveled metal structures using both electroless plating and electroplating.

8.8: Describe briefly different types of rapid prototyping (RP) and discuss how they can be scaled down. Which RP technique is of most interest to micromachinists?

8.9: Compare three major arraying methods for making DNA and protein arrays.

8.10: Detail the difference between scanning microstereolithography, parallel microstereolithography, and two-photon stereolithography.

References

1. Rath, K. 2005. Novel materials from solgel chemistry. *Science Technol Rev*, Lawrence Livermore National Laboratory, May 13, 2005.
2. Licari, J. J. 1970. *Plastic coatings for electronics.* New York: McGraw-Hill.
3. Howatt, G. N. 1952. *Method of producing high-dielectric high-insulation ceramic plates.* US Patent 2582993.
4. Jenkins, A. T. A., J. Hu, Y. Z. Wang, S. Schiller, R. Förch, and W. Knoll. 2000. Pulsed plasma deposited maleic anhydride thin films as supports for lipid bilayers. *Langmuir* 16:6381–84.
5. Bauer, M., T. Schwarz-Selinger, H. Kang, and A. v. Keudell. 2005. Control of the plasma chemistry of a pulsed inductively coupled methane plasma. *Plasma Sources Science Technol* 14:543–48.

6. Guckel, H., J. Uglow, M. Lin, D. Denton, J. Tobin, K. Euch, and M. Juda. 1988. Plasma polymerization of methylmethacrylate: a photoresist for 3-D applications. *Technical digest: 1988 solid state sensor and actuator workshop*. Hilton Head Island, SC.
7. Vossen, J. L., and W. Kern, eds. 1978. *Thin film processes.* Orlando, FL: Academic Press.
8. Biederman, H., ed. 2004. *Plasma polymer films.* London: Imperial College Press.
9. Acheson. 1991. Brochure. Scheemda, the Netherlands: Acheson Coloiden BV.
10. Pace, S. J., and M. A. Jensen. 1986. *Second international meeting on chemical sensors.* Bordeaux, France.
11. Belford, R. E., A. E. Owen, and R. G. Kelly. 1987. Thick-film hybrid pH sensors. *Sensors Actuators* 11:387–98.
12. Pace, S. J., P. P. Zarzycki, R. T. McKeever, and L. Pelosi. 1985. A thick-film multi-layered oxygen sensor. *Third international conference on solid-state sensors and actuators (Transducers '85)*. Philadelphia.
13. Karagounis, V., L. Lun, and C. Liu. 1986. A thick-film multiple component cathode three-electrode oxygen sensor. *IEEE Trans Biomed Eng* BME-33:108–12.
14. Lewandowski, J. J., P. S. Malchesky, M. Zborowski, and Y. Nose. 1987. *Proceedings: ninth annual conference of the IEEE engineering in medicine and biology society.* Boston.
15. Lambrechts, M., and W. Sansen. 1992. *Biosensors: microelectrical devices.* Philadelphia: Institute of Physics Publishing.
16. Weetall, H. H., and T. Hotaling. 1987. A simple, inexpensive, disposable electrochemical sensor for clinical and immuno-assay. *Biosensors* 3:57–63.
17. Cha, C. S., M. J. Shao, and C. C. Liu. 1990. Electrochemical behavior of microfabricated thick-film electrodes. *Sensors Actuators B* B2:277–81.
18. Chen, Y., J. Au, P. Kazlas, A. Ritenour, H. Gates, and M. McCreary. 2003. Electronic paper: flexible active-matrix electronic ink display. *Nature* 423:136.
19. Allara, D. L., and R. G. Nuzzo. 1985. Spontaneously organized molecular assemblies; II. Quantitative infrared spectroscopic determination of equilibrium structures of solution adsorbed n-alkanoic acids on an oxidized aluminum surface. *Langmuir* 1:52–66.
20. Allara, D. L., and R. G. Nuzzo. 1985. Spontaneously organized molecular assemblies; I. formation, dynamics, and physical properties of n-alkanoic acids adsorbed from solution on an oxidized aluminum surface. *Langmuir* 1:45–52.
21. Delamarche, E., B. Michel, H. A. Biebuyck, and C. Gerber. 1996. Golden interfaces: the surface of self-assembled monolayers. *Adv Mater* 8:719–29.
22. Liu, R. H., Q. Yu, J. M. Bauer, B.-H. Jo, J. S. Moore, and D. J. Beebe. 2000. *Micro total analysis systems 2000: proceedings, μTAS 2000 symposium.* Eds. A. van den Berg, W. Olthuis, and P. Bergveld. Enschede, the Netherlands: Kluwer.
23. Zubritsky, E. 2000. Microscale muscles. *Anal Chem* August:517–18.
24. Fodor, P. A. S., J. L. Read, M. C. Pirrung, L. Sryer, A. T. Lu, and D. Solas. 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251:767–73.
25. Britland, S. T., G. R. Moores, P. Clark, and P. Connolly. 1990. Patterning and cell adhesion and movement on artificial substrate: a simple method. *J Anat* 170:235–36.
26. Douglas, K., G. Deavaud, and N. A. Clark. 1992. Transfer of biologically derived nanometer scale patterns to smooth substrates. *Science* 257:642–44.
27. Singh-Gasson, S., R. D. Green, Y. Yue, C. Nelson, F. Blattner, M. R. Sussman, and F. Cerrina. 1999. Maskless fabrication of light-directed oligonucleotide micromirrors using a digital micromirror array. *Nat Biotechnol* 17:974–78.
28. Bertsch, A., J. Y. Jézéquel, and J. C. André. 1997. Study of the spatial resolution of a new 3D microfabrication process: the microstereophotolithography using a dynamic mask-generator technique. *Photochem Photobiol A: Chem* 107:275–81.
29. Okamoto, T., T. Suzuki, and N. Yamamoto. 2000. Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nat Biotechnol* 18:438–41.
30. St. John, P. M., R. Davis, N. Cady, J. Czajka, C. A. Batt, and H. G. Craighead. 1998. Diffraction-based cell detection using a microcontact printed antibody grating. *Anal Chem* 70:1108–11.
31. Madou, M., and T. Otagawa. 1992. *Tetrasulfonated metal phthalocyanine doped electrically conducting electrochromic poly(dithiophene) polymers.* US Patent 5,151,224.
32. Otagawa, T., and M. Madou. 1991. Permanently doped polyaniline and method thereof. US Patent 5,002,700, Osaka Gas Company.
33. Ikuta, K., and K. Hirowatari. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL.
34. Ehmann, K. F., D. Bourell, M. L. Culpepper, T. J. Hodgson, T. R. Kurfess, M. Madou, K. Rajurkar, and R. DeVor, eds. 2007. *Micromanufacturing: international research and development.* New York: Springer.
35. Ikuta, K., K. Hirowatari, and T. Ogata. 1994. *IEEE international workshop on micro electro mechanical systems (MEMS '94)*. Oiso, Japan: IEEE.
36. Takagi, T., and N. Nakajima. 1993. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL: IEEE.
37. Ikuta, K., T. Ogata, M. Tsubio, and S. Kojima. 1996. Development of mass productive micro stereo lithography. *Proceedings of IEEE International workshop on micro electro mechanical systems*. MEMS '96. San Diego, CA.
38. Takagi, T., and N. Nakajima. 1994. *IEEE international workshop on micro electro mechanical systems (MEMS '94)*. Oiso, Japan.
39. Ikuta, K., S. Maruo, and S. Kojima. 1998. New micro stereo lithography for freely movable 3-D micro structure – Super IH process with submicron resolution. *Proceedings of the 11th IEEE Workshop on micro electro mechanical systems (MEMS'98)*. Heidelberg, Germany.
40. Maruo, S., and S. Kawata. 1997. Two-photo-absorbed photopolymerization for three-dimensional microfabrication. *Proceedings of the IEEE micro electro mechanical systems workshop (MEMS '97)*. Nagoya, Japan.
41. Maruo, S., and S. Kawata. 1998. Two-photon-absorbed near-infrared photopolymerization for three-dimensional microfabrication. *JMEMS* 7:411–15.
42. Angelo, M. D., M. V. Chekhova, and Y. H. Shih. 2001. Two-photon diffraction and quantum lithography. *Phys Rev Lett* 87:013602.
43. Boto, A. N., P. Kok, D. S. Abrams, S. L. Braunstein, C. P. Williams, and J. P. Dowling. 2000. Quantum interferometric optical lithography: exploiting entanglement to beat the diffraction limit. *Phys Rev Lett* 85:2733–36.
44. Prasad, P. N. 2004. *Nanophotonics.* Hoboken, NJ: Wiley-Interscience.
45. Zhang, X., X. N. Jiang, and C. Sun. 1999. Micro-stereo lithography of polymeric and ceramic microstructures. *Sensors Actuators A* A77:149–56.

46. Taylor, C. S., P. Cherkas, H. Hampton, J. J. Frantzen, B. O. Shah, W. B. Tiffany, L. Nanis, P. Booker, A. Salahieh, and R. Hansen. 1994. A spatial forming—a three dimensional printing process. *Proceedings of IEEE MEMS '94*. Oiso, Japan.
47. Beluze, L., A. Bertsch, and P. Renaud. 1999. *Symposium on design, test, and microfabrication of MEMS and MOEMS*. Paris.
48. Bertsch, A., H. Lorenz, and P. Renaud. 1999. 3D microfabrication by combining microstereolithography and thick resist UV lithography. *Sensors Actuators A* 73:14–23.
49. Bertsch, A., P. Bernhard, C. Vogt, and P. Renaud. 2001. Rapid prototyping of small size objects. *Rapid Prototyping J* 6:259–66.
50. Sun, C., N. Fang, D. M. Wu, and X. Zhang. 2005. Projection micro-stereolithography using digital micro-mirror dynamic mask. *Sensors Actuators A* 121:113–20.
51. Romankiw, L. T., and T. A. Palumbo. 1987. Electrodeposition in the electronic industry. *Proceedings: symposium on electrodeposition technology, theory, and practice*. San Diego, CA.
52. Romankiw, L. T. 1984. Electrochemical technology in electronics today and its future: a review. *Oberflache-Surface* 25:238–47.
53. Mallory, G. O., and J. B. Hadju, eds. 1990. *Electroless plating: fundamentals and applications*. Orlando, FL: American Electroplaters and Surface Finishers Society (AESF).
54. Romankiw, L. T. 1976. Pattern generation in metal films using wet chemical techniques: a review. P*roceedings of the Symposium on Etching for Pattern Definition*. Ed. H. G. Hughes, and M. Rand. Pennington, NJ: The Electrochemical Society.
55. Ohno, I. 1988. Electroless deposition of metals and alloys. 172nd Meeting of The Electrochemical Society, Honolulu, HI. *ECS Proceedings, Vol. 88-12*. Eds. M. Paunovic, and I. Ohno. Pennington, NJ: ECS.
56. Harsch, S., W. Ehrfeld, and A. Maner. 1988. *Untersuchungen zur herstellung von mikrostrukturen grosser strukturhöhe durch galvanoformung im nickelsilfamatelektrolyten. KfK, Report No. 4455*. Karlsruhe, Germany. Kernforschungszentrum Karlsruhe.
57. van der Putten, A. M. T., and J. W. G. de Bakker. 1993. Anisotropic deposition of electroless nickel-bevel plating. *J Electrochem Soc* 140:2229–35.
58. van der Putten, A. M. T., and J. W. G. de Bakker. 1993. Geometrical effects in the electroless metallization of fine metal patterns. *J Electrochem Soc* 140:2221–28.
59. Maciossek, A. 1996. *SPIE: micromachining and microfabrication process technology II*. Austin, TX: SPIE.
60. Kenis, P. J. A., R. F. Ismagilov, and G. M. Whitesides. 1999. Microfabrication inside capillaries using multiphase laminar flow patterning. *Science* 285:83–85.
61. Romankiw, L. T., S. Krongelb, E. E. Castellani, J. Powers, A. Pfeiffer, and B. Stoeber. 1973. Additive electroplating technique for fabrication of magnetic devices. *Proceedings of the international conference on magnetics, ICM-73*., Vol. 6, pp. 104–111.
62. Romankiw, L. T., I. M. Croll, and M. Hatzakis. 1970. Batch-fabricated thin-film magnetic recording heads. *IEEE Trans Magn* 6:597–601.
63. Romankiw, L. T. 1989. *Proceedings: symposium on magnetic materials, processes, and devices*. Hollywood, FL.
64. Andricacos, P. C., and N. Robertson. 1998. Future directions in electroplated materials for thin-film recording heads. *IBM J Res Dev* 42:671.
65. Editorial. 1997. *Back to the future: copper comes of age. IBM Res* 35.
66. Madou, M. J., W. P. Gomes, F. Fransen, and F. Cardon. 1982. Anodic oxidation of p-type silicon in methanol as compared to glycol. *J Electrochem Soc* 129:2749–52.
67. Miller, C. J., and M. Majda. 1986. Microporous aluminium oxide films at electrodes. Part II: studies of electron transport in the Al_2O_3 matrix derivated by adsorption of poly (4-vinylpyridine). *J Electroanal Chem* 207:49–72.
68. Ehrfeld, W. 1994. LIGA at IMM. Notes from handouts. Banff, Canada.
69. Maner, A., S. Harsch, and W. Ehrfeldg. 1988. Mass production of microdevices with extreme aspect ratios by electroforming. *Plating Surface Finishing* March: 60–65.
70. Harsch, S., D. Munchmeyer, and H. Reinecke. 1991. *Proceedings: 78th AESF annual technical conference (SUR/FIN '91)*. Toronto.
71. Thomes, A., W. Stark, H. Goller, and H. Liebscher. 1992. Erste ergebnisse zur galvanoformung von LIGA mikrostrukturen aus eisen-nickel legierungel. *Symp. Mikroelektrochemie Friedrichsroda*. Germany.
72. Cohen, A., G. Zhang, F. Tseng, U. Frodis, F. Mansfeld, and P. Will. 1999. EFAB: rapid, low-cost desktop micromachining of high aspect ratio true 3-D MEMS. *Proceedings of the 12th IEEE micro electro mechanical systems workshop*. Orlando, FL, January 1999, p. 244.
73. Madden, J. D., and I. W. Hunter. 1996. Three-dimensional microfabrication by localized electrochemical deposition. *J Microelectromechan Syst* 5:24–32.
74. Angus, J. C., U. Landau, S. H. Liao, and M. C. Wang. 1986. Controlled electroplating through gelatin films. *J Electrochem Soc* 133:1152–60.
75. Bard, A. J. 1994. *Integrated chemical systems*. New York: Wiley.
76. Clementi, E., G. Corongiu, M. H. Sarma, and R. H. Sarma, eds. 1985. *Structure and motion: membranes, nucleic acids, and proteins*. Guilderland, NY: Adenine Press.
77. Editorial. 1993. Biograting. *Res Dev* 51.
78. Madou, M. J., and J. Joseph. 1993. Immunosensors with commercial potential. *Immunomethods* 3:134–52.
79. Connolly, P., G. R. Moores, W. Monaghan, J. Shen, S. Britland, and P. Clark. 1992. Microelectronic and nanoelectronic interfacing techniques for biological systems. *Sensors Actuators B* B6:113–21.
80. Heller, M. J., A. H. Forster, and E. Tu. 2000. Active microelectronic chip devices which utilize controlled electrophoretic fields for multiplex DNA hybridization and other genomic applications. *Electrophoresis* 21:157–64.
81. Edman, C. F., D. E. Raymond, D. J. Wu, E. Tu, R. G. Sosnowski, W. F. Butler, M. Nerenberg, and M. J. Heller. Electric field directed nucleic acid hybridization on microchips. *Nucleic Acid Res* 25:4907–14.
82. Heller, M. J. 1996. An active microelectronic device for multiplex DNA analysis. *IEEE Eng Med Biol* 15: 100–04.
83. Gilles, P. N., D. J. Wu, C. B. Foster, P. J. Dillon, and S. J. Chanock. 1999. Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips. *Nat Biotechnol* 17:365–70.
84. Cheng, J., E. L. Sheldon, L. Wu, A. Uribe, L. O. Gerrue, J. Carrino, M. J. Heller, and J. P. O'Connell. 1998. Preparation and hybridization analysis of DNA/RNA from E. coli on microfabricated bioelectronic chips. *Nat Biotechnol* 16:501–46.
85. Westin, L., X. Xu, C. Miller, L. Wang, C. F. Edman, and M. Nerenberg. 2000. Anchored multiplex amplification on a microelectronic chip array. *Nat Biotechnol* 18:190–204.
86. Kawata, S., H.-B. Sun, T. Tanaka, and K. Takada, 2001. Finer features for functional microdevices. *Nature* 412: 697–698.

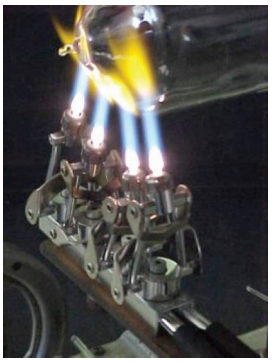
9

Thermal Energy-Based Forming Techniques—Thermoforming

A metal laser sintering/milling hybrid machine, the LUMEX 25C, developed by Matsuura Machinery Corporation in Japan, successfully combines freeform manufacturing and high-speed milling. The integration of laser sintering of metallic powder and high-speed cutting eliminates the finish machining operations (http://www.matsuura.co.jp).

Outline

Introduction
Spray Pyrolysis
Screen Printing/Silk Screening
Doctor's Blade or Tape Casting
Plasma Spray Deposition
Plasmas and Nanoparticles
Laser and Electron Beam Deposition
Questions
References



A four-jet lathe burner for glass blowing. This type of burner is usually mounted on a movable carriage attached to a glassblower's lathe.

Introduction

Thermal forming (or thermoforming) is a method where thermal energy provided by a heat source "transforms" a material's structure and/or shape. Structural material changes occur, for example, during crystallization, sintering, alloying, annealing, decomposition, and pyrolizing. Shaping in thermoforming may be free form, such as in glass blowing and in selective laser sintering (SLS) (see inset in the heading of this chapter). The latter thermoforming method is a rapid prototyping (RP) technique for solid parts, made by solidifying powder-like materials layer by layer through exposure of the surface of a powder bed to a laser beam. However, thermoforming may also involve a step that combines heat and pressure to conform to a final shape in a mold, such as in vacuum forming, a method for thermally shaping a flat plastic sheet into a three-dimensional (3D) shape by heating the sheet and withdrawing the air between the sheet and the mold.

Thermal and chemical forming techniques are often intertwined because heat (and catalysts) usually aids a chemical reaction. The relative contribution of thermal to chemical in a given process may range from 100% thermal to 100% chemical but falls often somewhere in between. Crystallization (see Volume I, Chapter 4), alloying, and sintering, for example, do not involve a chemical reaction and could be viewed as purely thermal forming techniques, but annealing, decomposition, and pyrolizing all typically involve both heat and a chemical reaction. The predominantly thermal techniques detailed in this chapter are spray pyrolysis, silk screening of ceramics, tape casting, plasma spraying, laser and electron beam deposition, and selective laser sintering (SLS). Spray pyrolysis is a simple form of chemical vapor deposition (CVD) and, together with sol-gel, a chemical forming technique that was covered in Chapter 8, a contender for nanoparticle manufacturing. Silk-screening and tape casting technology constitute two important thermal forming processes often referred to as *hybrid* manufacturing methodologies. Tape casting is also known as *doctor blading* or *knife coating*, and the process is well known under these names in many industries, including paper, plastic, and paint manufacturing. Laser and electron beam deposition can be subtractive or additive, and these two options are typically available in the same laser or electron beam reactor. SLS is a rapid prototyping technique (RP) (see also Chapter 8, Table 8.3) method for solid parts, made by solidifying powder-like materials layer by layer through exposure of the surface of a powder bed to a laser beam. The method might also integrate high-speed computer numerical controlled cutting, effectively combining an additive and a subtractive manufacturing technique, in which case RP and real manufacturing are both options.

Spray Pyrolysis

Introduction

In spray pyrolysis, a precursor reagent dissolved in a carrier liquid is sprayed through a nozzle or set of nozzles to form an aerosol, and the droplets in the aerosol experience thermal decomposition, i.e., pyrolysis in a hot zone. In the case of film formation, the hot zone is simply a hot substrate surface, and in the case of particle formation, the hot zone may be a furnace. Spray pyrolysis is also known as *aerosol decomposition synthesis*, in which a droplet-to-solid particle conversion takes place.

The production of droplets and their dispersion into a gas environment is called *atomization*. There are a number of different kinds of atomizers in use, as listed in Table 9.1. The atomization variables listed here are droplet size, which determines the size of the final solid particle, the atomization rate, which affects the scalability of the process, and the droplet velocity, which affects the residence time of the droplets in the hot zone. One other important parameter, not listed, is the droplet size dispersion, which relates to the homogeneity of the product.

Spray pyrolysis is a gas-phase synthesis method and constitutes a form of chemical vapor deposition (CVD) without the expensive vacuum reactors. During CVD, as we saw in Chapter Seven, the constituents of a vapor phase, often diluted with an inert carrier gas, react at a hot surface (typically higher than 300°C) to deposit a solid thin film. The CVD

Table 9.1 Characteristics of Atomizers Typically Used in Spray Pyrolysis

Atomizer	Droplet Size (μm)	Atomization Rate (cm^3/min)	Droplet Velocity (m/s)
Pressure	10–100	3–no limit	5–20
Nebulizer	0.1–2	0.5–5	0.2–0.4
Ultrasonic	1–100	<2	0.2–0.4
Electrostatic	0.1–10		

Source: Messing, G. L., S.-C. Zhang, and G. V. Jayanthi. 1993. Ceramic powder synthesis by spray pyrolysis. *J Am Cer Soc* 76:2707–26.[1]

technologies mostly evolved around the integrated circuit (IC) industry. However, spray pyrolysis, the simplest form of CVD, has never been a contender in that arena. Rather, it is a viable technology for large-area films in solar cells and window panes and, more recently, for the production of nanopowders. Nanoparticle synthesis is a burgeoning field with a broad range of applications. The "holy grail" of research in this field is the ability to generate particles with controlled and narrow size distribution and adequate throughput; thus, the particle size can be tailored as required by the application. In spray pyrolysis when used for nanoparticle synthesis rather than film formation, an organic precursor (liquid or gas) is forced through an orifice at high pressure and burned in a furnace. The resulting ash is air classified to recover the nanoparticles (typically metal oxides).

How Does It Work?

As mentioned above, spray pyrolysis produces films or powders. A simple spray pyrolysis setup for film deposition is shown in Figure 9.1.[2] Metallic salt solutions are sprayed onto a hot substrate where droplets experience thermal decomposition. Spray pyrolysis tends to work best for ceramics [inorganic, nonmetallic materials that are typically crystalline in nature and are compounds formed between metallic and nonmetallic elements, such as aluminum and oxygen (alumina-Al_2O_3)]. All spraying processes display the same significant variables: the substrate temperature, ambient temperature, chemical composition of the carrier gas and/or environment, carrier gas flow rate, nozzle-to-substrate distance, droplet radius, solution concentration, solution flow rate, and—for continuous processes—substrate motion.[2] Because individual droplets

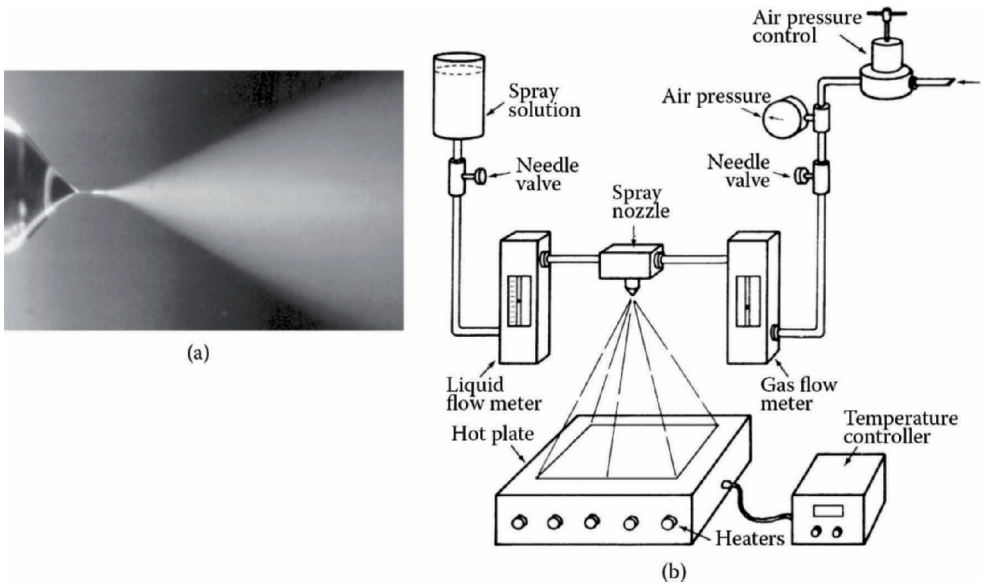


FIGURE 9.1 (a) Atomizer and aerosol. (b) Spray pyrolysis setup. (Mooney, J. B., and S. B. Radding. 1982. Spray pyrolysis processing. *Ann Rev Mater Sci* 12:81–101. Courtesy of Jack Mooney, SRI International, Menlo Park, CA.[2])

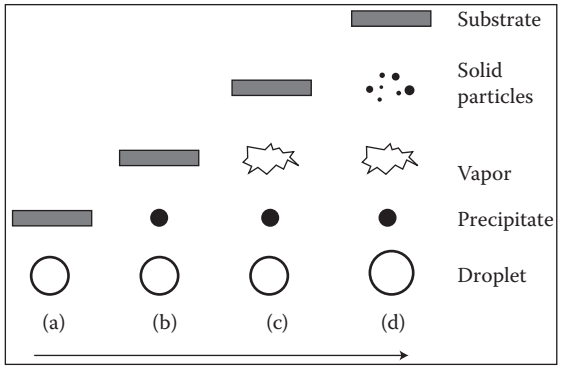


FIGURE 9.2 Steps involved in the decomposition of the dissolved precursors as the temperature of the substrate is ramped up. Steps a to d are explained in the text.

evaporate and react quickly, grain sizes can be very small (e.g., as small as 0.1 µm). The small grains pose a disadvantage for most semiconductor applications but not necessarily for sensor applications; for example, in SnO_2-based gas sensors (Taguchi sensor; see Example 9.1) where sensitivity resides in the intergranular contacts (necks).

In Figure 9.2 we detail the possible scenarios of how droplets laden with solid-particle precursors arrive at a hot substrate as the temperature of the substrate is ramped up. If the temperature is too low, the droplets splash onto substrate and decompose there (Figure 9.2a). At somewhat higher temperature, the solvent evaporates and dry precipitate hits the substrate and decomposes (Figure 9.2b). At higher substrate temperatures, the solvent evaporates and the precipitate melts and vaporizes without decomposition and vapor diffuses to the substrate (Figure 9.2c). At yet higher temperatures, the precursor decomposes to make solid particles before they reach the substrate (Figure 9.2d).

At a fixed substrate temperature, smaller droplets may vaporize before hitting the substrate, and because droplets decrease in size as they approach the substrate, the concentration gradient within the approaching droplet keeps on changing. If the surface concentration of the solute exceeds the solubility limit, precipitation occurs at the surface of the particles, and hollow spheres result, increasing the film roughness.

Spray pyrolysis is a simple, inexpensive, high-throughput technique used to produce several compound semiconductors with utility in various devices, such as solar cells, gas sensors, antireflection coatings, and ion selective electrodes.[3] The technique does not require high-quality targets or substrates and does not need a UHV system. The process produces relatively thick films, is difficult to control, and is not compatible with IC processing. It also tends to form hollow particles (see above) and consumes lots of solvent.

In case spray pyrolysis is used to make a powder instead of a film, a setup as shown in Figure 9.3 may be used. In this apparatus, the precursor solution is atomized into a diffusion dryer, then a thermolysis,

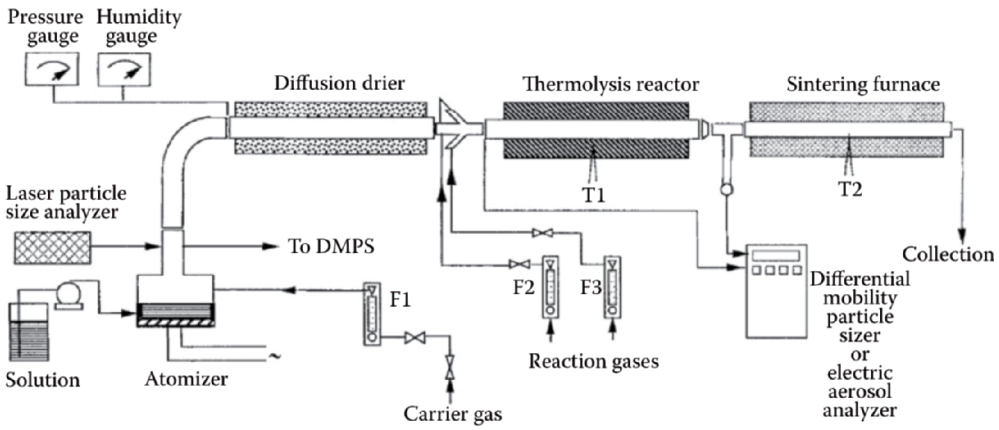


FIGURE 9.3 The diffusion dryer is used to dry and remove water vapor from the aerosol generated by the atomizer. Thermolysis of the precipitate particle at higher temperature to form a microporous particle occurs in the thermolysis furnace and sintering of the microporous particle to form a dense particle in the sintering furnace. The setup is also equipped with a differential mobility particle sizer (DMPS). (From Messing, G. L., S.-C. Zhang, and G. V. Jayanthi. 1993. Ceramic powder synthesis by spray pyrolysis. *J Am Cer Soc* 76:2707–26.[1])

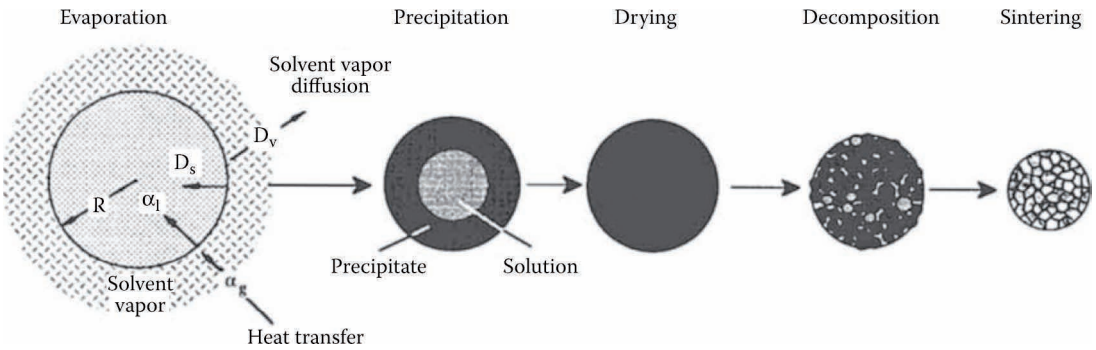


FIGURE 9.4 Droplet evolution in spray pyrolysis. Thermal decomposition or pyrolysis—forms a nanoporous structure and finally sintering—involves the adhesion/solidification of the crystallites. (From Messing, G. L., S.-C. Zhang, and G. V. Jayanthi. 1993. Ceramic powder synthesis by spray pyrolysis. *J Am Cer Soc* 76:2707–26.[1])

and finally a sintering furnace. In this series of reactors, the aerosol droplets undergo evaporation and solute condensation within the droplet, by drying, thermolysis of the precipitate particle at higher temperature to form a microporous particle, and sintering of the microporous particle to form a dense particle. The setup is also equipped with a differential mobility particle sizer (DMPS). Depending on the droplet size and solution concentration, particles sized from some submicrometers to some hundreds of micrometers are available by spray pyrolysis.

The stages the aerosol droplets go through are sketched in Figure 9.4.

During evaporation of solvent from the droplet surface, there is diffusion of solvent vapor away from the droplet (D_v), change in droplet temperature, diffusion of solute toward the center of the droplet (D_S), and change in droplet size (R). Precipitation/drying involves volume precipitation or surface precipitation of the solute, followed by the evaporation of the solvent through the nanoporous crust. If the solute concentration at the center of the drop is less than the equilibrium saturation of the solute at the droplet temperature, then precipitation occurs only in that part of the drop where the concentration is higher than the equilibrium saturation, i.e., surface precipitation (see also Figure 9.5).

It is one of the unique characteristics of spray pyrolysis that chemical reactions occur within micrometer to submicrometer-sized liquid droplets, microcapsule chemical reactors so to speak. Besides controlling the precipitation in these microreactors

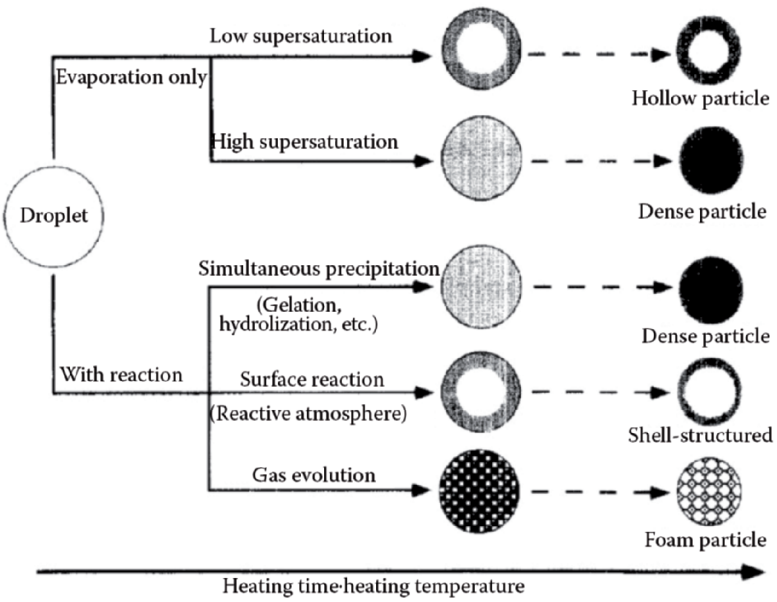


FIGURE 9.5 Droplet pyrolysis with and without chemical reaction.

by simple evaporation, precipitation can also be induced by a chemical reaction, as illustrated in Figure 9.5. Reaction schemes relied on to induce precipitation might involve gelation, hydrolyzation, gas evolution, reacting with a reactive atmosphere, and so on. We encountered another example of microcapsule reactions in Volume I, Chapter 6 where we describe droplet manipulation fluidic platform. In this case, the droplet microreactors are injected in a nonmixing fluid.

Salt-Assisted Spray Pyrolysis

Spray pyrolysis, like sol-gel (see Chapter 8), is a contender for making nanoparticles, used for chemical-mechanical polishing (CMP), magnetic recording tapes, sunscreens, and automotive catalysts. Spray pyrolysis is an attractive method for preparing uniform size and composition of nanoparticles because of its high production rate, continuous unit operation, and simple apparatus. However, with conventional spray pyrolysis the production of submicrometer particles faces some challenges. The concept derives from the principle of one particle from one droplet; thus, very fine droplets are needed to produce nanoparticles. The atomizers most commonly used to generate droplets tend to produce average droplet diameters in the range of several micrometers. For a droplet to dry from a typical initial diameter of 5 μm into a solid particle with a diameter of 0.1 μm, the initial volume fraction of dissolved involatile solute must be less than 0.0008%. In practice, these low solution concentrations lead to a low particle generation rate and make it difficult to maintain a high purity of the final particle product. Conventional spray pyrolysis produces particles that have multiple nanosized crystallites, but these crystallite aggregates are virtually inseparable because of the formation of a 3D network within each drop. The salt-assisted spray pyrolysis (SASP) developed by Dr. K. Okuyama from the Hiroshima University overcomes these limitations associated with conventional spray pyrolysis.[4] In the SASP method, salts such as $NaNO_3$, KNO_3, $LiNO_3$, and KCl are added to the spray precursor solution where the salt promotes nucleation and prevents nanoparticles from agglomerating. As a result, uniform-sized nanoparticles are produced. The diameters of nanoparticles produced by the SASP method range from several nanometers to several tens of nanometers. These sizes are about one-thirtieth to one-eightieth those of fine particles produced by traditional spray pyrolysis. Size distribution of SASP particles appears to be independent of their parent solution droplets; moreover, the SASP process is less energy consuming because it runs at lower temperatures.

In Figure 9.6 SASP is compared with traditional spray pyrolysis (SP).

Flame Spray Pyrolysis

Flame spray synthesis has been established as a state-of-the-art method for the production of high-purity mixed metal oxides for a vast range of applications in catalysis, optics, and electronics (Figure 9.7). It is primarily useful for making oxides. In flame spray pyrolysis, one directly sprays a liquid precursor into a flame, and particle synthesis occurs within the flame by a combustion reaction. This allows for the use of low-vapor-pressure precursors. Products are affected by the flame configuration, process temperature, oxidant composition, and additives. With metalorganic precursors, flame spray pyrolysis constitutes a commercial approach for the manufacture of silica and titania powders. Using heavy residual oils as precursor, carbon black is produced (millions of metric tons per year). The production of TiO_2 powder from the flame spray pyrolysis of $TiCl_4$ (the chloride process) accounts for ~60% of TiO_2 production (Dupont is the world's largest producer with about ~22% of global production). The TiO_2

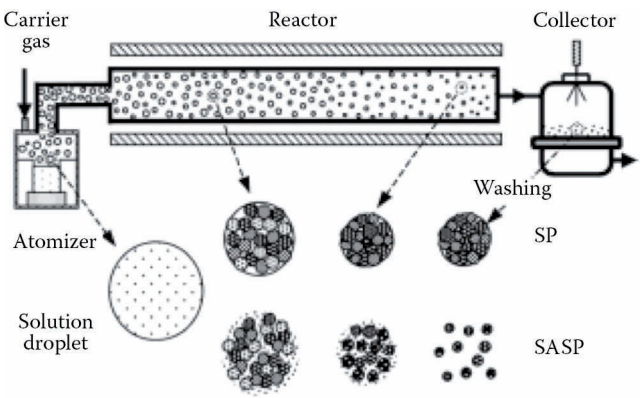


FIGURE 9.6 Comparison of SASP with traditional SP.

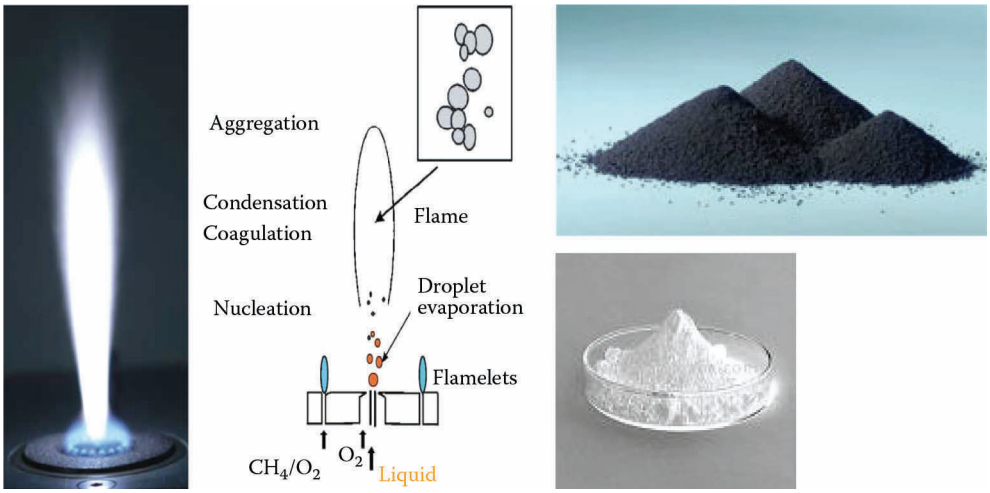


FIGURE 9.7 Flame spray pyrolysis. Using heavy residual oils as precursor, carbon black is produced using flame spray pyrolysis (millions of metric tons per year). The white powder is TiO_2.

powder made this way is the least expensive submicrometer powder available (see bottom right in Figure 9.7). Because of its high index of refraction, it is an ideal opaque dye for paper, paint, pigment, food, and cosmetics. In the chloride process, an electric field can be used to control particle size in the flame.[5] The flame is held between two plates, with one plate connected to a DC source and the other grounded as sketched in Figure 9.8. A flame filled with newly formed TiO_2 particles creates both positive and negative ions. As a result, these charged particles that otherwise would have collided with each other and grown to form larger particles are prevented to do so by the presence of the electrical field. The strength of the E field determines the velocity of the ions and essentially the size of the product particles. This method allows very precise control of the size (within 1 nm) and aggregate structure and crystallinity.

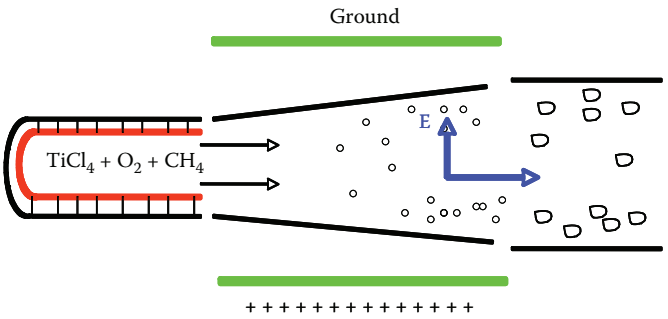


FIGURE 9.8 Electrical field control of particle size in flame spray pyrolysis of TiO_2 nanoparticles.

Pros of this method are its large capacity, proven technique, and high purity of products. Cons include the fact that, unless electrical fields are applied, agglomerates form, and it features a wide particle size distribution.

Electrospray Pyrolysis

Electrospraying is capable of atomizing a liquid into ultrafine droplets and offers yet another option for converting conventional spray pyrolysis into a nanoparticle production means. In this method, the meniscus of a conducting solution supported at the end of a capillary tube becomes conical when charged to a high-enough voltage (several kV) with respect to a counter electrode. When the electrostatic force becomes stronger than the surface tension of the liquid, the solution exits the capillary as a highly charged jet of fluid. As the solvent evaporates from the droplets, particles form and are deposited onto a heated collector. The droplets are stably formed by the continuous breakup of a jet extending from this liquid cone (Figure 9.9).

Electrospray (ES) pyrolysis can produce monodisperse droplets/particles over a wide size range, from a few nanometers to hundreds of micrometers, depending on liquid flow rate, applied voltage, and liquid electric conductivity. Especially in the small size range, the capability of producing monodisperse particles with relative ease is unmatched by any other aerosol generation scheme. Zinc sulfide

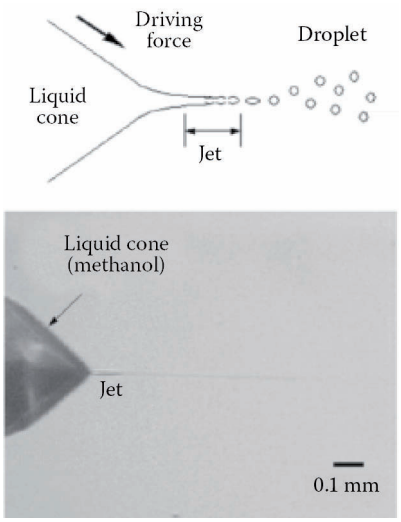


FIGURE 9.9 Droplet formation in the electrospray technique.

FIGURE 9.10 Electrospray from an array of nozzles etched in silicon, with a density of 250 sources/cm^2. (From Deng, W., J. F. Klemic, X. Li, M. A. Reed, and A. Gomez. 2006. Increase of electrospray throughput using multiplexed microfabricated sources for the scalable generation of monodisperse droplets. *Aerosol Sci* 37:696–714.[7])

particles 20–40 nm in diameter were prepared by electrically driven spray pyrolysis from solutions of ethyl alcohol with zinc nitrate [$Zn(NO_3)_2$] and thiourea [$SC(NH_2)_2$] at concentrations from 0.0025–0.2 M and electrical conductivities between 10^{-4} and 10^{-1} Sm^{-1}.[6]

A crucial drawback that has hampered wider ES pyrolysis applications is the low flow rate at which the cone-jet mode is established. Multiplexing the spray source, that is, operating several ESs in parallel is a possible way to address this problem. Deng et al.[7] demonstrated the microfabrication and testing of a compact multiplexed system of ESs as an array of nozzles etched in silicon, with a density of 250 sources/cm^2 (see Figure 9.10).

It is projected that this arrayed ES technology, once optimized, will be applied in various fields, including the synthesis of biological nanoparticles and biomedical coatings for controlled/targeted drug delivery, the synthesis of nanoparticle ceramics, superconductors, quantum dots, and thin films.

The ES method has also had a dramatic technological impact in the field of analytical chemistry. Electrospray ionization (ESI) is a technique used in mass spectrometry to produce ions. It is especially useful in producing ions from macromolecules because it overcomes the propensity of these molecules to fragment when ionized. The development of ESI for the analysis of biological macromolecules was rewarded with the attribution of the Nobel Prize in Chemistry to John Bennett Fenn in 2002. In the same electro-spraying set-up of Figure 9.9 and 9.10, higher molecular weight solutions enable one to produce continuous fibers instead of beads. At low viscosity, the columbic force dominates the elastic force, leading to a rapid breakup of the polymer jet. However, increased viscoelasticity engendered by chain entanglements does not allow the breakup of the jet, resulting in continuous fibers in the so-called electrospinning process. The electrospinning process is indeed a simple, versatile, and widely used method of producing all types of polymer nanofibers.[8–14] As in the case of particle spraying, pyrolysis may further convert the polymer nanofiber precursor to a final desirable product (e.g., carbon nano fibers through the carbonization process, i.e., the high temperature heating in the absence of oxygen; see also C-MEMS in Volume III, Chapter 5).

Among the precursors used in the synthesis of carbon nanofibers by electrospinning, polyacrylonite (PAN)/N,N-dimethylformamide (DMF) solution is the most commonly employed. Carbon nanofibers with diameters in the range of 80–750 nm are typically obtained from this solution by electrospinning. Our team has fabricated carbon nanofibers by electrospinning SU-8 (an epoxy-based negative photoresist) as the polymer precursor and subsequent pyrolysis of that precursor.[15] We have shown that the electrospun material retains its photoresist properties and that it can be directly

patterned in a wide variety of patterns by lithography before pyrolysis. The pyrolysis of SU-8 nanofibers to produce carbon nanofibers will enhance the C-MEMS toolbox considerably and allow for the facile integration of high-surface-area fibers with other photoresist derived carbon structures. This not only will allow for the making of interesting photopatternable carbon composites but also will enable the fabrication of multiscale hierarchal assemblies such as fractal-like electrodes where a high external surface area is coupled with a minimized internal resistance, which is a feature of great utility, in batteries and fuel cells.[16] The biocompatibility of SU-8-derived carbon[17] provides further motivation for the development of multiscale hierarchal scaffolds for tissue engineering and in vivo biosensors.

Example 9.1: Spray Pyrolysis
Application: Taguchi Gas Sensor

The use of a ceramic cylindrical tube in a classical Taguchi sensor (Figure 9.11) maximizes the utilization of heater power; thus, most of the power is used to heat the gas-sensitive tin oxide (SnO_2). A tin oxide paste is applied by dip coating the outside ceramic body to cover the thick film resistance measuring electrodes and is then sintered at high temperature. The resistance of tin oxide at high temperature gives a measure for the amount of reducing gases in the contacting atmosphere. Sintering stabilizes the intergranular contacts (necks) where the sensitivity of the gas sensor resides.[18] The design shown in Figure 9.11 is not suited for mass production because it involves excessive hand labor. The problems with the thick film structure mainly reside in the areas of reproducibility:

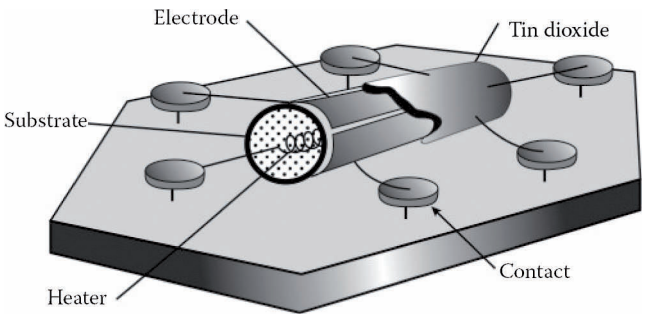


FIGURE 9.11 Classical Taguchi gas sensor.

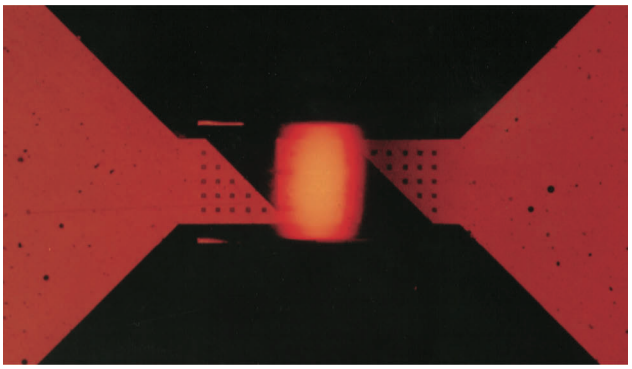


FIGURE 9.12 Heater element.

compressing and sintering a powder, the deposition of the catalyst, and the use of binders and other ceramics (e.g., for filtering). Application of IC techniques could improve the state of the art dramatically if a thin semiconductor oxide film (e.g., tin oxide) with the same sensitivity as the traditional thick sintered film could be made. Micromachined heater elements, as shown in Figure 9.12, improve the reproducibility and the absolute power budget needed to bring the sensor to the required temperature. Thermal efficiency and low power budget are obtained by thermally isolating the heater element on a thin, suspended membrane. Whereas low power budget sensors have proven feasible, it has been more difficult to make a compatible thin film as sensitive as the ceramic-type, thick film devices. When making thin films with physical vapor deposition (PVD) methods, the films have a lower surface area and buried intergranular contacts (see Figure 9.13). The buried intergranular contacts cannot be reached by the contacting gas as easily as in the powder case, making not only for lower sensitivity but also slower responding gas sensors. The slower response of the thin film sensors is because oxygen diffusion in the grain boundaries between compacted grains involves long time constants. When the grain boundaries are blocked on purpose, the sensor reacts faster but displays even less sensitivity.[18]

Using silk screening of thick films (see following section) on a thin membrane would seem to provide the solution—a low-energy heater combined with a fast and sensitive thick tin oxide film. Unfortunately, during our attempts to deposit thick films this way, usually the thin Si membrane broke as a result of the pushing action of the squeegee (see Figure 9.14).[19] Spray pyrolysis

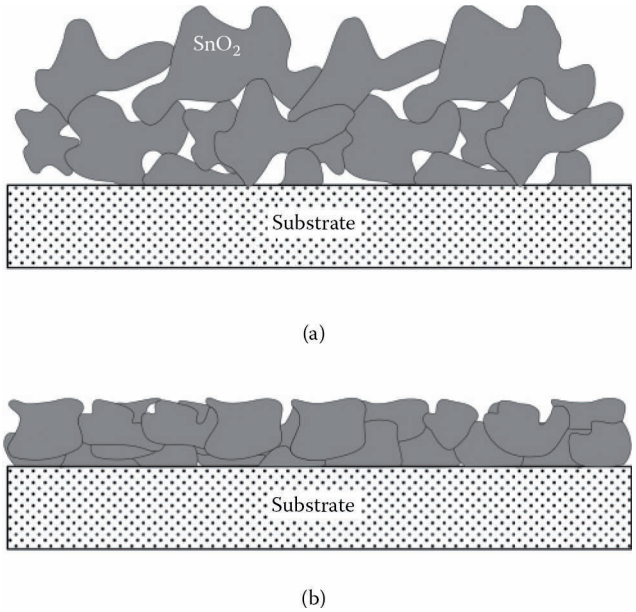


FIGURE 9.13 Grain structure in a sintered powder, thick SnO_2 film (a) and a PVD, SnO_2 thin film (b).

provides the possible answer here because its deposition does not involve mechanical pressure on the silicon membrane. Fast-responding, planar Taguchi SnO_2 gas sensors were made by spraying, with oxygen as the carrier gas, an organic solution of $(CH_3COO)_2SnCl_2$ in ethyl acetate onto quartz, glass, and Al_2O_3 ceramic substrates, heated to 300°C.[3] This result carries significance; other micromachining attempts based on thin PVD SnO_2 film fail to reach the sensitivity attained with the classical thick films and sintered SnO_2 films. Spray pyrolysis, with its very small grain-size deposits, holds the promise of reaching the objective of providing films with freely accessible polycrystalline grains. As a planar technology, it is also suited for large substrates and considered a viable alternative to the dip-coating process.

Screen Printing/Silk Screening

Introduction

Screen printing presents a more cost-effective means of depositing a wide variety of films on planar substrates than does IC technology, especially when fabricating devices at relatively low production volumes. The technique constitutes one of several thick film or hybrid technologies used for selective coating of flat surfaces (e.g., a ceramic substrate). The technology was originally developed for the production of miniature, robust, and (above all) cheap electronic circuits. The up-front investment in a thick film facility is low compared with that of IC manufacturing. For disposable chemical sensors, recent industrial experience indicates that screen printing thick films is a viable alternative to Si thin film technologies. Materials based on low-temperature organic polymers, to be cured at low temperatures or using UV radiation, as well as high-temperature enamel coatings with ceramic paints are coated in screen printing; the latter need sintering at high temperatures. Typical film thicknesses are in the range from several 10 to several 100 µm. Therefore, the coefficients of thermal expansion of the coat have to be matched well to that of the substrate.

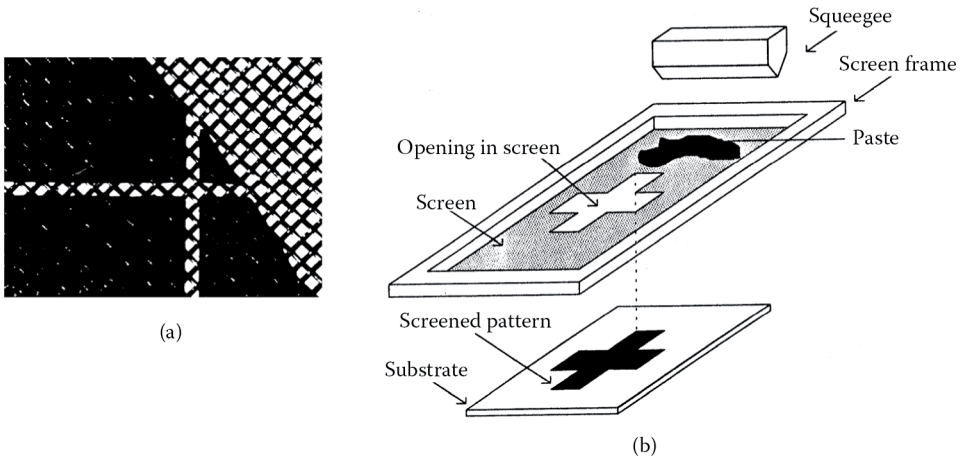


FIGURE 9.14 Screen-printing. (a) Screen with a 0.002-in. line opening oriented at 45° to the mesh weave. (b) Schematic representation of the screen-printing process. (From Lambrechts, M., and W. Sansen. 1992. *Biosensors: microelectrical devices.* Philadelphia: Institute of Physics Publishing. With permission.[20])

How Does It Work?

The elements of screen printing are the screen with openings defining the pattern to be printed, the screen frame, a squeegee, the ink material, and the substrate. A paste or ink is pressed onto the substrate through the openings in the photoemulsion fashioned on the stainless steel screen (see Figure 9.14a).[20] The mesh openings are created using photolithography on the photoemulsion in the screen frame. The paste consists of a mixture of the material of interest, an organic binder, and a solvent. The organic vehicle determines the flow properties (rheology) of the paste. The bonding agent provides adhesion of particles to one another and to the substrate. The choice of active particles makes the ink a conductor, a resistor, or an insulator. The lithographic pattern in the screen emulsion is transferred onto a substrate by forcing the paste through the mask openings with a squeegee (Figure 9.14b). This process depends on the mesh opening size and the viscosity of the ink. The squeegee further keeps the mesh in contact with the substrate, a process that depends on the screen tension and the off-contact distance. Squeegees are made from a stiff material like polyurethane to avoid deformation during printing. Relatively hard squeegees are required to withstand the high pressure needed to print at high speed, for example, high-viscosity materials on glass plates. However, the harder the material, the lower the ability to print an even layer of ink on an uneven surface. The squeegee edge must be sharp. The angle of the squeegee to the mesh defines the amount of material to be printed. Pulling the squeegee at 15–20° off vertical ensures an even deposit of material. The setting of the exact angle depends in part on the squeegee's profile and its hardness. The possible speed of the squeegee depends on the hardness of the squeegee and the viscosity of the printed material.

In a first step, paste is put down on the screen (Figure 9.15Aa); then the squeegee lowers and pushes the screen onto the substrate, forcing the paste through openings in the screen during its horizontal motion (Figure 9.15Ab).[20] During the last step, the screen snaps back; the thick film paste that adheres between the screen and the substrate shears; and the printed pattern is formed on the substrate (Figure 9.15Ac). Heavier frames cause the screen to be less deflected and minimize distortions and achieve a better resolution of the printed pattern. Throughput and resolution depend on the dimensions of the mesh openings (holes in the screen) and the nature of the pastes/inks. This leads to compromises in selecting the mesh size: a coarse mesh enables fast processing and high throughput but at low resolution, whereas a fine mesh enables better resolution but causes problems when pressing highly viscous materials through the screen. With a 325-mesh screen (i.e., 325 wires/in. or 40-μm

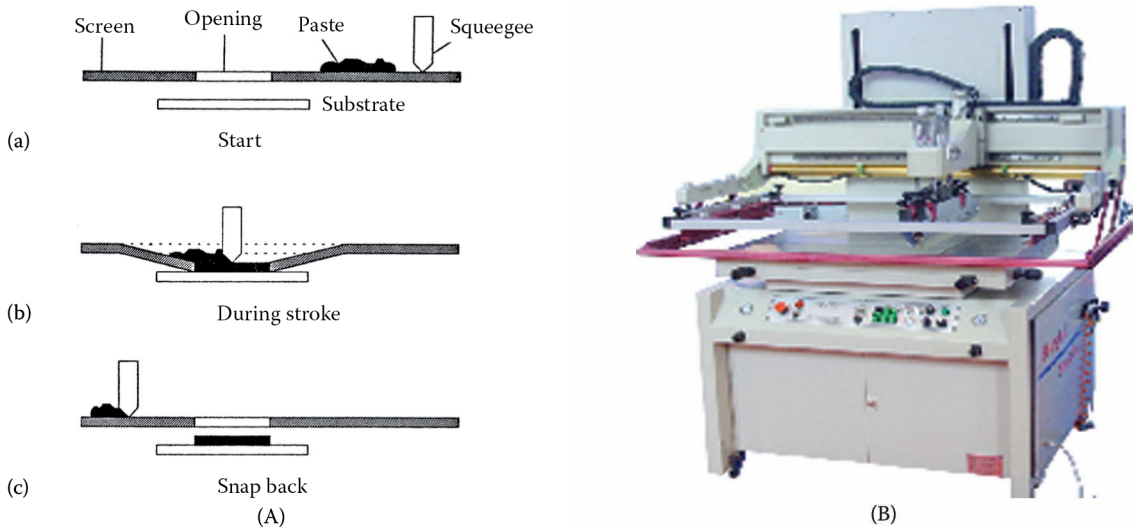


FIGURE 9.15 (A) The three different steps of the silk screening process. (From Lambrechts, M., and W. Sansen. 1992. *Biosensors: microelectrical devices.* Philadelphia: Institute of Physics Publishing. With permission.[20]) (B) Shanghai Shuoxing Screen Printing Equipment Co., Ltd.

holes) and a typical paste, a lateral resolution of 100 µm can be obtained. For difficult-to-print pastes, a stencil mask may be used, such as a thin metal foil with cut openings. However, the resolution of this method is inferior (>500 µm). Lower-cost and high-throughput printing techniques, such as screen printing, have a great promise in the fabrication of organic light-emitting devices (OLEDs). Jabbour et al. demonstrated, for the first time, the use of screen printing to deposit ultrathin layers in OLEDs of less than 15 nm with root mean square surface roughness of less than 1.5 nm.[21] After printing, the wet films are allowed to settle for 15 min to flatten the surface while drying. This removes the solvents from the paste. Subsequent firing burns off the organic binder; metallic particles are reduced or oxidized; and glass particles are sintered. Typical temperatures range from 500–1000°C. After firing, the thickness of the film ranges from 10–50 µm.

A typical silk screening setup is the one from Universal Instruments Corporation (http://www4.uic.com), the DEK 4265 Horizon Screen (see also The Shanghai Shuoxing screen printing machine in Figure 9.15B). Dupont provides technical tips for screen printing at http://www2.dupont.com/MCM/en_US/techinfo/techtips.html, and the Screen Printing Magazine can be found at http://www.screenweb.com/home.php.

Types of Inks/Pastes and Applications

Inks are formulated to exhibit pseudoplastic* behavior, as illustrated in Figure 9.16[22,23] (see also Volume I, Chapter 6 on fluidics), to prevent flowing through the screen until the squeegee applies sufficient pressure. Almost all materials compatible with the high firing temperature and the other ink constituents can be used to screen print. Different pastes—conductive (e.g., Au, Pt, Ag/Pd), resistive (e.g., RuO_2, IrO_2), overglaze, and dielectric (e.g., Al_2O_3, ZrO_2)—are commercially available.

The conductive pastes are based on metal particles, such as Ag, Pd, Au, Pt, or a mixture of these combined with glass. Glass is necessary for the adhesion of the metal conductor to the ceramic (Al_2O_3) substrate. The following conductive pastes can be distinguished by the bonding mechanism used.

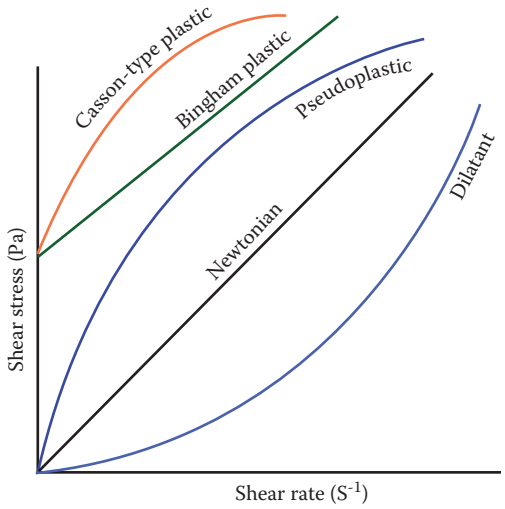


FIGURE 9.16 Pseudoplastic behavior of inks for silk screening (see also Volume I, Chapter 6).

- Glass-bonded or fritted pastes are inks where adhesion of the metal is achieved with the addition of a glass mixture (30%) (a typical glass composition is 65% PbO, 25% SiO_2, and 10% Bi_2O_3). A fritted ink contains powdered glass that binds the ink to the substrate material when fired at a temperature of 850°C. Thus, a fritted ink will generally consist of the glass frit, a material defining the desired ink property, and an organic vehicle that renders the ink printable. Vehicles typically consist of solvents mixed with slightly more viscous materials, such as resins, in a ratio designed to give the optimum overall viscosity to the ink.
- Oxides or fritless-bonded inks adhere to the metal via the addition of copper oxide (3%).
- Mixed-bonded pastes are inks for which adhesion is achieved by use of both glass and copper oxide.

Resistive pastes are based on RuO_2 or $Bi_2Ru_2O_7$ mixed with glass (65% PbO, 25% SiO_2, 10% Bi_2O_3). The resistivity is determined by the mixing ratio. Overglaze and dielectric pastes are based on glass mixtures. According to composition, different melting temperatures can be achieved.

Thick film technology with the above type of traditional inks has application in the construction of

* In pseudo-plastic fluids viscosity diminishes as the applied velocity gradient increases (see also Volume I, Chapter 6).

a wide variety of hybrid sensors, such as sensors for radiant signals, pressure sensors, strain gauges, displacement sensors, humidity sensors, thermocouples, capacitive thick-film temperature sensors, and pH sensors (see Middlehoek et al.[23] and references therein). Also, with silicon-based sensors (e.g., pressure sensors and accelerometers) die-mounted on a ceramic substrate, thick film resistors are used for calibration by trimming resistors on the ceramic substrate.

New inks, specifically developed for chemical and biological sensor applications, are available or under development. For example, SnO_2 pastes incorporating Pt, Pd, and Sb dopants have been developed for the construction of high-temperature (>300°C) semiconductor gas sensors for reducing gases (so-called *Taguchi sensors*) (see Example 9.1).[18] Thick metal phthalocyanine films have been deposited on alumina to form the active material in relatively low-temperature (<180°C) gas sensors.[24]

In Chapter 8 we saw that for biosensor applications, thick film technology based on pastes (polymer-based inks) that can be deposited at room temperature have been developed.

Doctor's Blade or Tape Casting

Silk screening and tape casting technology constitute two important thermal forming processes, often referred to as *hybrid* manufacturing methodologies. Here we detail tape casting as yet another example of thermal forming. Tape casting is also known as *doctor blading* or *knife coating*, and the process is well known under these names in many industries, including paper, plastic, and paint manufacturing. The "doctor's blade" is a scraping blade for the removal of excess substance from a moving surface being coated. Howatt was the first to obtain a patent to use the technique for "forming ceramic materials into flat plates, especially useful in the electric and radio fields."[25] This is still the principal application of tape casting today, although its use extends far beyond the field envisioned in 1952 (http://www.drblade.com). Since then, tape casting has been used to manufacture a variety of multilayer ceramic package products, such as low-temperature cofired ceramics, solid oxide fuel cells, and piezoelectric smart devices.

The ceramic application of doctor blading, also referred to as *green tape* technology, is a thermal forming process well suited for fabrication of mesoscale devices (layers with a thickness between several micrometers and millimeters). In green tape technology, a ceramic slurry, either aqueous or organic solvent-based, is continuously dispensed onto a moving substrate from a hopper or other appropriate plumbing. An important requirement of the slurry is that its viscosity is in the 100–130-Poise (10^{-13} Ns/m^2) range. The substrate and dispensed product then move under the doctor's blade, an adjustable gate with a precisely controlled height on a smooth casting/coating surface. The purpose of the doctor's blade is to allow a known thickness of material to pass under the blade. In today's technology, very thin ceramic layers are defined in micrometers, with tapes as thin as 5 μm reported by equipment manufacturers (http://www.drblade.com). Standard tape casting machines come with thickness ranges from 12 μm to more than 3 mm. Precise, smooth control of the carrier or web speed is also essential to ensure uniform films. Next, the substrate and bladed ceramic slurry are moved through a precision oven for partial drying. The result is a flexible green ceramic tape that can be easily machined before final drying and sintering. Ceramic in green tape form may be cut, for example, with lasers or razor blades into any desired shape. Tape stacks are produced by lamination of several single green tapes under heat and pressure. The optimum dried thickness range for tape casting is generally accepted as being from 0.025–1.27 mm. The green tape process flow is summarized in Figure 9.17. In Figure 9.18 we show a photo of a typical tape caster.

Tape casting is the best method for forming large-area, thin, flat ceramic parts and even for certain thin metal foils. Such thin parts are virtually impossible to press and are very difficult to extrude. The difficulties are compounded in dry pressing when the plate is to be pierced with numerous holes because of the increased possibility of nonuniform die fill. On the other hand, punching via holes and slots of various sizes and shapes into green tape is relatively easy and essential to the many types of multilayered ceramic packages designed and manufactured today. Tape casting has been

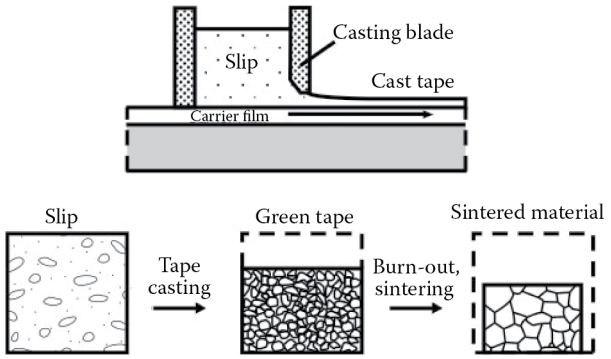


FIGURE 9.17 Tape casting. The top schematic drawing illustrates tape casting. The bottom sketch shows the different stages during the processing: the slip consisting of water, ceramic particles and binder; the cast, dried green sheet; and finally, the microstructure of the sintered material.

expanded to nonceramic, polymer, and metal applications, such as polyvinylidene fluoride (PVDF) battery separator films and tape automated bonding (TAB) semiconductor adhesive-metals. Tape casting is often the preferred choice in manufacturing these products because of its relative simplicity and low implementation cost compared with other available methods, such as dry pressing, slip casting, or injection molding. Tape casting is a straightforward method of forming uniformly thin sheets of film that is inexpensive, scaleable, and may be used with any ceramics, metals, or polymers that are readily mixed in a liquid suspension or slip.

An industrial green tape setup costs between $100,000 and $120,000. Dupont Microcircuit Materials is a typical supplier of green tape materials for a wide variety of electronic and electrical applications in the wireless and wired telecommunications, automotive, military, medical, instrumentation, industrial, data processing, components, and consumer industry segments (Green Tape™; http://www.mcm.dupont.com).

FIGURE 9.18 A typical tape caster. HED Lab-Cast 7-ft. tape caster.

An excellent recent review of tape casting can be found in *Tape Casting: Theory and Practice* by Richard E. Mistler and Eric R. Twiname.[26] More background information on tape casting is also available at http://www.drblade.com.

Example 9.2: Green Tape Technology: Multilayer Planar Oxygen Sensor

Green tape methodology is of special interest in the fabrication of high-temperature gas sensors. As an example, we show Figure 9.19, a multilayer planar oxygen sensor manufactured by NGK Spark Plug Co., Ltd. Also known as a *wide range air-to-fuel ratio* sensor, it incorporates two oxygen-pumping cells and one potentiometric gauge. The operating principle of the NGK sensor is as follows. The potentiometric cell—a lambda sensor—measures EMF between the exhaust in the gap and the atmospheric air. The measured voltage V_{EMF} is compared with V_o (corresponding to EMF of the stoichiometric point). The amplifier applies an appropriate pumping voltage to the pumping cell after comparing V_{EMF} and V_o (if $V_{EMF} > V_o$ then the exhaust is rich; if $V_{EMF} < V_o$ then the exhaust is lean). It is fabricated by laminating and cofiring, at high temperatures, several layers of ceramic green tape, some of which are metal coated and have openings through them.[27] Because both metal and ceramic layers constitute the sensor, the materials should be carefully prepared, and the cofiring process must be tightly controlled to avoid metal diffusion or reaction with the ceramic. Because of the process complexities, this sensor turned out too expensive for common use.

Plasma Spray Deposition

Introduction

Because of the need to incorporate more and different materials in a thickness ranging anywhere from monolayers to a few hundred micrometers,

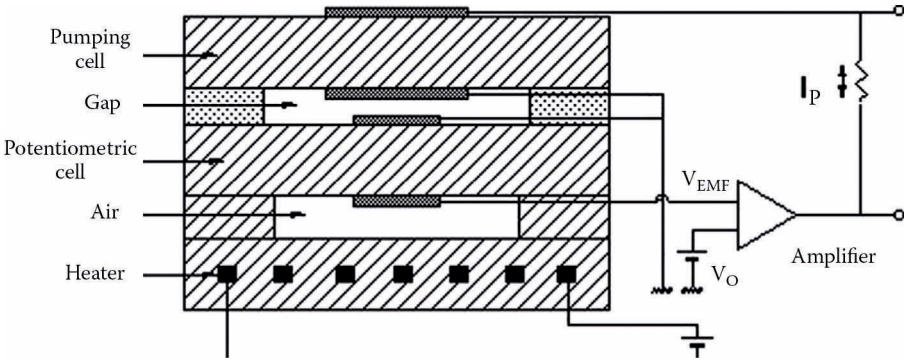


FIGURE 9.19 Planar Oxygen Sensor from NGK, Spark Plug Co., Ltd.

micromachinists are broadening their horizon beyond IC deposition techniques and increasingly are incorporating hybrid thick film methodology in their tool box. Microstructures can be crafted economically with non-IC equipment and materials, and thick film silk screening, spray pyrolysis, sol-gel, and tape casting are all attractive additive techniques in this regard. Spray pyrolysis (see earlier), drop delivery systems (see Chapter 8), and pick-and-place technology (see Volume III, Chapter 4) are but a few examples. Here we review plasma spraying as another example of a non-IC thermal forming tool. With plasma spraying almost any material can be coated on many types of substrates. Applications include corrosion- and temperature-protective coatings, superconductive materials, and abrasion resistance coatings.[28] Today, turbine blades and other components of aircraft engines are plasma coated with corrosion- and temperature-resistant coatings. In this section, we propose that the technique be expanded to the mass manufacture of ceramic-based sensors.

How Does It Work?

The CVD and PVD techniques discussed in Chapter 7, except for the cluster deposition method, rely on atomistic deposition, that is, atoms or molecules are individually deposited onto a surface to form a coating. Plasma spraying, like spray pyrolysis, is a typical particle deposition method; particles a few micrometers to 100 µm in diameter are transported from source to substrate. The basic configuration of a plasma-arc torch setup for layer deposition and a plasma spray nozzle are shown in Figures 9.20 and 9.21, respectively.

In plasma spraying, a high-intensity plasma arc is operated between a stick-type cathode and a nozzle-shaped, water-cooled anode as illustrated in Figure 9.21. For atmospheric spraying, one typically works at power levels from 10–100 kW. Plasma gas, pneumatically fed along the cathode, is heated by the arc to plasma temperatures and leaves the anode nozzle as a plasma jet or plasma flame. Argon and mixtures of argon with other noble (He) or molecular gases (e.g., H_2, N_2, O_2) are frequently used for plasma spraying. Fine powder suspended in a carrier gas is injected into the plasma jet where the particles are accelerated and heated. The plasma jet may reach temperatures of 20,000 K and velocities up to 1,000 ms^{-1}. The temperature of the particle surface is lower than the plasma temperature, and the dwelling time in the plasma gas is very short. The lower surface temperature and short duration prevent the spray particles from being vaporized in the gas

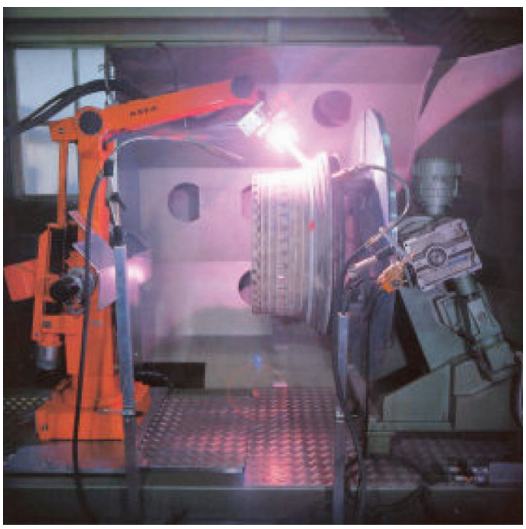


FIGURE 9.20 Setup for plasma spray. With a robot arm and a sound-proof spraying booth.

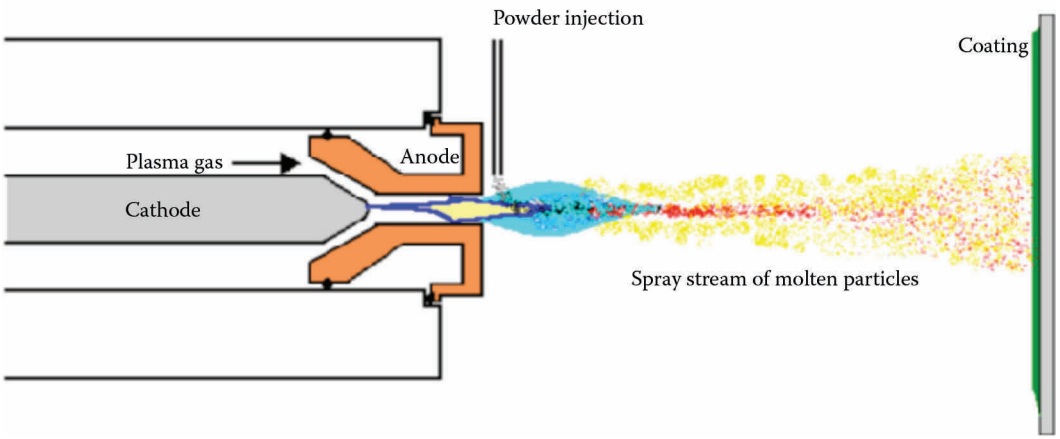


FIGURE 9.21 Plasma spray nozzle.

plasma. The particles in the plasma assume a negative charge, owing to the different thermal velocities of electrons and ions. As the molten particles splatter with high velocities onto a substrate, they spread, freeze, and form a more or less dense coating, typically forming a good bond with the substrate. The resulting coating is a layered structure (lamellae), as shown in Figure 9.22. Typical coatings for high-temperature applications involve an oxidation-resistant coating and a thermal barrier coating (TBC). The role of TBCs is, as their name suggests, to provide thermal insulation of the blade. A coating of about 1–200 μm can reduce the temperature by up to 200°C. The oxidation-resistant coating is also called *bond coat* because it provides a layer on which the ceramic TBC can adhere.

As shown in Figure 9.23, a particle leaving the spray nozzle goes through different regions of temperature and flow velocity. Ideally, the particles should arrive at the substrate at high velocities in a completely molten state to form the densest coating with little porosity. To produce porous films for the fabrication of oxygen sensors in our own work,[19] we positioned the substrate somewhere between regions 4 and 5. To produce dense films, we positioned the substrate in region 4. With this technique, a minimum thickness is about 25 μm, and very thick coats up to a few millimeters thick are possible. When attempting to deposit gas-sensitive layers, such as ZrO_2, on thermally isolated, thin Si membranes to make a power-efficient gas sensor, we found that the kinetic energy of the plasma was too high, and it broke the thin suspended Si membranes. The high temperature and kinetic energy preclude the potential for integrating Si with high-temperature plasma spraying. In Example 9.3 below we demonstrate how plasma spraying of yttria-stabilized zirconia may be used in the batch fabrication of all solid-state oxygen sensors.

Plasma spray deposition is perhaps the most flexible of all the thermal spray processes because it can develop sufficient energy to melt any material. Because it uses powder as the coating feedstock, the number of coating materials that can be used in the plasma spray process is almost unlimited. Plasma spraying equipment may be purchased, for example, from Sulzer Metco (http://www.sulzermetco.com/index.html).

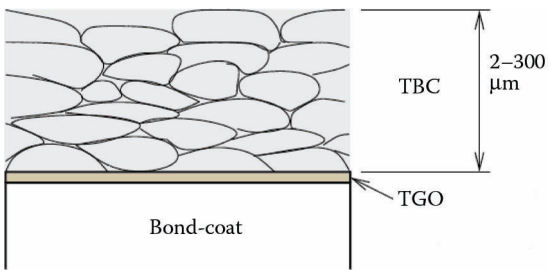
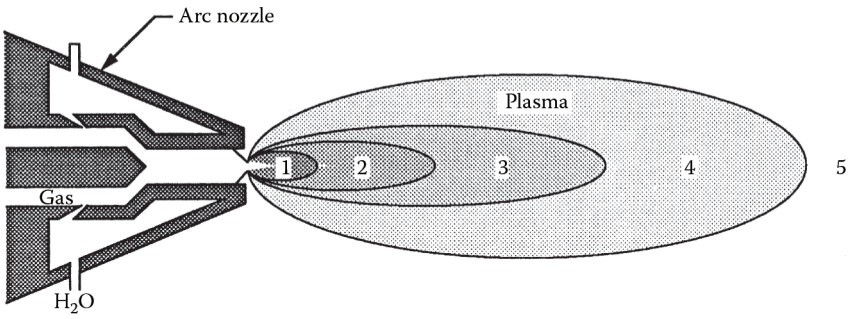


FIGURE 9.22 Schematic microstructure of a TBC obtained by APS on a TGO.

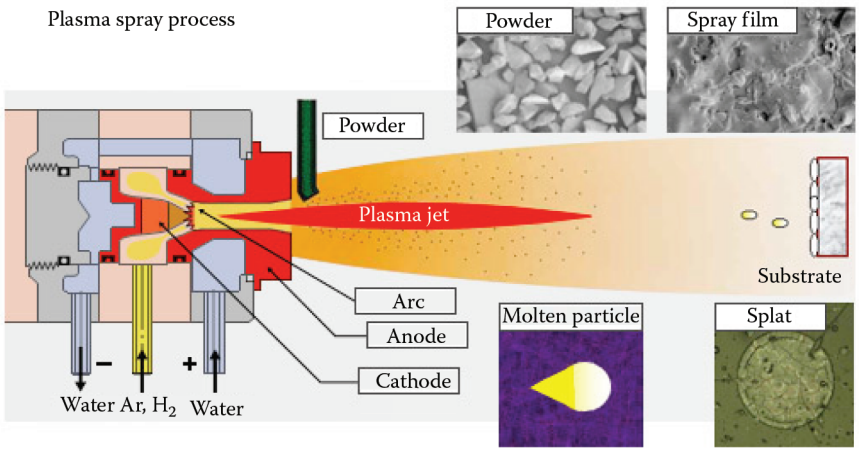
Example 9.3: Plasma Spraying: Multilayer Planar Oxygen Sensor

Here we present an example of how to use plasma spray technology to batch produce solid-state oxygen sensors based on yttria-stabilized ZrO_2 (YSZ) solid electrolyte films.[29] Plasma spraying is a particulate method geared toward



Region	Distance from nozzle	Temperature range (K)	Linear flow velocity (m/s)
1	< 1 cm	1.5×10^4 — 1×10^4	~ 400
2	1 < d < 5 cm	1×10^4 — 5×10^3	~ 400 — 200
3	5 < d < 10 cm	5×10^3 — 2×10^3	~ 200 — 100
4	10 < d < 20	2×10^3 — 1×10^3	< 100
5	20 < d	< 500	

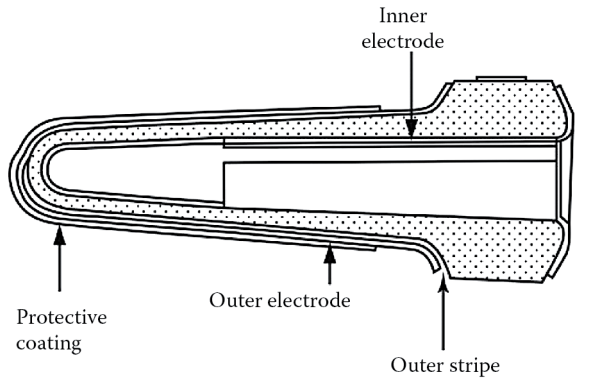
(a)



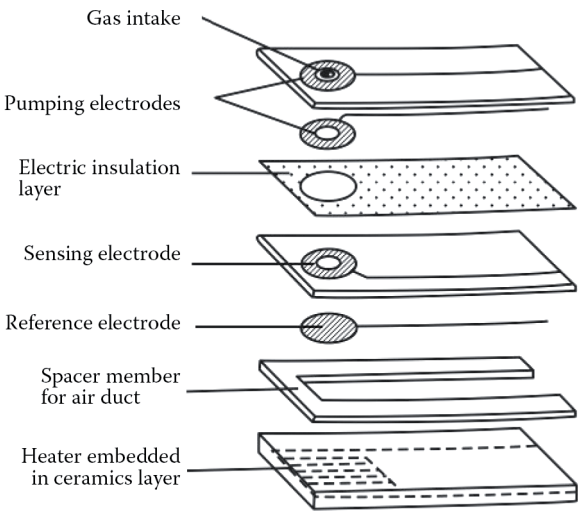
(b)

FIGURE 9.23 (a) Typical ranges of temperature and flow velocity with distance from the nozzle. (b) Plasma spray nozzle.

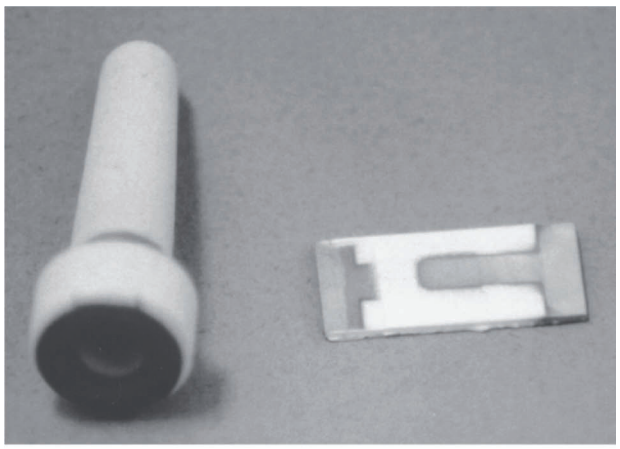
fast deposition of thicker films (>30 μm), and it might enable the batch fabrication of solid-state oxygen sensors at a fraction of the current cost ($2 vs. $12 and up). Traditional automotive solid-state oxygen sensors for combustion control (so-called *lambda probes*) are nose-shaped, 3D structures fabricated by sintering a molded green tape zirconia body (see Figure 9.24a). A plasma-sprayed, porous, gas-diffusion barrier, typically a spinel structure oxide, protects the resulting dense ceramic YSZ solid electrolyte and the silk-screened oxygen-sensing Pt electrode contacting it. Newer designs for oxygen sensors, called *wide range air-to-fuel ratio sensors*, incorporate two oxygen pumping cells and one potentiometric gauge. As we saw in Example 9.2, these sensors are planar and fabricated by laminating and cofiring at high temperatures several layers of ceramic green tape, some of which are metal coated and have openings through them (see Figures 9.19 and 9.24b).[27] Because both metal and ceramic layers constitute the sensor, the materials should be well prepared, and the cofiring process must be tightly controlled to avoid metal diffusion or reaction with the ceramic. Because of the process complexities, this sensor is too expensive. In Figure 9.24c, an alternative planar oxygen sensor fabricated using plasma-spray deposition is compared with the traditional cone-shaped oxygen sensor.[29,30] In this planar oxygen sensor, the metal electrodes are deposited by sputtering,



(a)



(b)



(c)

FIGURE 9.24 Solid-state oxygen sensors. (a) Traditional oxygen probe only protective, gas diffusion barrier is applied by plasma spray. (b) Wide-range air-to-fuel ratio sensor. (c) Planar oxygen probe, sensor, and gas diffusion layer made by plasma spray and compared with a classical oxygen probe. [(b) from Suzuki, S., T. Sasayama, M. Miki, M. Ohsuga, S. Tanake, S. Ueno, and N. Ichikawa. 1986. SAE Paper 860408. With permission.[27]]

and the plasma-sprayed YSZ film acts both as a gas-diffusion layer and as an oxygen-conducting electrolyte. The ionic conductivity of the plasma-deposited YSZ films does not reach the same level as YSZ electrolytes sintered at high temperatures because the films are not as dense. However, the relatively thin film geometry of the present sensor allows for using the plasma-sprayed YSZ film as an oxygen-pumping cell for the wide-range, air-to-fuel ratio sensor. The major challenge in the manufacture is the control of the porosity gradient in the gas diffusion barrier and electrolyte material, which is accomplished by controlling the size of the spraying powder and spraying conditions.[30] To change the temperature and flow velocity of the particles, we changed the distance between the nozzle and the substrate. With reference to Figure 9.23, to produce porous YSZ, we positioned the substrate somewhere between regions 4 and 5; to produce dense films, we positioned the substrate in region 4. The plasma-sprayed films adhere very well to the substrates and have exceptionally high integrity. The plasma method produces almost fully activated YSZ films onto a substrate carrying the thin film sputter-deposited Pt electrodes, and there is no need for additional sintering. This straightforward manufacture of solid-state oxygen sensors can be performed in large batches by using simple shadow masks and laser cutting the separate sensor elements.

We believe that this approach—a combination of thin and thick film methods—opens up the potential for planarization of many types of gas sensor devices. In the manufacture of chemical sensors, thick film technology on hybrid substrates is more prevalent than in the IC industry, and plasma deposition of all types of chemical sensor materials is, in the author's opinion, fertile ground for research. Plasma deposition of sensor materials such as ZrO_2, SnO_2, and ZnO on large inert carrier substrates could provide wafers coated with chemical sensor material very quickly and inexpensively.

Plasma Arc Deposition

In the plasma reactors discussed above, the energy necessary for evaporation and reaction is provided by an electric arc that forms between a stick-type cathode and a nozzle-shaped, water-cooled anode

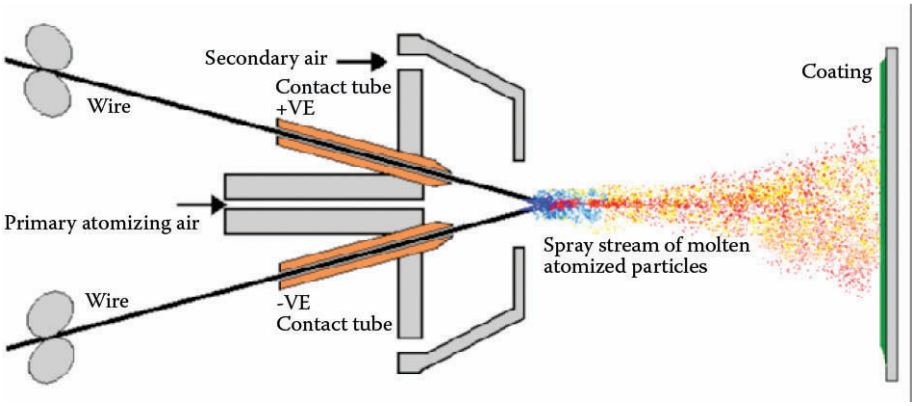


FIGURE 9.25 Plasma arc spray deposition process.

as illustrated in Figure 9.21. The material to be deposited is a fine powder suspended in a carrier gas and is injected into the plasma jet where the particles are accelerated and heated. The electrodes themselves do not participate in the reaction itself; in contrast, in the arc spray deposition process, illustrated in Figure 9.25, the electrodes are made of the metal to be vaporized. The electrodes are polarized in the presence of an inert background gas until the breakdown voltage is reached, and an arc is formed across the electrodes that vaporizes them. The molten material is propelled by compressed air toward the substrate surface. The impacting molten particles on the substrate rapidly solidify to form a coating. High spray rates and efficiency make plasma arc deposition a good tool for spraying large areas and high production rates. Disadvantages of the electric arc spray process are that only electrically conductive electrodes can be sprayed. The main traditional applications of the arc spray process are anticorrosion coatings of zinc and aluminum.

Plasmas and Nanoparticles

In the section above on spray pyrolysis, we discovered that this technique can be used for either film formation or for nanopowder generation. The same is true for the plasma equipment discussed above. Plasma temperatures are in the order of 10,000 K; thus, precursors injected in those plasmas are generally decomposed fully into atoms. To form nanoparticles from these, the atoms must then react or condense to form small solid particulates when cooled by mixing with a cool gas or by expansion through a nozzle. We saw an example of that in the ion cluster beam technology discussed in Chapter 7. In ion cluster beam technology the cooling occurs when the vapor exits a special evaporation cell. As shown in Figure 7.14, the heating of the evaporant in an evaporation cell with a small opening causes an adiabatic expansion of more than 100–10^{-5} or 10^{-7} mbar on exiting that cell. The expansion causes a sudden cooling, inducing the formation of atom clusters.

A very prominent application of arc deposition is that of fullerenes. The first large-scale synthesis of fullerenes was discovered in 1989 by Huffman and Kratschmer; their technique consisted of the arc evaporation of graphite electrodes via resistive heating within an atmosphere of approximately 100 atm of helium (see Volume III, Chapter 3). Spark discharge deposition was also the very technique that led to the discovery of carbon nanotubes in 1991. Carbon nanotubes were found in the carbon soot of graphite electrodes during arc discharge by using a current of 100 amps. In this process, the negative electrode sublimates because of the high temperatures caused by the discharge. Because nanotubes were initially discovered using this technique, it is still the most widely used method for nanotube synthesis. The yield for this method is up to 30% by weight, and it produces both single- and multiwalled nanotubes with lengths of up to 50 μm. The method is also used, for example, to produce fused silica. Silica sand is first vaporized in arc plasma at atmospheric pressure, and the resulting mixture of plasma gas and silica vapor is rapidly cooled by quenching with oxygen, thus ensuring the quality of the fumed silica produced.

In an RF induction plasma, energy coupling to the plasma is accomplished through the electromagnetic field generated by an induction coil. The plasma gas does not come in contact with electrodes, thus eliminating possible sources of contamination and allowing the operation of such plasmas with a wide range of gases, including inert, reducing, oxidizing, and other corrosive atmospheres. The working frequency is typically between 200 kHz and 40 MHz. Because the residence time of the injected feed droplets in the plasma is very short, it is important that the droplet sizes are small enough to obtain complete evaporation. The RF plasma method has been used to synthesize different nanoparticle materials, for example, various ceramic nanoparticles such as oxides and nitrides of Ti and Si.

Laser and Electron Beam Deposition

Introduction

Laser and electron beam machining are thermal techniques that are used as either subtractive tools in laser and electron beam etching (see Chapter 5) or as additive tools such as in the applications covered in this section, i.e., laser-assisted CVD (LCVD), SLS (a RP technique), and electron-assisted CVD (ECVD). Lasers also are used in stereolithography to make polymeric microstructures; this additive photochemical technology was reviewed in Chapter 8 and will not be repeated here. Pulsed laser deposition (Chapter 5) is also not covered here; although it uses intense laser radiation (thermal) to erode a target, it deposits the eroded material onto the adjacent substrate mechanically, not thermally.

Forming Processes with Laser Machining

Laser-Assisted Chemical Vapor Deposition

Laser-assisted chemical vapor deposition (LCVD) is an additive technique that involves the deposition of solids by localized chemical reaction using a laser beam (see Figure 9.26). The method has been used for growing 3D microstructures and thin films in elemental or compound form, and at least 17 elements of the periodic chart have been deposited (e.g., W, Ni, Ti).[31] The reactions are either pyrolytic,

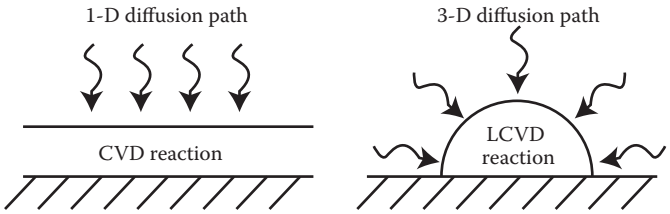


FIGURE 9.26 Comparison of a CVD and an LCVD reaction.

i.e., thermally activated, or photolytic, i.e., nonthermally activated. The latter method fills the need for a low-temperature deposition process. Because this chapter is about thermal forming, we will emphasize the former types of reactions.

The setup shown in Figure 5.25 is used to both laser etch and to laser deposit materials. Lasers and electron and ion beams add another important capability to materials processing, namely, the ability to process selected materials at specific sites and at low bulk temperatures. Of the different techniques, the laser is the more versatile. It can be operated in air, is not affected by surface charging, and can more easily be incorporated in a manufacturing assembly line.[32] With a laser source, it is possible to write a pattern on a surface directly by scanning the microsized light beam over the substrate in the presence of the suitable reactive gases. The species to be deposited is usually chemically encapsulated in a precursor gas with carbonyls or alkyl "backbones." Typical laser-driven deposition reactions are:

- Tungsten deposited by:

$$WF_6 + 3H_2 \rightarrow W + 6HF \quad \text{Reaction 9.1}$$

- Nickel deposited by:

$$Ni(CO)_4 \rightarrow Ni + 4CO \quad \text{Reaction 9.2}$$

- Silicon deposited by:

$$SiH_4 \rightarrow Si + 2H_2 \quad \text{Reaction 9.3}$$

Doped polysilicon lines with conductivities as low as 10^{-2}–10^{-3} Ω-cm may be obtained by adding PH_3 to the silane. Typical deposition rates are in the order of micrometers per second. By optimizing the process, very high deposition rates have been obtained in LCVD. For example, in the case

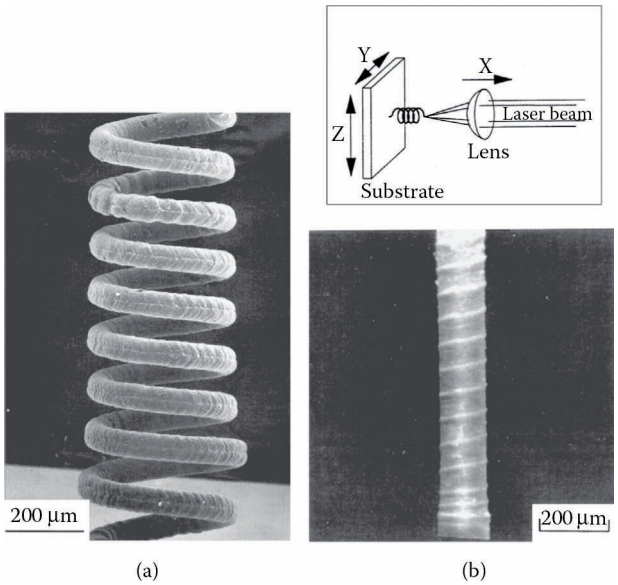


FIGURE 9.27 Examples of LCVD microfabricated shapes made by adjusting the focal point of a laser (tungsten helix) (a) and rotating goniometer (Si microsolenoid with tungsten helix) (b). [From Boman, M., H. Westberg, S. Johansson, and J.-Å. Schweitz. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92).* Travemunde, Germany: IEEE. With permission.[33]]

of silicon, deposition rates of 500 µm/s have been reported by using 100% SiH_4 at 1 atm, and the writing speed of tungsten on silicon substrates can exceed several centimeters per second.[33]

Not only "flat" lines but also more complex 3D structures such as fibers and springs can be grown by LCVD (Figure 9.27). This is accomplished by adjusting the focal point of the laser continuously by moving the substrate using a 3D linear micro-positioning system.[33] Microscale Si and boron rods and helical structures, freestanding tungsten coils, and a tungsten helix on a cylindrical silicon substrate (for a microsolenoid) have been demonstrated this way. For deposition of the tungsten helix on the Si rod, a rotating goniometer is used.

Table 9.2 lists examples of localized, electroless, and maskless laser depositions. Besides 2D and 3D writing patterns, there are two other site-specific actions that are possible by the laser irradiation technique: 1) the driving of dopants into semiconductor materials, and 2) the inducing of oxidation at a surface. Examples of the latter two laser machining applications are listed in Table 9.2 as well.

Selective Laser Sintering

Selective laser sintering (SLS®, registered trademark by DTM™ of Austin, TX) is a process that was patented in 1989 by Carl Deckard, a University of Texas graduate student.

SLS is an additive layered rapid manufacturing method that creates solid, 3D objects by using a high-power laser (for example, a carbon dioxide laser) to fuse small particles of plastic, metal, or ceramic powders into a mass representing a desired 3D object. A thin layer of powder material is laid down, and the laser "draws" on the layer, sintering together the particles hit by the laser. The layer is then lowered a small amount, and a new layer of powder is placed on top. This process is repeated one layer at a time until the part is complete. As shown

TABLE 9.2 Examples of Localized Laser Deposition, Doping, and Oxidation

Material Deposition		
Deposit	Precursor	Laser
Ti	$TiCl_4$	Ar^+
Ti/Al	$TiCl_4$, $Al(CH_3)_3$	Ar^+
Cr	$Cr(CO)_6$	Ar^+
Laser-induced doping		
Dopant/substrate or film	Precursor	Laser
B in Si	Evaporated B	Ruby
P in polysilicon	P-doped glass	N_2 dye
As in Si	AsH_3	XeCl
Laser oxidation of substrates		
Substrate	Oxide species	Laser
Ti	TiO_2	Nd:YAG
Zr	ZrO_2	CO_2

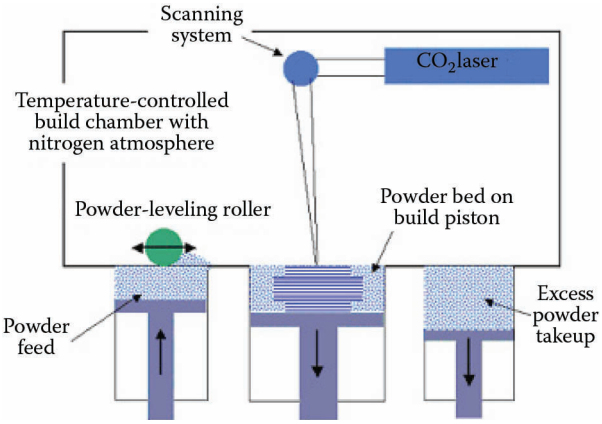


FIGURE 9.28 Selective laser sintering.

in Figure 9.28, an SLS machine consists of two powder magazines on either side of the work area. A leveling roller moves powder over from one magazine, crossing over the work area to the other magazine. The laser selectively fuses powdered material by scanning cross-sections generated from a 3D digital description of the part (for example, from a CAD file or scan data) on the surface of a powder bed. After each cross-section is scanned, the powder bed is lowered by one layer thickness; a new layer of material is applied on top; and the process is repeated until the part is completed.

The chief advantage of SLS over other RP methods such as stereolithography (SLA) revolves around material properties. Many materials are possible, and these materials can approximate the properties of thermoplastics such as polycarbonate, nylon, or glass-filled nylon.

A new RP method combines additive and subtractive techniques. The laser sintering/milling hybrid machine, the LUMEX 25C, developed by Matsuura Machinery Corporation, successfully combines free-form manufacturing and high-speed milling (see inset in the heading of this chapter). This is the world's first "metal laser modeling hybrid machining unit." The integration of laser sintering of metallic powder and high-speed cutting eliminates the finish machining operations. It layers a powder mixture (90% steel and 10% copper) over the part and then sinters it with a 300-W (500 W maximum) CO_2 laser. The edges of the part are milled after sintering five layers, with tolerances in the 25-µm range. For the generated part, a surface roughness of 15 µm can be obtained.

Conclusions

Additive laser machining presents several application opportunities in micromachining because of its versatility, site-specific operation, and RP capability. Ultrafast laser machining, when lasers become less expensive, will make laser micromachining yet more attractive. There are some obvious drawbacks with these direct write techniques; they are serial processes, and the deposition rate limits the speed of the micromanufacture. However, for RP, mold fabrication, and site-specific manufacturing, laser machining has a very bright future, and, as we have remarked before, some microsystems might carry a bigger price tag than an IC, making more expensive manufacturing technology acceptable.

Electron Beam Forming Processes: Electron Beam-Assisted CVD

Electron beam-induced metal deposition is a slow but flexible and versatile process for the fabrication of microstructures with high lateral accuracies. For example, using a precursor gas at a pressure of 2×10^{-2} mbar, W/C needles 0.2 µm in diameter have been deposited. The diameter of the needles is bigger than the electron beam diameter, which is less than 100 nm, because of electron scattering. To increase the growth rate, directing the precursor gas through a small nozzle onto the surface increases the local gas flow. The total gas flow needs to be kept below a maximum value to maintain the base pressure less than 10^{-5} mbar. For the deposition of other metals, numerous gases have been reported in the literature as precursors (mainly, for ion beam-induced deposition): $Al(CH_3)_3$ for Al, $C_7H_7O_2F_6Au$ for Au, and (methylcyclopentadienyl)trimethyl platinum for Pt.[34] A schematic of an electron beam fabrication setup is shown in Figure 9.29, along with an SEM photograph of W/C needles produced in this type of setup.[34] The deposition time of the needles shown was 1–2 min. Using the same type of technology, Matsui[35,36] demonstrated tips with a diameter of 15 nm for use in a tunneling microscope.

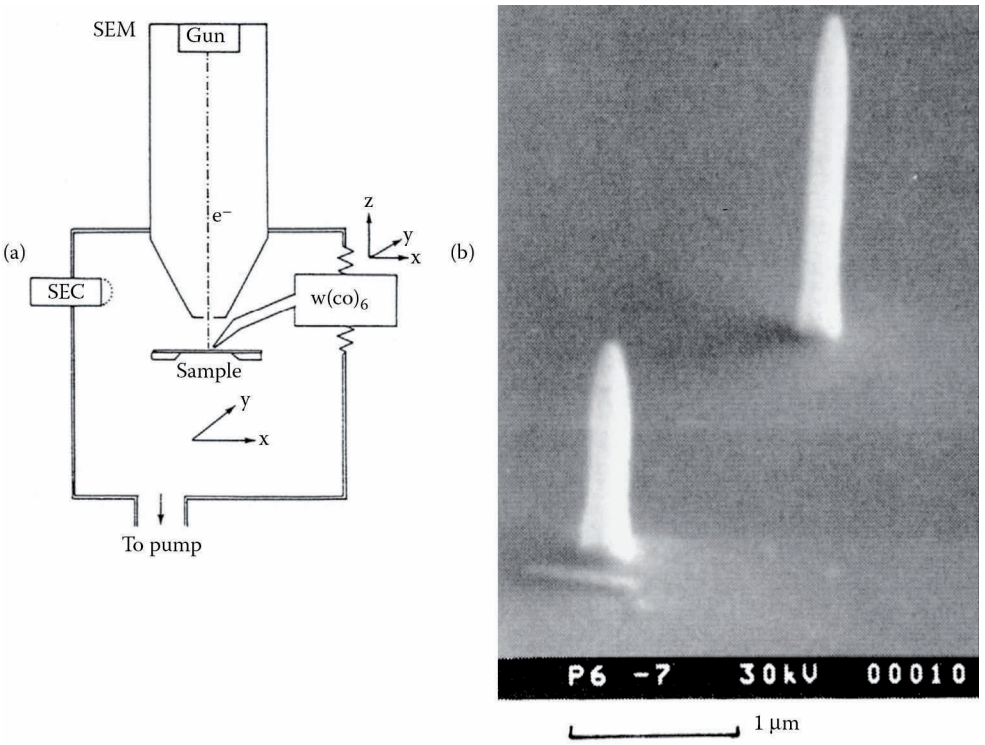


FIGURE 9.29 Electron beam machining setup (a) and example of an electron beam-formed W/C microstructure (b). [From Brunger, W. H., and K. T. Kohlmann. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92).* Travemunde, Germany: IEEE. With permission.[34]]

For processes such as controlled submicrometer repair of masks, focused ion beam-induced deposition appears to be the preferred technique, although both lasers[37,38] and electron beams[39] have been used as well.

Questions

Questions by Dr. Madou assisted by Mr. Omid Rohani, UC Irvine

9.1: Explain how salt-assisted spray pyrolysis (SASP) overcomes some of the limitations of traditional spray pyrolysis.

9.2: Explain the flame spray pyrolysis process. What are the advantages and disadvantages of using this method?

9.3: What are typical applications of thick film technology?

9.4: Suppose that you want to form a thin flat ceramic part with a large surface area and various holes and shapes cut into it. Which manufacturing method do you suggest for this purpose? Why?

9.5: Sketch a multilayer ceramic planar oxygen sensor. Explain how this sensor works.

9.6: What might the possible advantages be of plasma depositing (plasma spraying) sensor materials in the manufacture of chemical sensors?

9.7: Explain the plasma spray deposition manufacturing process briefly.

9.8: How do we get nanoparticles from plasma atoms?

9.9: Explain the selective laser sintering (SLS) process.

9.10: How do you convert your chemical vapor deposition (CVD) reactor into a reactor that produces nanoparticles?

9.11: Why does one need a pseudoplastic material to function as an ink in silk screening?

9.12: What is a doctor's blade, and where is it used?

9.13: Compare laser-assisted CVD (LCVD) with E-beam assisted CVD (ECVD).

9.14: Name some of the applications of TiO_2 powders.

References

1. Messing, G. L., S.-C. Zhang, and G. V. Jayanthi. 1993. Ceramic powder synthesis by spray pyrolysis. *J Am Cer Soc* 76:2707–26.
2. Mooney, J. B., and S. B. Radding. 1982. Spray pyrolysis processing. *Ann Rev Mater Sci* 12:81–101.
3. Tomar, M. S., and F. J. Garcia. 1988. Spray pyrolysis in solar cells and gas sensors. *Prog Cryst Growth Charact* 4:221–48.
4. Xia, B., I. W. Lengorro, and K. Okuyama. 2001 Particle formation processes of conventional spray pyrolysis (CSP) and salt-assisted spray pyrolysis (SASP). *Adv Mater* 13:1579.
5. Vemury, S., S. E. Pratsinis, and L. Kibbey. 1977. Electrically-controlled flame synthesis of nanophase TiO_2, SiO_2, and SnO_2 powders. *J Mater Res* 12:1031–42.
6. Lenggoro, I. W., K. Okuyama, J. F. d. l. Mora, and N. Tohge. 2000. Preparation of ZnS nanoparticles by electrospray pyrolysis. *Aerosol Sci* 31:121–36.
7. Deng, W., J. F. Klemic, X. Li, M. A. Reed, and A. Gomez. 2006. Increase of electrospray throughput using multiplexed microfabricated sources for the scalable generation of monodisperse droplets. *Aerosol Sci* 37:696–714.
8. Dzenis, Y. 2004. Spinning continuous fibers for nanotechnology. *Science* 304:1917.
9. Sutasinpromprae, J., S. Jitjaicham, M. Nithitanakul, C. Meechaisue, and P. Supaphol, P. 2006. Preparation and characterization of ultrafine electrospun polyacrylonitrile fibers and their subsequent pyrolysis to carbon fibers. *Polym Int* 55:825.
10. Zussman, E., X. Chen, W. Ding, L. Calabri, D. A. Dikin, J. P. Quintana, and R. S. Ruoff. 2005. Mechanical and structural characterization of PAN-derived carbon nanofibers. *Carbon* 43:2175.
11. Wang, Y., S. Serrano, and J. J. Santiago-Aviles. 2002. Conductivity measurement of electrospun PAN-based carbon nanofiber. *J Mater Sci* 21:1055.
12. Doshi, J. and H. D. Reneker. 1995. Electrospinning process and applications of electrospun fibers. *J Electrost* 35:151.
13. Shin, Y. M., M. M. Hohman, M. P. Brenner, and G. C. Rutledge. 2001. Electrospinning: A whipping fluid jet generates submicron polymer fibers. *Appl Phys Lett* 78:1149.
14. Theron, S. A., E. Zussman, and A. L. Yarin. 2004. Experimental investigation of the governing parameters in the electrospinning of polymer solutions. *Polymer* 45:2017.
15. Sharma, C. S., A. Sharma, and M. Madou. 2010. Multiscale carbon structures fabricated by direct micropatterning of electrospun mats of SU-8 photoresist nanofibers. *Langmuir* 26(4):2218–22.
16. Turon Teixidor, G., B. Park, P. Mukherjee, Q. Kang, and M. Madou. 2009. Modeling fractal electrodes for Li-ion batteries. *Electrochim Acta*, 54(24):5928–36.
17. Turon Teixidor, G., R. Gorkin, P. Tripathi, G. Bisht, M. Kulkarni, T. Maiti, T. Battacharyya, J. Subramaniam, A. Sharma, B. Park, and M. Madou. 2008. Carbon-MEMS as a substratum for cell growth. *J Biomed Mater* 3(3):034116.
18. Madou, M. J., and S. R. Morrison. 1989. *Chemical sensing with solid state devices.* New York: Academic Press.
19. Madou, M. 1994. *NIST special publication 865*. Gaithersburg, MD: NIST.
20. Lambrechts, M., and W. Sansen. 1992. *Biosensors: microelectrical devices.* Philadelphia: Institute of Physics Publishing.
21. Jabbour, G. E., R. Radspinner, and N. Peyghambarian. 2001. Screen printing for the fabrication of organic light-emitting devices. *J IEEE Selected Topics Quantum Electronics* 7:769–73.
22. Riemer, D. E. 1988. Analytical engineering model of the screen printing process: part I. *Solid State Technol* August:107–11.
23. Middlehoek, S., D. J. W. Noorlag, and G. K. Steenvoorden. 1983. Silicon and hybrid micro-electronic sensors. *Electrocomponent Sci Technol* 10:217–29.
24. White, N. M., and A. W. J. Cranny. 1987. Design and fabrication of thick film sensors. *Hybrid Circuits* 12:32–36.
25. Howatt, G. N. 1952. *Method of producing high-dielectric high-insulation ceramic plates.* US Patent 2582993.
26. Mistler, R. E., and E. R. Twiname. 2000. *Tape casting: theory and practice.* Westerville, OH: American Ceramic Society.
27. Suzuki, S., T. Sasayama, M. Miki, M. Ohsuga, S. Tanake, S. Ueno, and N. Ichikawa. 1986. Air-fuel ratio sensor for rich, stoichiometric and lean ranges. SAE Paper 860408.
28. Pfender, E. 1988. Fundamental studies associated with the plasma spray process. *Surf Coat Technol* 34:1–14.
29. Oh, S., and M. Madou. 1992. Planar-type, gas diffusion-controlled oxygen sensor fabricated by the plasma spray method. *Sensors Actuators B* B14:581–82.
30. Oh, S. 1994. A planar-type sensor for detection of oxidizing and reducing gases. *Sensors Actuators B* B20:33–41.
31. Eden, J. G. 1986. Photochemical processing of semiconductors: new applications for visible and ultraviolet lasers. *IEEE Circuits Devices Mag* 2:18–24.
32. Helvajian, H. 1995. *Microengineering technology for space systems.* Ed. H. Helvajian. El Segundo, CA: Aerospace Corporation.
33. Boman, M., H. Westberg, S. Johansson, and J.-Å. Schweitz. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92)*. Travemunde, Germany.
34. Brunger, W. H., and K. T. Kohlmann. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92)*. Travemunde, Germany: IEEE.
35. Matsui, S. 1987. Direct writing onto Si by electron beam simulated etching. *Appl Phys Lett* 51:1498–99.
36. Matsui, S., and K. Mori. 1986. New selective deposition technology by electron beam induced surface reaction. *J Vac Sci Technol* B4:299–304.
37. Bauerle, D. 1986. *Chemical processing with lasers.* Berlin, Germany.
38. Ehrlich, D. H., and J. Y. Tsao. 1983. A review of laser–microchemical processing. *J Vac Sci Technol* B1:969–84.
39. Kunz, R. R., and T. M. Mayer. 1987. Catalytic growth rate enhancement of electron beam deposited iron films. *Appl Phys Lett* 50:962–64.

10

Micromolding Techniques—LIGA

The Graduate, starring Dustin Hoffman, was released by Embassy Pictures in 1967. A memorable poolside scene from the movie:

Mr. McGuire:	*Come with me for a minute. I want to talk to you. I just want to say one word to you. Just one word.*
Ben:	*Yes, sir.*
Mr. McGuire:	*Are you listening?*
Ben:	*Yes sir, I am.*
Mr. McGuire:	**PLASTICS.**
Ben:	*Exactly how do you mean?*
Mr. McGuire:	*There is a great future in plastics. Think about it. Will you think about it?*
Ben:	*Yes I will.*

Outline

Introduction

Plastic Molding

LIGA—Background

LIGA and LIGA-Like Process Steps

Comparison of Master Micromold Fabrication Methods

Alternative Molding Materials in LIGA and Pseudo-LIGA

Examples

Questions

References

Introduction

A wide variety of micro- and nanofabrication techniques such as x-ray lithography (Chapter 2), computer numerical control (CNC) machining (Chapter 6), laser machining (Chapter 5), and many others are either too slow or too expensive for the mass production of inexpensive devices. This is where micromolding/replication technology comes into play because of its capacity for very large volume and inexpensive manufacture.[1]

A key example involving micromolding is LIGA, which, we learned in Chapter 2, is the German acronym for x-ray lithography (*x-ray lithographie*), electrodeposition (*galvanoformung*), and molding (*abformtechnik*). The process involves a thick layer of x-ray resist (from micrometers to centimeters), high-energy x-ray radiation exposure, and development to

arrive at a three-dimensional (3D) resist structure. Subsequent metal deposition fills the resist mold with a metal, and, after resist removal, a freestanding metal structure results.[2] The metal shape may be a final product or serve as a mold insert for precision plastic molding. Molded plastic parts may in turn be final products or lost molds (see Figure 2.9). The plastic mold retains the same shape, size, and form as the original resist structure but is produced quickly and inexpensively as part of an infinite loop. The plastic lost mold may generate metal parts in a second electroforming process or generate ceramic parts in a slip casting process.

The bandwidth of possible sizes in all three dimensions renders LIGA useful for manufacture of microstructures (micrometer and submicrometer dimensions) and packages for these microstructures (millimeter and centimeter dimensions), and even for the connectors from those packages to the "macro world" (e.g., electrical, through-vias or physical and gas, in- and outlets). Once LIGA was established in the research community, interest in other micro- and nanomolding/replication methods became more pronounced. Given the cost of the LIGA equipment, various LIGA-like (pseudo-LIGA) processes took center stage. These pseudo-LIGA methods involve micromolding/replication of masters created by alternate means such as deep reactive ion etching (DRIE) and novel deep ultraviolet (UV) thick photoresists (e.g., SU-8). This more generalized lithography and replication procedure is illustrated in Figure 10.1.

Micro- and nanofabrication of molds is reshaping manufacturing approaches for a wide variety of mass-produced small parts. Frequently, integrated circuit (IC)-based batch microfabrication methods are considered together with more traditional, serial machining methods. In this evolution, LIGA and pseudo-LIGA processes constitute "handshake-technologies," bridging IC and classical manufacturing technologies. The capacity of LIGA and pseudo-LIGA for creating a wide variety of shapes from different materials makes these methods akin to classical machining with the added benefit of unprecedented aspect ratios and absolute tolerances rendered possible by lithography or other high-precision mold fabrication techniques.

LIGA and LIGA-like processes can be used directly to fabricate prototype microdevices, but, for commercial-scale manufacture of microdevices, replication by molding from a LIGA or LIGA-like mold is required. If the feature sizes are larger than

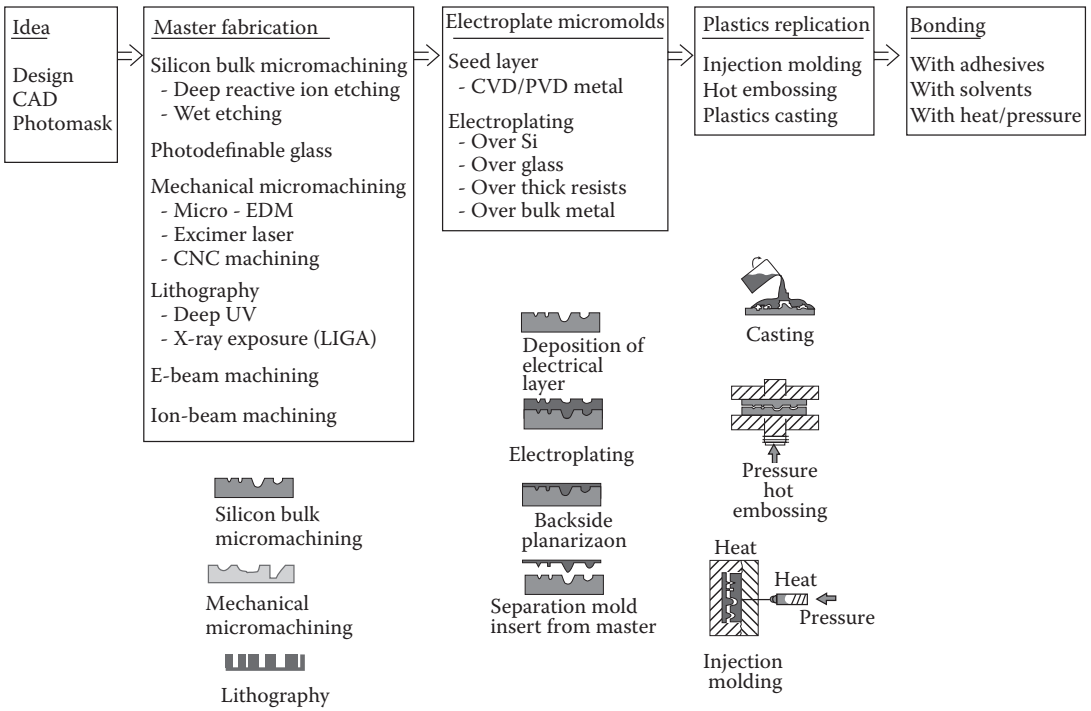


FIGURE 10.1 Process flow for plastic microfabrication.

50 μm and several levels of feature depths are involved, even CNC machining may be used to manufacture a prototype device. For example, in Figure 10.2, we show a CNC-machined plastic (polycarbonate) fluidic compact disc (CD) platform prototype. Compared with LIGA or LIGA-like processes, CNC machining does not provide as good a surface finish or dimensional control (see Chapter 6). However, it does not have any material limitation; various metals and plastics can be used. For feature sizes smaller than 50 μm (but larger than several micrometers), ultraprecision milling or laser ablation can be used. These prototype processes, just like LIGA and LIGA-like processes, are all intensive, slow, and costly, and for these microfabrication techniques to become economically viable, microstructures must be able to be replicated/micromolded successively, in an "infinite loop," without remaking the master mold. For example, in the plastic molding process, a metal structure produced by metal deposition serves as a mold insert and is used over and over without remaking the metal master. For mass-produced plastic devices, one of the following molding techniques is suitable: liquid resin molding, thermoplastic injection molding (IM), and compression (embossing/imprinting) molding. Our results of IM of a plastic fluidic platform, as shown in Figure 10.2, from a micromachined mold using pseudo-LIGA is shown in Figure 10.6.

In this chapter, we first review the most popular industrially available polymer molding techniques, and then, as examples that push the technology to its extremes, we detail LIGA and LIGA-like processes.

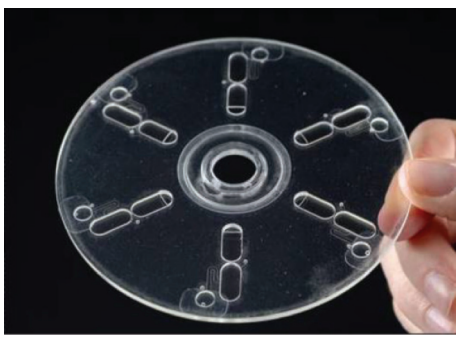


FIGURE 10.2 CNC-machined plastic lab CD platform: prototype cell disrupter disk, made from machined plastic disks and cut transfer adhesive films. (UCI and RotaPrep, Drs. Madou and Kido.)

The miniature features and the high-aspect ratios of the mold inserts generated from LIGA and LIGA-like methods present new problems compared with the molding and demolding processes of macrocomponents and compared with the production of small but low-aspect-ratio products such as CDs. A variety of the special processes developed to accommodate these new needs are reviewed. We start with a historical introduction to LIGA and analyze the process steps. The different applications and technical characteristics of x-rays from a synchrotron are discussed, and then we present an introduction to the crucial issues involved in making x-ray masks and their alignment for LIGA. This is followed by a treatise on LIGA substrate choice, resist requirements, exposure, development, electroplating, and finally molding. The electroplating and molding sections apply to LIGA and pseudo-LIGA approaches. The section on LIGA is followed by a review of competing mold fabrication techniques and a short treatise on ceramic molding.

Micromolding with an elastomeric mold (e.g., polydimethylsiloxane), or soft lithography, was reviewed in Chapter 2, where we stressed its low cost and convenience in research. Compared with other micromolding techniques reviewed in the current chapter, soft lithography is especially useful for thin microstructure fabrication.

Bonding of plastic molded microparts, including plastic welding, organic solvent bonding, and bonding with thermoset adhesives, is covered in Volume III, Chapter 4 in its discussion of packaging techniques.

Plastic Molding

Liquid Resin Molding Techniques

Introduction

Because mold inserts made by photolithography techniques are typically limited to rather soft metals like nickel or brittle materials such as silicon and glass, micromolding based on low-viscosity liquid resins (instead of high-viscosity polymer melts) is a very attractive approach. During liquid resin molding, the low-viscosity reactive polymer components are mixed shortly before injection into the mold

cavity, and polymerization takes place during the molding process. Under liquid resin molding techniques, we distinguish between reaction IM (RIM) and transfer molding.

Reaction Injection Molding and Transfer Molding

Two typical setups for RIM and transfer molding are shown in Figure 10.3a and b, respectively.[3] In the RIM setup shown in Figure 10.3a, two or more highly reactive liquid resins impinge and mix inside a mixer and are injected into the mold. In RIM, the polymerization is usually mixing activated; therefore, it can quickly react at room temperature or slightly above and convert from liquid to solid. The transfer molding setup in Figure 10.3b, for a thermally activated system, is simpler and less expensive. Here, the reactive species have been premixed at room temperature, and the mixture is poured into the "transfer pot." A plunger squeezes the mixture through the sprue into the mold cavity. To cure the mixture, the mold is heated or radiated, depending on the resin system. With the lowest viscosity resins, vacuum-assisted transfer molding is used, which is even less expensive than mechanical transfer molding.

Two complications associated with RIM are 1) an internal mold release agent must be used and 2) polymer shrinkage during polymerization. Release agents are added to the bulk polymer and tend to reduce the thermomechanical properties of the molded polymer. External mold release agents, coating the surfaces of the mold insert, are unsuitable because they are difficult to introduce into all the microfeatures of the mold insert. Because the resin must be fully cured at high temperatures, the cycle time is typically more than 10 min. Hanemann et al.[4] developed a photoinitiated RIM process by using the methylmethacrylate (MMA) resin. They found that the cycle time can be greatly reduced (to 2–3 min), but other disadvantages, such as internal mold release agent and polymerization shrinkage, still exist.

Summary: Pros and Cons of Liquid Resin Molding

The pros and cons of RIM are summarized in Table 10.1. Typical materials are optically clear reactive liquid resins, e.g., epoxies, urethanes, silicone rubbers, cross-linkable acrylics, and hydrogels.

Injection Molding

Introduction

Injection molding (IM) is based on heating a thermoplastic material until it melts, thermostatting the mold parts, injecting the melt with a controlled injection pressure into the mold cavity, and cooling the manufactured goods. IM is probably the most widely used technique in macroscopic production of polymer parts. Besides conventional IM, one variant of IM—a good candidate for microfabrication—is

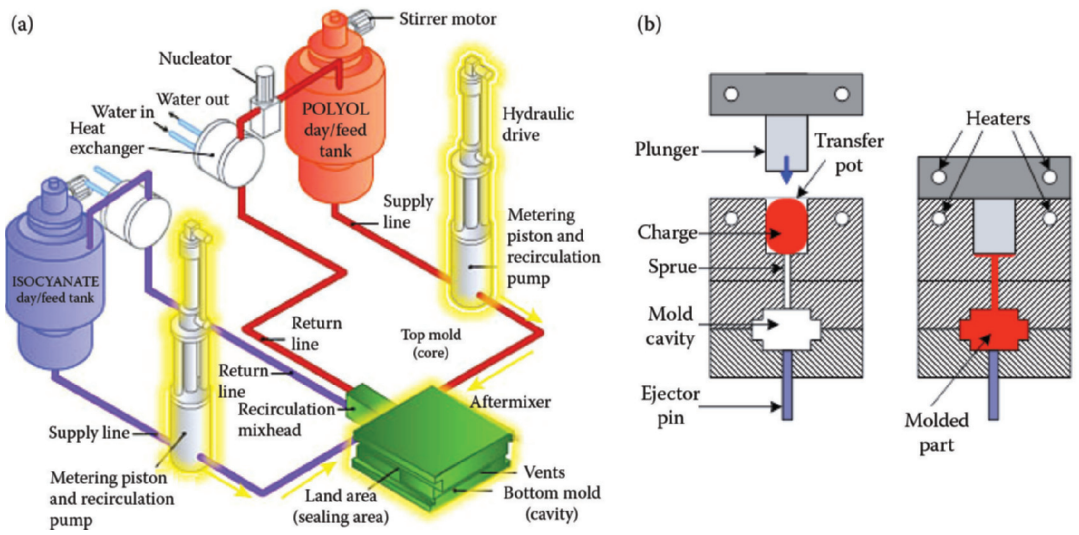


FIGURE 10.3 Two typical setups for RIM (a) and transfer molding or TM (b).

TABLE 10.1 Pros and Cons of Liquid Resin Molding

Pros	Cons
Ease of mold filling (low η)	Long cycle time (minutes, sometimes even hours)
Low stress on master (low η)	Polymerization shrinkage
High chemical and thermal resistance (because of the cross-linking)	Contamination (resin residue if the reaction is not 100%)
Replication extremely good for small, high-aspect ratio, and 3D features	Production cost medium to high

detailed here[3]: thin-wall IM. A typical injection compression molding setup is shown in Figure 10.4. The mold insert in IM, as in RIM, must have extremely smooth walls to be able to demold the plastic structures. Such smooth walls cannot be produced with a classical mold-making technique such as spark erosion. For some polymers, such as polymethylmethacrylate (PMMA), the polymer pellets need to be dried for 4–6 h at 70°C before use in IM because PMMA absorbs water in air (up to 0.3%), leading to poorer wall quality.

Conventional Injection Molding

In conventional IM of microparts, the mold remains closed, and the walls are kept at a uniform temperature above the glass transition temperature (T_g) of the polymer; thus, the injected molten plastic mass does not harden prematurely. To be able to fill very small mold features with aspect ratios higher than 10, the temperature of the tool holding the insertion mold should be higher than what is typically used in IM applications of parts with lower aspect ratios. However, one should stay well below the melting temperature (e.g., 170°C for PMMA with a melting temperature of 240°C) because the microstructure might show temperature-induced defects if one works at a temperature closer to the melting temperature.[5]

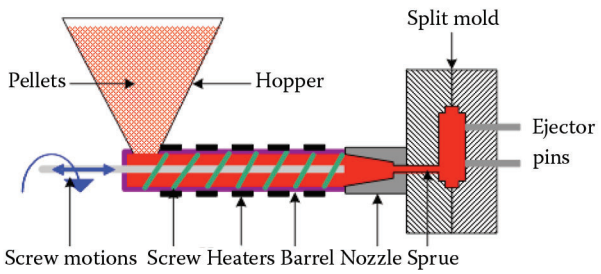


FIGURE 10.4 Typical injection molding setup.

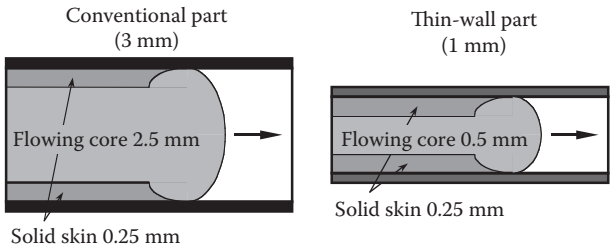


FIGURE 10.5 Schematics of solid skin buildup in conventional and thin-wall injection molding. (OSU, Drs. Madou and Lee.)

One of the most important factors for replication fidelity in IM is the thickness of the skin, which forms on the mold surface during filling of the cavity (see Figure 10.5).[3]

The injection pressure must penetrate this skin to replicate the shape of the master accurately. The faster the injection, the thinner this skin on the insert mold surfaces. The upper shapes of the master forming the bottom of the parts are usually perfectly formed. It is the bottom of the master and especially inner corners that cause problems. Because in conventional IM the mold is kept closed during the entire mold-filling process, high clamping forces on the mold parts are required, and the method is used for relatively thick substrates with high replication demands and low cost. Parts made this way often end up having high molded-in stresses.

Conventional thermoplastic IM is the technique used, for example, to make compact discs (CDs), involving features about 0.1 μm in depth and minimum lateral dimensions of 0.6 μm; that is, an aspect ratio of 0.16. CDs are made by IM of polycarbonate in a cavity formed between a mirror surface block and a stamper (i.e., mold or master). Conventionally, electroforming from a glass master generates nickel stampers. The principal attributes required of a stamper substrate are toughness, thermal shock resistance, thermal conductivity, and hardness. For IM in particular, toughness, thermal shock resistance, and thermal conductivity are critical. Thermal shock resistance is the product of a material's modulus of rupture or tensile strength and its thermal conductivity, divided by the product of its coefficient of thermal expansion and its modulus of elasticity. In Table 10.2, we compare some materials for their merit as stamper substrates. Ion machining is one of the methods that can be used to make mold

Table 10.2 Materials Selection for Stamper/Insert Mold Substrates

Substrate	Knoop Hardness (kg/mm^2)	Fracture Toughness ($MPa\ m^{1/2}$)	Thermal Shock Resistance (W/m)	Thermal Conductivity (W/m K)	Ion Machinability (Surface Quality)
Nickel	100	~100	7,138	80	Poor
Al_2O_3	2,100	4	3,225	30	Very good
SiC	2,500	3	19,149	90	Excellent
Glassy carbon	500	2	135,517	120	Excellent
Corning 9647 (glass)	450	1.3	546	2.5	Excellent

Source: Bifano, T. G., H. E. Fawcett, and P. A. Bierden. 1997. Precision manufacture of optical disc master stampers. *Precision Eng* 20: 53–62.[6]

inserts (see Figure 10.1 under Master fabrication), and surface quality after ion etching is also listed (ion machinability).

Nickel CD stampers are normally 138 mm in diameter and 0.3 mm thick with an average roughness less than 10 nm. For ceramic stampers, a thickness of 0.9 mm is chosen to avoid fracture in IM. Typical conditions for IM are summarized in Table 10.3.

IM of parts with small features and low-aspect ratios (e.g., CDs) have been widely applied. The main challenge in microelectromechanical system (MEMS) and nanoelectromechanical system (NEMS) replication attempts is to extend this technique to the fabrication of components with the same small feature size but much larger aspect ratios, as needed in many medical and biochemical applications (see later discussions of LIGA and pseudo-LIGA).

Thin-Wall Injection Molding

Thin-wall IM is basically high-speed IM.[3,7] Speeds can be 10 times faster than in conventional IM, reducing the thickness of the polymer skin on the mold (see Figure 10.5). The polymer melt can quickly be injected into the mold to fill all the detail structure of the cavity before solidification occurs.

In our own thin-wall IM experiments for replication of BioMEMS devices, such as the fluidic CD (about 1 mm thick)[3,8] shown in Figure 10.6, we have used a Sumitomo 200-ton high-pressure and high-speed machine.[7] Only at injection speeds of 50 mm/s or higher does the quality of the replication become adequate. A birefringence technique was used to qualitatively examine the stress in the polymer parts as demonstrated in Figure 10.7. Samples showing sharp color contrast have large molded-in stresses. At the lower flow rates the residual stress is higher because, during the flow, the polymer has been stretched and solidified, meaning that there is a lot of flow-induced stress. Increasing the flow

Table 10.3 Typical Conditions for Injection Molding

Mold temperature	85°C
Polycarbonate temperature	330°C
Clamping force	60 tons
Injection time	1 s
Cooling time	2 s

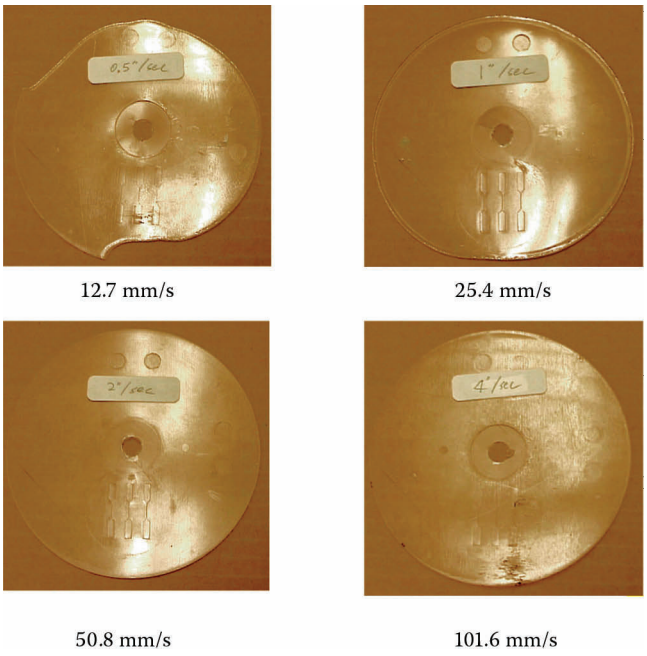


FIGURE 10.6 Microfluidic CD devices made at different injection rates (OQPC). (From Lee, L. J., C.-H. Shih, Y.-J. Juang, J. Garcia, M. J. Madou, and K. W. Koelling. 1999. *Novel microfabrication options for BioMEMS conference.* San Francisco: The Knowledge Foundation;[3] and Madou, M. J., Y. Lu, S. Lai, Y.-J. Juang, L. J. Lee, and S. Daunert. 2000. *Proceedings of the IEEE: solid-state sensor and actuator workshop.* Hilton Head Island, SC.[8])

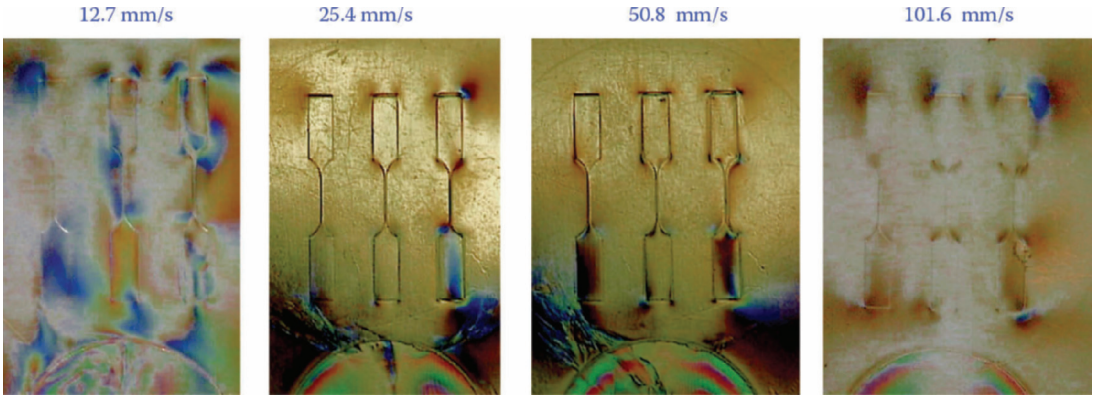


FIGURE 10.7 Birefringence of microfluidic devices made at different injection rates. (From Madou, M. J., Y. Lu, S. Lai, Y.-J. Juang, L. J. Lee, and S. Daunert. 2000. *Proceedings of the IEEE: solid-state sensor and actuator workshop.* Hilton Head Island, SC;[8] and Lee, L. J., C.-H. Shih, Y.-J. Juang, J. Garcia, M. J. Madou, and K. W. Koelling. 1999. *Novel microfabrication options for BioMEMS conference.* San Francisco: The Knowledge Foundation.[3])

rate decreases the amount of residual stress because the polymer can quickly fill in the mold and relax before solidification occurs. Nevertheless, there is some remaining stress that may cause warpage or less chemical resistance. By comparison, in casting there is essentially no flow: it is stress free.

A comparison of sample birefringence from different replication methods is given in Figure 10.8.[3] One observes that reactive casting with an epoxy resin causes the least amount of residual stress (least sharp color contrasts).

Another example of thin-walled structures made with thin-wall IM is shown in Figure 10.9. High injection speed is needed to completely fill the mold cavities for a thin-wall relay case (0.45 mm thick, flow length is 50.4 mm).

Injection Molding Equipment

For manufacturing MEMS structures, conventional IM machines used for CD fabrication are expanded with special features for the molding of microsized parts or parts with high-aspect ratio. The IM machine periphery then includes a vacuum unit for the evacuation of the split mold cavity and a precise temperature control unit. It is desirable to use one temperature control cycle on each surface of the split mold (see Figure 10.4). The flow temperatures are adapted to the respective half of the tool or to the molding and demolding process cycle to achieve a relatively short overall cycle time and a homogeneous tool temperature. To keep the cycle time as short as possible, the thermal mass of the tool sections to be heated is minimized and

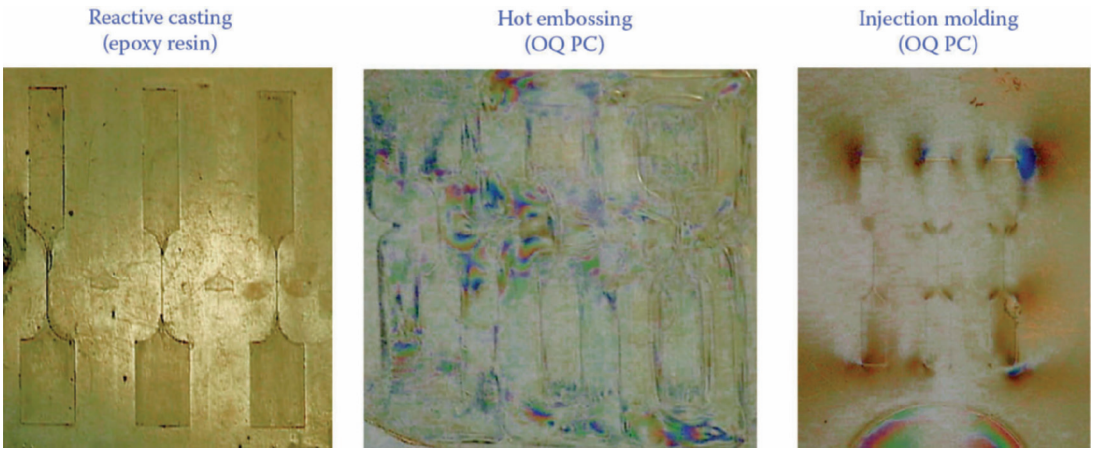


FIGURE 10.8 Comparison of birefringence patterns of microfluidic platforms made by different replication methods. (From Lee, L. J., C.-H. Shih, Y.-J. Juang, J. Garcia, M. J. Madou, and K. W. Koelling. 1999. *Novel microfabrication options for BioMEMS conference.* San Francisco: The Knowledge Foundation.[3])

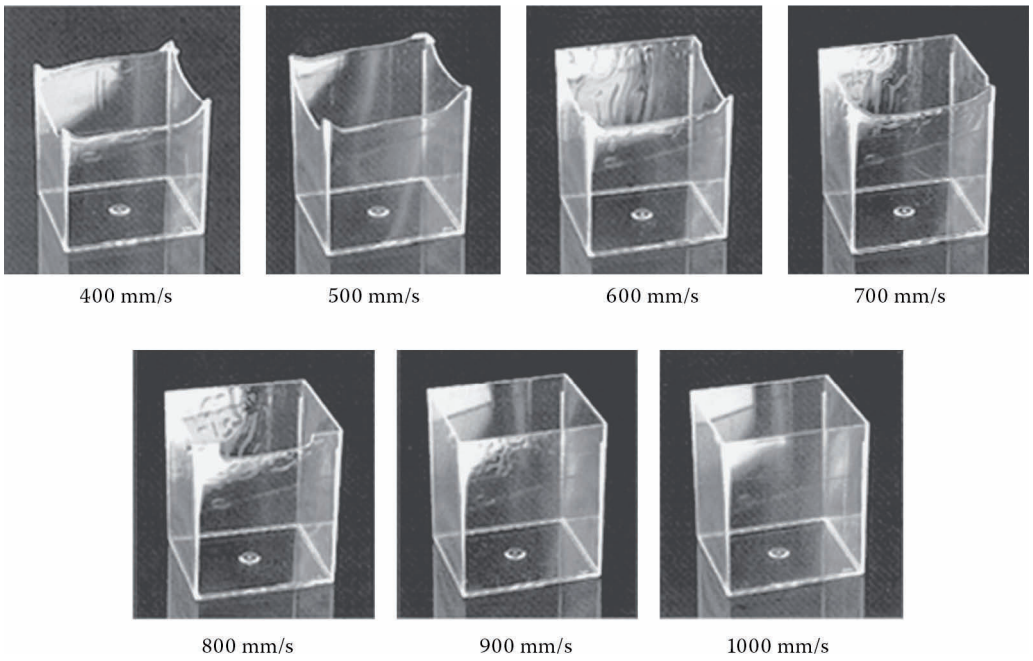


FIGURE 10.9 Relay case made by thin-wall injection molding at different injection rates (Nissei). (Courtesy Dr. Jim Lee, OSU.)

thermally insulated from the adjacent assemblies. The guiding mechanisms of the tool halves and the ejector system may not be fitted too loosely because small transverse movements might damage the microstructures during demolding. The machines and tools to be evacuated have to comply with tolerances in the micrometer range.[9] Demolding of the microstructures takes place by means of ejector pins, which are located on the sole of the mold tool, the guiding mechanism, or directly at extremities of the microstructure.[9] Under laboratory conditions, the microstructures are removed manually. In small series production, a handling robot is used. It may be equipped with a feed picker. In addition, the machines used for small series and mass production are provided with a granulate drying, supply, and metering system such that fully automatic molding in personnel-free shifts is possible.[10]

Commercial IM machines designed for microsized parts or parts with microsized features are available from Battenfeld (http://www.battenfeld.de), Boy Machines (http://www.dr-boy.de/rightmachine.htm), Engel (http://www.engel.at), Ferromatik Milacron (http://www.ferromatik.com), Murray (http://www.murrayeng.com), and Nissei (http://www.nisseiamerica.com). Figure 10.10 shows a photograph of Battenfeld's Microsystem 50.

Summary: Pros and Cons of Injection Molding

Pros and cons of IM are summarized in Table 10.4. Example materials in IM include polycarbonate (PC), polymethylmethacrylate (PMMA), polystyrene (PS), polyvinylchloride (PVC), polypropylene (PP), polyethylene (PE), and polyacrylnitrilbutadienstyrol (ABS).

Compression Molding (Also Relief Imprinting or Hot Embossing)

Introduction

The basic principle of compression molding (also *hot embossing* and *relief imprinting*) is that a polymer substrate is first heated above its glass transition temperature, T_g (or softening temperature). A mold (or master) is then pressed against the substrate, fully transferring the pattern onto it (embossing). After a certain time of contact between the mold and the substrate, the system is cooled down below T_g, followed by separation of the mold and the substrate (de-embossing). Importantly, the hot embossing process can be achieved in either a cyclic process or a continuous process. Compression molding[11,12] provides several advantages compared with IM, such as relatively low costs for embossing tools, simple process, and high replication accuracy for small features.

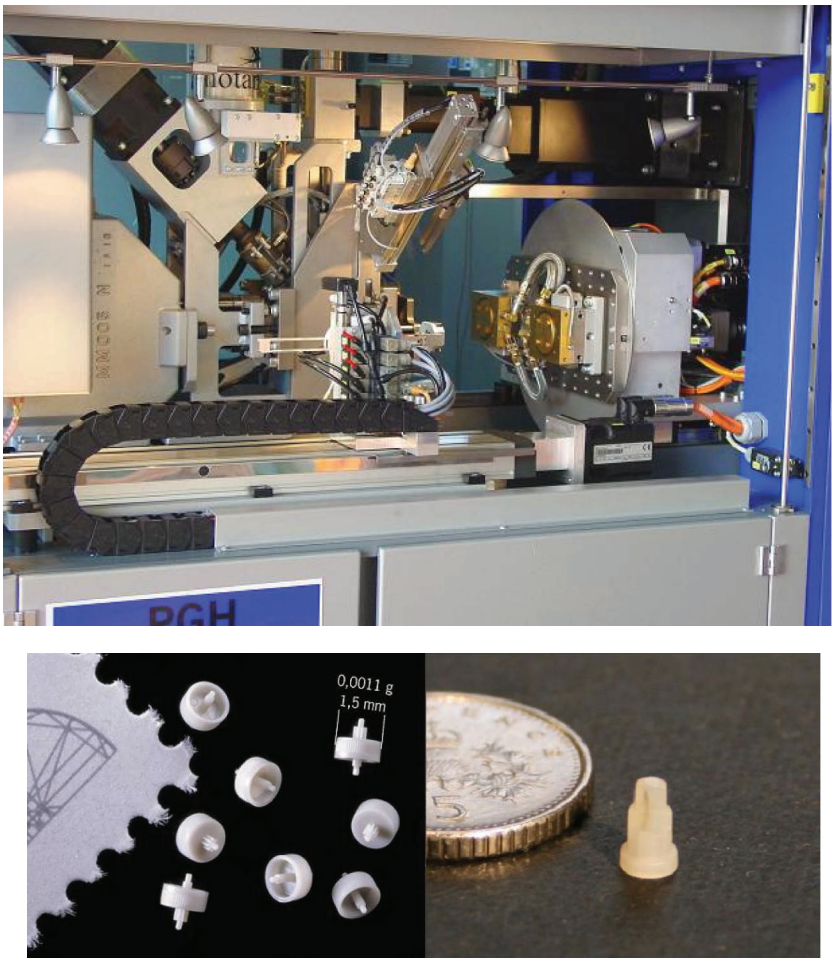


FIGURE 10.10 (a) Detail of the Battenfeld Microsystem 50 IM machine. (b) Examples of micromolded components with microfeatures.

A typical setup for cyclic and continuous embossing is shown in Figure 10.11a and Figure 10.11b, respectively.[3] In a cyclic process (Figure 10.11a), a metal master is put in a hydraulic press, and applying the appropriate force, thus replicating the structure from the master to the polymer, stamps a heated polymer sheet. This constitutes a low-cost method for making prototypes. For mass production, a continuous process is preferred (Figure 10.11b). Here, a polymer sheet stretches through a temperature chamber, and several masters, mounted on a conveyor belt, continuously produce parts. The process also may incorporate a lamination station to enclose certain features. This continuous-type process is very much in keeping with a vision for continuous manufacture of MEMS and NEMS devices.

TABLE 10.4 Pros and Cons of Injection Molding

Pros	Cons
Good for small structures with low aspect ratio, e.g., CD and DVD	Only low-molecular-weight polymers (may reduce mechanical and thermal strength)
Good for large, high-aspect ratio, and 3D features	More expensive equipment
Excellent dimensional control	Cyclic process only
Short cycle time (as low as 10 s)	High stress on master
High productivity	High residual stresses on molded parts
Closed mold process enables packing pressure application	

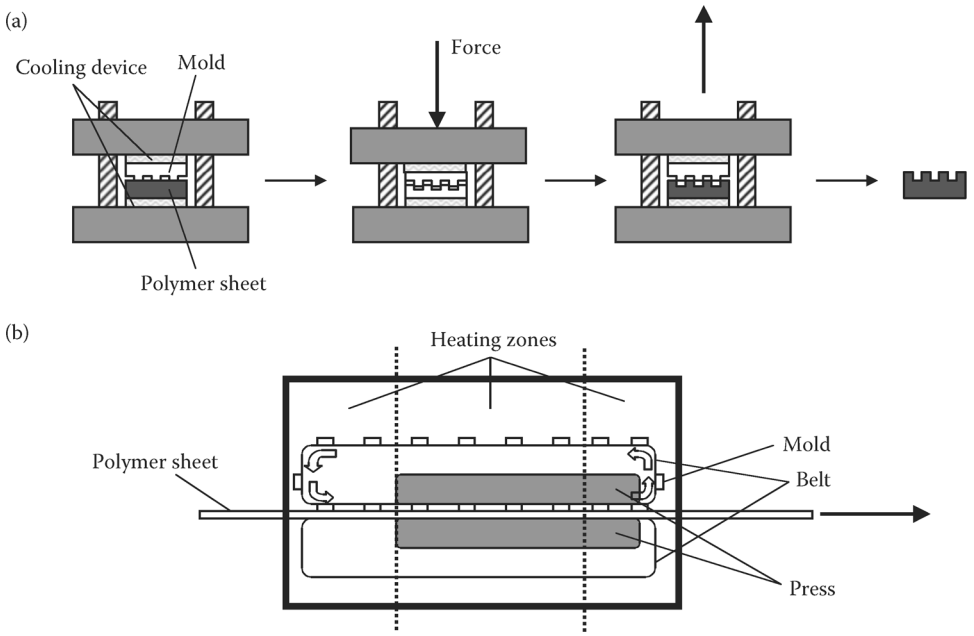


FIGURE 10.11 Schematic of cyclic (a) and continuous (b) embossing processes. (From Lee, L. J., C.-H. Shih, Y.-J. Juang, J. Garcia, M. J. Madou, and K. W. Koelling. 1999. *Novel microfabrication options for BioMEMS conference.* San Francisco: The Knowledge Foundation.[3])

Hot Embossing

As seen from Figure 10.11a, hot embossing takes place in a machine frame similar to that of a press. The force frame delivers the embossing force—on the order of 5–20 tons. An upper boss holds the molding tool, and the lower boss holds the substrate. Processing parameters include thermal cycle, compression force, and compression speed. The temperature difference between embossing and de-embossing determines the thermal cycle time, typically from 25–40°C. In principle, one could, after hot embossing, cool down the whole device to room temperature before de-embossing, or, at the other extreme, one could de-emboss just below or at the glass-transition temperature. A compromise is needed; the quality of the replication may not be good if one tries to remove the master when the polymer is still soft, whereas cooling all the way down to room temperature takes too long. A small temperature cycle leads to smaller induced thermal stress. Such a smaller temperature cycle also reduces replication errors as a result of different thermal expansion coefficients of tool and substrate. By actively heating and cooling the upper and lower bosses, one gets cycle times of about 5 min. A vacuum for hot embossing ensures longer lifetime for the mold, absorbs any water released during the embossing process from the polymer material, and prevents bubbles from entrapped gases. A Jenoptik hot embossing machine is shown in Figure 10.12.

Replication of micro- and nanosized structures has been successfully achieved using hot embossing.[12–18] Adding an antiadhesive film to reduce the interaction between the mold and the replica during embossing

FIGURE 10.12 JENOPTIK HEX 04 hot embossing system manufactured by Jenoptik Mikrotechnik GmbH (http://www.jenoptik.com).

Table 10.5 Pros and Cons of Compression Molding (Imprinting or Hot Embossing)

Pros	Cons
Low polymer flow	More difficult for structures with high-aspect ratio (near T_g processing)
High-molecular-weight polymers (with better mechanical and thermal properties)	Less dimensional control (open mold process)
Simple process	Planar features only
Continuous or cyclic (see Figure 10.7)	High residual stresses on molded parts
Good for small structures	Difficult for large parts and multiple feature depth (too high a pressure and temperature are required)

has also been studied.[19,20] Instead of the conventional nickel molds, the possibility of using silicon molds because of their excellent surface quality and easy mold release has been demonstrated.[21,22] Also, the use of a plastic mold in the embossing process was recently illustrated.[23]

At the Kernforschungszentrum Karlsruhe, or KfK, hot embossing is mainly used for small series production, whereas IM is applied to mass production.

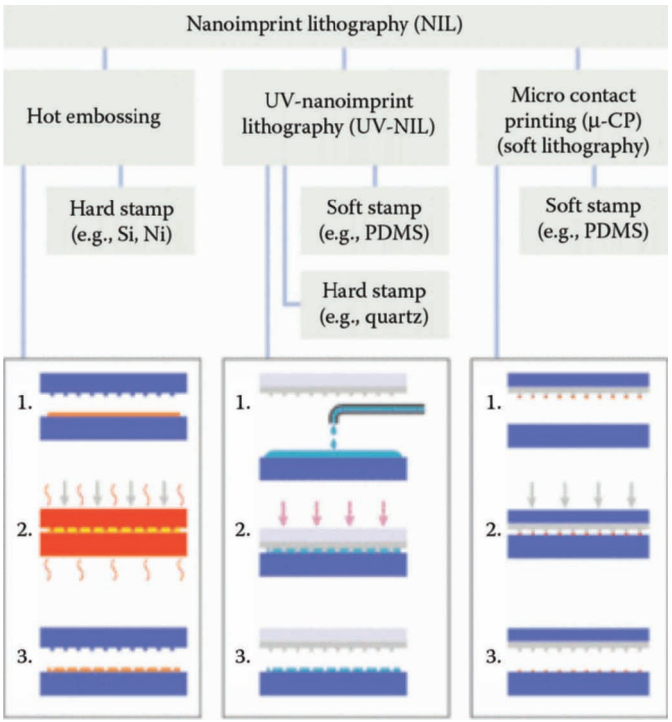


FIGURE 10.13 Nanoimprint lithography (NIL). Comparison of hot embossing with UV-nanoimprint lithography and microcontact printing.

Both techniques are applied to amorphous (PMMA, PC, PSU) and semicrystalline thermoplastics (POM, PA, PVDF, PFA). RIM is used for molding of thermoplastics (PMMA, PA), duroplastics (PMMA, epoxide), or elastomers (silicones).[24]

Summary: Pros and Cons of Compression Molding

Pros and cons of compression molding are summarized in Table 10.5. In Figure 10.13 we compare hot embossing with UV-nanoimprint lithography (see Chapter 2) and microcontact printing (soft lithography, Chapter 2). These techniques can all be grouped under nanoimprint lithography (NIL).

LIGA—Background

History

LIGA combines the sacrificial wax molding method, known since the time of the Egyptians, with x-ray lithography and electrodeposition. Combining electrodeposition and x-ray lithography was first carried out by Romankiw and coworkers at IBM as early as 1975.[25] These authors made high-aspect-ratio metal structures by plating gold in x-ray-defined resist patterns of up to 20 μm thick. They had, in other words, already invented "LIG"; that is, LIGA without the *abformung* (molding).[25] This IBM work was an extension of through-mask plating, also pioneered by Romankiw et al. in 1969, and was geared toward the fabrication of thin film magnetic recording heads (see Figure 2.4a).[26] The addition of plastic molding to the lithography and plating process was realized by Ehrfeld et al.[27] at the Kernforschungszentrum Karlsruhe (KfK) in 1982. By adding molding, these pioneers recognized the broader implications of LIGA as a new means of low-cost manufacturing of a wide variety of microparts with unprecedented accuracies from various materials previously impossible to batch fabricate.[27] In Germany, LIGA originally developed almost completely outside of the semiconductor industry. In the United States, it was the late Henry Guckel who, starting in 1988, repositioned the field in light of semiconductor process capabilities and brought it closer to standard manufacturing processes.

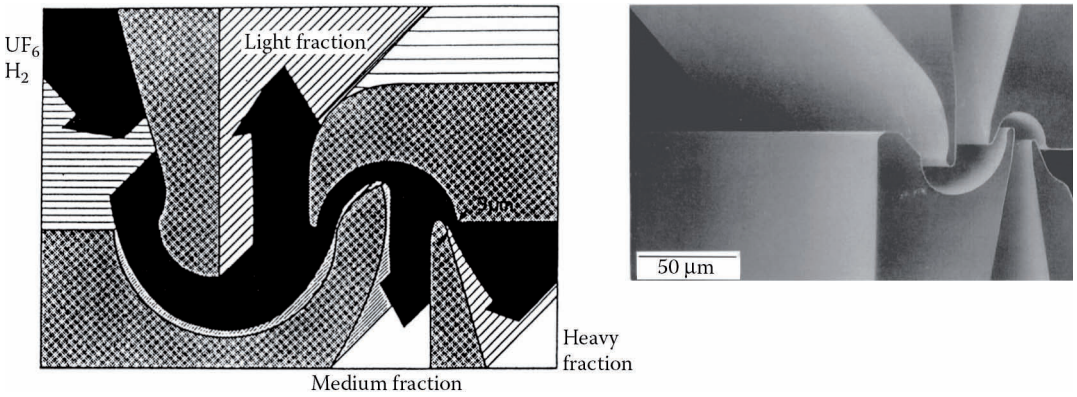


FIGURE 10.14 Schematic and scanning electron micrograph of a separation nozzle structure produced by electroforming with nickel using a micromolded PMMA template. This nozzle represents the first product ever made by LIGA. (From Hagmann, P., W. Ehrfeld, and H. Vollmer. 1987. *First meeting of the European Polymer Federation, European symposium on polymeric materials.* Lyon, France. http://opac.fzk.de:81/de/oai_frm.html?titlenr=24694&server=//127.0.0.1&port=81; Archiv *KAROLA–OA–Volltextserver_des_Forschungszentrums Karlsruhe* (Germany). With permission.[28])

The development of the LIGA process initiated at KfK was intended for the mass production of micrometer-sized nozzles for uranium-235 enrichment (see Figure 10.14).[27] The German group used synchrotron radiation from a 2.5-GeV storage ring for the exposure of the poly(methylmethacrylate) (PMMA) resist.

Today LIGA and especially LIGA-like processes are researched in many laboratories around the world, and developing the ideal means for fabricating micromolds for the large-scale production of precise micromachines remains an elusive holy grail. In LIGA, mold inserts are made via x-ray lithography, but, depending on the dimensions of the microparts, the accuracy requirements, and the fabrication costs, mold inserts may also be realized by electron beam (e-beam) writing, computer numerically controlled (CNC) machining, wet Si bulk micromachining, deep UV resists, deep reactive ion etching (DRIE), ultrasonic cutting, excimer laser ablation, electrical discharge machining (EDM), and laser cutting (see Figure 10.1).

Synchrotron Orbital Radiation

Introduction

Lithography based on synchrotron radiation, also called *synchrotron orbital radiation* (SOR), is primarily pursued with the aim of adopting the technology as an industrial tool for the large-scale manufacture of microelectronic circuits with characteristic dimensions in the submicrometer range (see Chapter 2).[29,30] Synchrotron radiation sources "outshine" electron impact and plasmas sources for generating x-rays. They emit a much higher flux of usable collimated x-rays, thereby allowing shorter exposure times and larger throughputs. Pros and cons of x-ray radiation for lithography in IC manufacture are summarized in Table 10.6.

Despite the many promising features of x-ray lithography, the technique today lacks mainstream acceptance in the IC industry, and continued improvements in optical lithography outpace the industrial use of x-ray lithography for IC applications. However, its use for prototype development on small product runs will no doubt continue its course. In 1991, experts projected that x-ray lithography would be in use by 1995 for 64-Mb dynamic random

TABLE 10.6 Pros and Cons of SOR X-Ray Lithography for IC Manufacture

Pros	Cons
Lithography process insensitive to resist thickness, exposure time, development time (large DOF)	Resist not very sensitive (not too important because of the intense light source)
Absence of backscattering results in insensitivity to substrate type, reflectivity and topography, pattern geometry and proximity, dust and contamination	Masks very difficult and expensive to make
High resolution, <0.2 µm	Very high start-up investment
Some have suggested high throughput	Not proven as a system Radiation effects on SiO_2

access memory (DRAM) manufacture with critical dimensions around 0.3–0.4 μm. With more certainty, they projected that the transition to x-rays would occur with the 0.2–0.3-μm critical dimensions of 256-Mb DRAMs, which came online by 1998.[29] Both dates obviously passed without materialization of the wide-scale industrial use of x-rays.

In addition to being an option for the next-generation IC lithography as discussed in more detail in Chapter 2, x-rays are also used in the fabrication of 3D microstructures. In LIGA, synchrotron radiation is used in the lithography step only. However, other micromachining applications for SOR exist. Urisu and his colleagues, for example, explored the use of synchrotron radiation for radiation-excited chemical vapor deposition and etching.[31] Micromachinists are hoping to piggyback x-ray lithography research and development efforts for the fabrication of micromachines onto major IC projects. Today the use of x-ray lithography for fabricating microdevices, other than ICs, does not present a large business opportunity by itself. Not having a major IC product line associated with x-ray lithography makes it extra hard to justify the use of x-ray lithography for micromachining, especially because other, less-expensive micromachining technologies are only very recently starting to open up the type of mass markets one is used to in the IC world. The fact that the x-rays used in LIGA are shorter wavelength than in the IC application (2–10 Å vs. 20–50 Å) also puts micromachinists at a disadvantage. For example, soft x-rays in the IC industry may eventually be generated from a much less-expensive source, such as a transition radiation source.[32] Also, nontraditional IC materials are frequently used in LIGA. The fabrication of x-ray masks poses more difficulties than masks for IC applications. Rotation and slanting of the x-ray masks to craft nonvertical walls further differentiate LIGA exposure stations. All these factors make exploring LIGA a challenge. However, given sufficient research and development money, large markets are likely to emerge in the next 10 years. These markets could be in the manufacture of devices with stringent requirements imposed on resolution, aspect ratio, structural height, and parallelism of structural walls. Optical applications for the information technology field seem particularly attractive early product targets.

Thus far, it is the research community that has primarily benefited from the availability of SOR photon sources. With its continuously tunable radiation across a very wide photon range, highly polarized, and directed into a narrow beam, SOR provides a powerful probe of atomic and molecular resonances. Other types of photon sources prove unsatisfactory for these applications in terms of intensity or energy spread. As can be concluded from Table 10.7, applications of SOR beyond lithography range from structural and chemical analysis to microscopy, angiography, and even to the preparation of new materials.

Table 10.7 SOR Applications

Application Area	Instruments / Technologies Needed
Structural analysis	
Atoms Molecules Very large molecules Proteins Cells Crystals Polycrystals	Photoelectron spectrometers Absorption spectrometers Fluorescent spectrometers Diffraction cameras Scanning electron microscope (to view topographical radiographs) Time-resolved x-ray diffractometers
Chemical analysis	
Trace Surface Bulk	Photoelectron spectrometers (Secondary ion) mass spectrometer Absorption/fluorescence spectrometers Vacuum systems
Microscopy	
Photoelectron X-ray	Photoemission microscopes X-ray microscopes SEM (for viewing) Vacuum systems
Micro-/nanofabrication	
X-ray lithography Photochemical deposition of thin films Etching	Steppers, mask making Vacuum systems LIGA process
Medical diagnostics	
Radiography Angiography and tomography	X-ray cameras and equipment Computer-aided display
Photochemical reactions	
Preparation of novel materials	Vacuum systems Gas-handling equipment

Source: Muray, J. J., and I. Brodie. 1991. *Report no. 2019.* Menlo Park, CA: SRI International.[33]

Synchrotron Radiation: Technical Aspects

Some important concepts associated with synchrotron radiation (such as the bending radius of the synchrotron magnet, magnetic field strength, beam current, critical wavelength, and total radiated power) require introduction. Figure 10.15 presents a schematic of an x-ray exposure station. Electrons are injected into the ring, where they are maintained at energies anywhere from 10^6–10^9 eV. The cone of radiation shown in this figure is the electromagnetic radiation emitted by the circling electrons as a result of the radial acceleration that keeps them in the orbit of the electron synchrotron or storage ring. For high-energy particle studies, this radiation, emitted tangentially to the circular electron path (*Bremsstrahlung*[*]), limits the maximum energy the electrons can attain. Bremsstrahlung is a nuisance for studies of the composition of the atomic nucleus in which high-energy particles are smashed into the nucleus. To minimize the Bremsstrahlung, physicists desire ever bigger synchrotrons.[†] The energy lost in the emission process is made up in a radiofrequency (RF) cavity, where electrons are accelerated back up to the storage ring energy. Injection of electrons must be repeated a few times a day because the electron current slowly decays as a result of leakage. For x-ray lithography applications, electrical engineers want to maximize the x-ray emission and build small synchrotrons instead (e.g., the radius of curvature for a compact, superconducting synchrotron is 2 m). The operating cost of these magnets is primarily that of the liquid helium refrigeration. Once high-T_c (critical temperature, the temperature at which the resistance falls to zero) superconducting materials can be made in bulk, compact storage rings will become very attractive. The angular opening of the radiation cone in Figure 10.15 is determined by the electron energy, E, and is given by:

$$\theta \approx \frac{mc^2}{E} = \frac{0.5}{E(\text{GeV})}(\text{mrad}) \qquad (10.1)$$

The x-ray light bundle with the cone opening, θ, describes a horizontal line on an intersecting substrate because the x-ray bundle is tangent to the circular electron path. In the vertical direction, the intensity of the beam exhibits a Gaussian distribution, and the vertical exposed height on the intersecting substrate can be calculated knowing θ and R, the distance from the radiation point, P, to the substrate. With E = 1 GeV, θ = 1 mrad, and R = 10 m, the exposed area in the vertical direction measures about 0.5 cm. To expose a substrate homogeneously over a wider vertical range, the sample is moved vertically through the irradiation band with a precision scanner, for example, at a speed of 10 mm/s over a

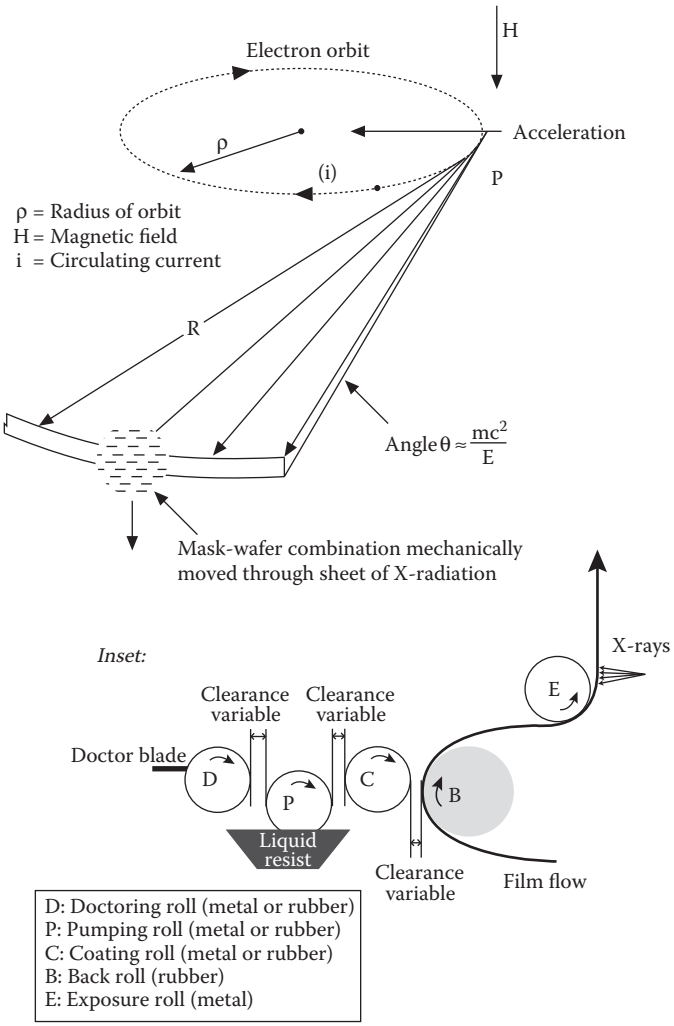


FIGURE 10.15 Schematic of an x-ray exposure station with a synchrotron radiation source. The x-ray radiation cone (opening θ) is tangential to the electron's path, describing a line on an intersecting substrate. Inset, a vision for the future of x-ray lithography: continuous micromanufacturing.

[*] German for "braking radiation."

[†] The Superconducting Super Collider Laboratory (SSCL or SSC for short), which was never finished, was to be an oval-shaped accelerator, the circumference of which would have been 54 miles. The collider oval was to be located about 200 ft. underground and would have surrounded the city of Waxahachie, Texas.

100-mm scanning distance. Usually, the substrate is stepped up and down repeatedly until the desired x-ray dose is obtained. It is interesting to note that an SOR setup affords continuous lithography, a prospect, this author believes, that will make disposable ICs and MEMS and NEMS a possibility. Rolls of dry x-ray photoresist or x-ray photoresist covered metal foils could pass through the exposure beam, resulting in a continuous lithography process, that is, a "beyond batch" type of approach (see Figure 10.15, inset).

The electron energy E in Equation 10.1 is given by:

$$E(GeV) = 0.29979\ B(Tesla)\ \rho\ (meters) \quad (10.2)$$

with B the magnetic field and ρ the radius of the circular path of the electrons in the synchrotron.

The total radiated power can be calculated from the energy loss of the electrons per turn and is given by:

$$P(kW) = \frac{88.47E^4 i}{\rho} \quad (10.3)$$

with i the beam current.

The emission of the synchrotron electrons is a broad spectrum without characteristic peaks or line enhancements, and its distribution extends from the microwave region through the infrared, visible, ultraviolet (UV), and into the x-ray region. The critical wavelength, λ_c, is defined so that the total radiated power at lower wavelengths equals the radiated power at higher wavelengths, and is given by:

$$\lambda_c = \frac{5.59\rho(m)}{E^3(GeV)} \quad (10.4)$$

Equation 10.3 shows that the total radiated power increases with the fourth power of the electron energy. From Equation 10.4, we appreciate that the spectrum shifts toward shorter wavelengths with the third power of the electron energy.

The dose variation absorbed in the top versus the bottom of an x-ray resist should be kept small so that the top layer does not deteriorate before the bottom layers are sufficiently exposed. Because the depth of penetration increases with decreasing wavelength, synchrotron radiation of very short wavelengths is needed to pattern thick resist layers. To obtain good aspect ratios in LIGA structures, the critical wavelength ideally should be 2 Å. Bley et al.,[34] at KfK, designed a new synchrotron optimized for LIGA. They proposed a magnetic flux density, B, of 1.6285 T; a nominal energy, E, of 2.3923 GeV; and a bending radius, ρ, of 4.9 m. With those parameters, Equation 10.4 results in the desired λ_c of 2 Å and an opening angle of radiation, based on Equation 10.1, of 0.2 mrad (in practice this angle will be closer to 0.3 mrad because of e-beam emittance).

The x-rays traveling from the ring to the sample site are held in a high vacuum. The sample itself is either kept in air or in a helium atmosphere. The inert atmosphere prevents corrosion of the exposure chamber, mask, and sample by reactive oxygen species, and removal of heat is much faster than in air (the heat conductivity of He is high compared with air). In He, the x-ray intensity loss is also 500 times less than in air. A Be window separates the high vacuum from the inert atmosphere. For wavelengths shorter than 1 nm, Be is very transparent—that is, an excellent x-ray window. A 25-µm-thick Be window can withstand a 1-atm pressure differential across a small diameter (<1 in.). For large area exposures, windows up to 6-cm diameter have been developed. Be windows age with x-ray exposure and must be replaced periodically.

Access to the Technology

Today, the construction cost for a typical synchrotron totals more than $30 million, restricting the access to LIGA. Obviously, a less-expensive alternative for generating intense x-rays is preferred. Along this line, in Japan, Ishikawajima-Harima Heavy Industries (IHI) is building compact synchrotron x-ray sources (e.g., an 800-MeV synchrotron of about 30 ft./side) (http://www.ihi.co.jp).

By the end of 1993, eight nonprivately owned synchrotrons were in use in the United States. The first privately owned synchrotron was put into service in 1991 at IBM's Advanced Semiconductor Technology Center (East Fishkill, NY). Table 10.8 lists the eight US synchrotron facilities.

Most of the facilities listed in Table 10.8 allow LIGA work to continue. For example, Cronos Integrated Microsystems, Inc., a JDS Uniphase Company and a spinoff from MCNC (Research Triangle Park, NC),

Table 10.8 Access to Synchrotron Radiation Is or Will Soon Be Available at the Following Facilities in the United States

Facility	Institute	URL
Advanced Photon Source (APS)	Argonne National Laboratory	http://www.aps.anl.gov
Cornell High Energy Synchrotron Source (CHESS)	Cornell University	http://www.tn.cornell.edu
National Synchrotron Light Source (NSLS)	Brookhaven National Laboratory	http://www.nsls.bnl.gov
Stanford Synchrotron Radiation Laboratory (SSRL)	Stanford University	http://www-ssrl.slac.stanford.edu
Synchrotron UV Radiation Facility (SURF)	National Institute of Standards and Technology	http://physics.nist.gov/MajResFac/SURF
Synchrotron Radiation Center (SRC)	University of Wisconsin, Madison	http://www.src.wisc.edu
Center for Advanced Microstructures and Devices (CAMD)	Louisiana State University	http://www.camd.lsu.edu
Advanced Light Source (ALS)	Lawrence Berkeley Laboratory	http://www-als.lbl.gov

in collaboration with the University of Wisconsin, Madison, announced its first multiuser LIGA process sponsored by Advanced Research Projects Agency in September 1993 (http://www.memsrus.com). CAMD, at Louisiana State University, has three beam lines exclusively dedicated to micromachining work, and the Advanced Light Source at Berkeley has one beam line available for micromachining.

Like Cronos, Forschungszentrum Karlsruhe GmbH offers a multiuser LIGA service (LEMA, or LIGA experiment for multiple applications). The commercial exploitation of LIGA is pursued by at least three German organizations: microParts GmbH STEAG (http://www.microparts.de), IMM (http://www.imm.uni-mainz.de), and Forschungszentrum Karlsruhe, or KfK (http://www.fzk.de). In the United States, Louisiana State University's CAMD (http://www.camd.lsu.edu; Baton Rouge, LA) with its spinoff Mezzo Systems, Inc., founded in 2000, (http://www.mezzosystems.com) is leading the way.

LIGA and LIGA-Like Process Steps

X-Ray Masks

Introduction

With reference to Figure 10.16 we now detail each of the steps in the LIGA process.

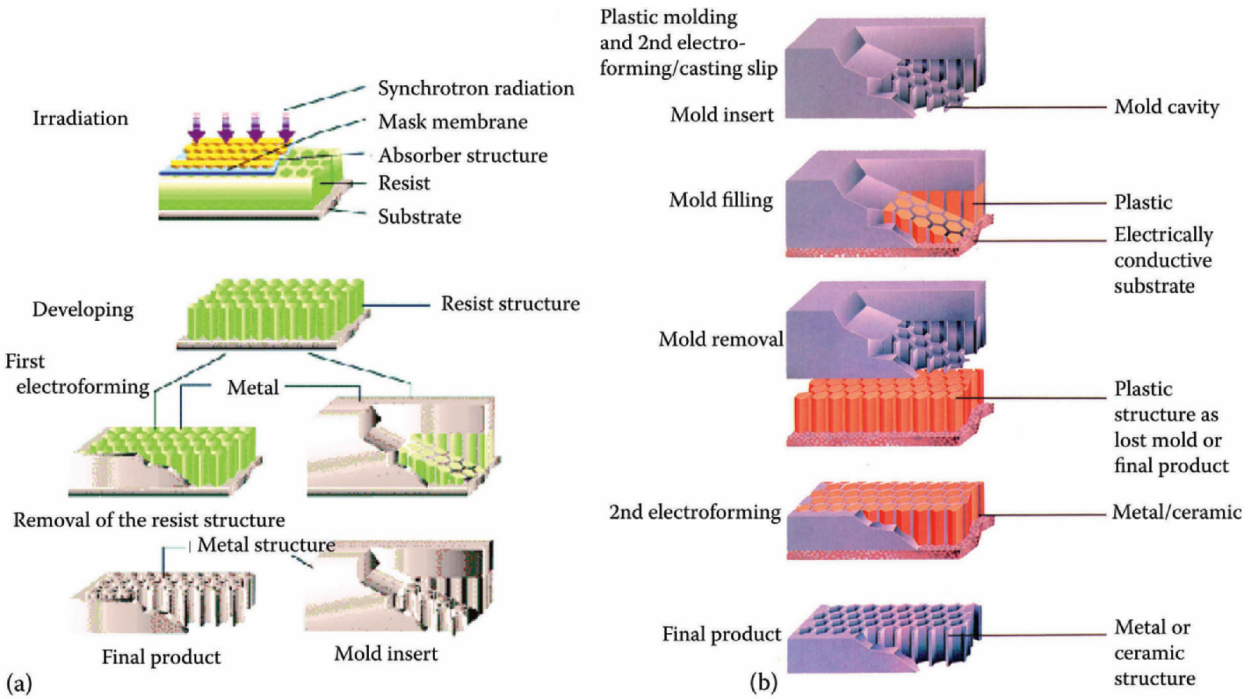


FIGURE 10.16 (a) Basic LIGA process step: x-ray deep-etch lithography and first electroforming. (b) Plastic molding and second electroforming/slip casting.

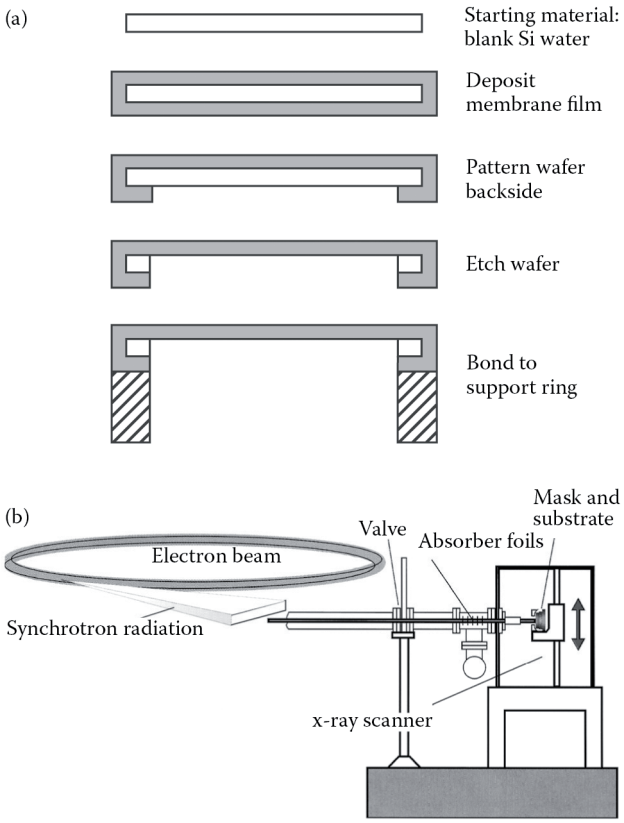


FIGURE 10.17 Schematic of a typical x-ray mask (a) and mask and substrate assembly in an x-ray scanner (b). [(b) from IMM, http://www.imm-mainz.de/. 1995. Commercial brochure. With permission.[2]]

X-ray mask production is one of the most difficult aspects of x-ray lithography. To be highly transmissive to x-rays, the mask substrate by necessity must be a low Z (atomic number) thin membrane. X-ray masks should withstand many exposures without distortion, be alignable with respect to the sample, and be rugged. A possible x-ray mask architecture and its assembly with a substrate in an x-ray scanner are shown in Figure 10.17. The mask shown here has three major components: an absorber, a membrane or mask blank, and a frame. The absorber contains the information to be imaged onto the resist. It is made up of a material with a high atomic number (Z), often Au, patterned onto a membrane material with a low Z. The high-Z material absorbs x-rays, whereas the low-Z material transmits x-rays. The frame lends robustness to the membrane/absorber assembly so that the whole can be handled confidently.

The requirements for x-ray masks in LIGA differ substantially from those for the IC industry. A comparison is presented in Table 10.9.[35] The main difference lies in the absorber thickness. To achieve a high contrast (>200), very thick absorbers (>10 μm vs. 1 μm) and highly transparent mask blanks (transparency >80%) must be used because of the low x-ray resist sensitivity and the great depth of the resist. Another difference focuses on the radiation stability of membrane and absorber. For conventional optical lithography, the supporting substrate is a relatively thick, optically flat piece of glass or quartz highly transparent to optical wavelengths. It provides a highly stable (>10^6 μm) basis for the thin (0.1 μm) chrome absorber pattern. In contrast, the x-ray mask consists of a very thin membrane (2–4 μm) of low-Z material carrying a high-Z thick absorber pattern.[36] A single exposure in LIGA results in an exposure dose 100 times higher than in the IC case.

We will look into these different mask aspects separately before detailing a process with the potential of obviating the need for a separate x-ray mask altogether, that is, through the use of conformal or transfer masks.

TABLE 10.9 Comparison of Masks in LIGA and the IC Industry

	Semiconductor Lithography	LIGA Process
Transparency	≥50%	≥80%
Absorber thickness	±1 μm	±10 μm
Field size	50 × 50 mm^2	100 × 100 mm^2
Radiation resistance	= 1	= 100
Surface roughness	<0.1 μm	<0.5 μm
Waviness	<±1 μm	<±1 μm
Dimensional stability	<0.05 μm	<0.1–0.3 μm
Residual membrane stress	~10^8 Pa	~10^8 Pa

Source: Ehrfeld, W., W. Glashauer, D. Munchmeyer, and W. Schelb. 1986. Mask making for synchrotron radiation lithography. *Microelectron Eng* 5:463–70.[35]

X-Ray Membrane (Mask Blank)

The low-Z membrane material in an x-ray mask must have a transparency for rays with a critical wavelength, λ_c, from 0.2–0.6 nm of at least 80% and should not induce scattering of those rays. To avoid pattern distortion, the residual stress, σ_r, in the membrane should be less than 10^8 dyn/cm^2. Mechanical stress in the absorber pattern can cause in-plane distortion of the supporting thin membrane, necessitating a high Young's modulus for the membrane material. Humidity or high deposited doses of x-ray might also distort the membrane directly. During one typical lithography step, the mask may be exposed to 1 MJ/cm^2 of x-rays. Because most membranes must be very thin for optimum transparency, a compromise has to be found between transparency, strength, and form stability. Important x-ray membrane materials are listed in Table 10.10. The higher radiation dose in LIGA prevents the use of boron nitride (BN) and compound mask blanks, which incorporate a polyimide layer. Those mask blanks are perfectly appropriate for classical IC lithography work but will not do for LIGA work. Mask blanks of metals such as Ti and Be were specifically developed for LIGA applications because of their radiation hardness.[35,37] In comparing titanium and beryllium membranes, beryllium can have a much greater membrane thickness, d, and still be adequately transparent. For example, a membrane transparency of 80%, essential for adequate exposure of a 500-µm-thick PMMA resist layer, is obtained with a thin 2-µm titanium film, whereas with beryllium, a thick 300-µm membrane achieves the same result. The thicker beryllium membrane permits easier processing and handling. In addition, beryllium has a greater Young's modulus E than titanium (330 vs. 140 kN/mm^2), and because it is the product of $E \times d$ that determines the amount of mask distortion, distortions resulting from absorber stress should be much smaller for beryllium blanks.[37,38] Thus, beryllium comes forward as an excellent membrane material for LIGA because of its high transparency and excellent damage resistance. Such a mask should be good for up to 10,000 exposures and may cost $20,000–30,000 ($10,000–15,000 in quantity). Stoichiometric silicon nitride (Si_3N_4) used in x-ray mask membranes may contain numerous oxygen impurities, absorbing x-rays and thus producing

Table 10.10 Comparison of Membrane Materials for X-Ray Masks

Material	X-Ray Transparency	Nontoxicity	Dimensional Stability	Remark
Si	0 (50% transmission at 5.5-µm thickness)	++	0 (thermal exp coefficient 2.6°C^{-1} 10^{-6}) Young's modulus = 1.3	Single-crystal Si, well developed, rad hard, stacking faults cause scattering, material is brittle
SiN_x	(50% transmission at 2.3-µm thickness)	++	(thermal exp coefficient 2.7°C^{-1} 10^{-6}) Young's modulus = 3.36	Amorphous, well developed, rad hard if free of oxygen, resistant to breakage
SiC	(50% transmission at 3.6-µm thickness)	++	(thermal exp coefficient 4.7°C^{-1} 10^{-6}) Young's modulus = 3.8	Poly and amorphous, rad hard, some resistance to breakage
Diamond	+(50% transmission at 4.6-µm thickness)	++	++(thermal exp coefficient 1.0°C^{-1} 10^{-6}) Young's modulus = 11.2	Poly, research only, highest stiffness
BN	(50% transmission at 3.8-µm thickness)		(thermal exp coefficient 1.0°C^{-1} 10^{-6}) Young's modulus = 1.8	Not rad hard, i.e., not applicable for LIGA
Be	++	—	++	Research, especially suited for LIGA, even at 100 µm the transparency is good, 30 µm typical, difficult to electroplate, toxic material
Ti	—		0	Research, used for LIGA, not very transparent, films must not be more than 2–3 µm thick

heat. This heat often suffices to prevent the use of nitride as a good LIGA mask. Single-crystal silicon masks have been made (1 cm^2 and 0.4 µm thick and 10 cm^2 and 2.5 µm thick) by electrochemical etching techniques. Nanostructures, Inc. (http://www.nanostructures.com) is one of the companies that make such thin Si masks. For Si and Si_3N_4, the Young's modulus is low compared with CVD-grown diamond, and for SiC films, the Young's modulus is as much as three times higher. These higher stiffness materials are more desirable because the internal stresses of the absorbers, which can distort mask patterns, are less of an issue. Unfortunately, diamond and SiC membranes are also the most difficult to produce.

Absorber Materials

The requirements on the absorber are high attenuation (>10 db), stability under radiation for extended periods, negligible distortion (stress <10^8 dyn/cm^2), ease of patterning, repairability, and low defect density. Typical absorber materials are listed in Table 10.11. Gold is used most commonly, and some research groups are looking at the viability of tungsten and other materials. In the IC industry, an absorber thickness of 0.5 µm might suffice, whereas LIGA deals with thicker layers of resist requiring a thicker absorber to maintain the same resolution.

Figure 10.18 illustrates how x-rays, with a characteristic wavelength of 0.55 nm, are absorbed along their trajectory through a Kapton preabsorber filter, an x-ray mask, and resist.[39] As we saw in Volume I,

TABLE 10.11 Comparison of Absorber Materials for X-Ray Masks

Material	Remark
Gold	Not the best stability (grain growth), low stress, electroplating only, defects repairable (thermal exp coefficient 14.2°C^{-1} 10^{-6}) (0.7 µm for 10 dB)
Tungsten	Refractory and stable, special care is needed for stress control, dry etchable, repairable (thermal exp coefficient 4.5°C^{-1} 10^{-6}) (0.8 µm for 10 dB)
Tantalum	Refractory and stable, special care is needed for stress control, dry etchable, repairable
Alloys	Easier stress control, greater thickness needed to obtain 10 dB

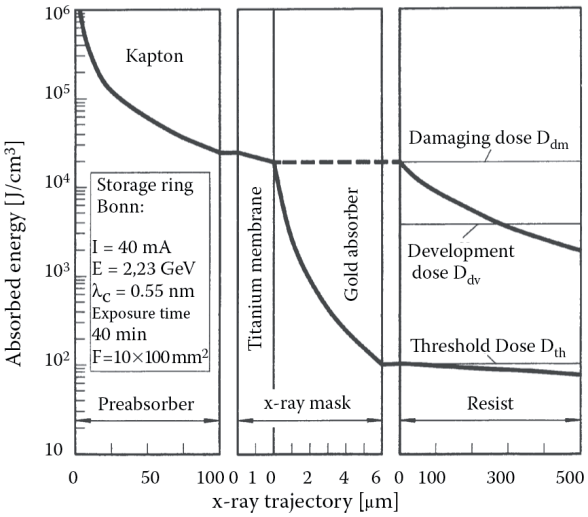


FIGURE 10.18 Absorbed energy along the x-ray trajectory including a 500-µm-thick PMMA specimen, x-ray mask, and a Kapton preabsorber. (From Bley, P., W. Menz, W. Bacher, K. Feit, M. Harmening, H. Hein, J. Mohr, W. K. Schomburg, and W. Stark. 1991. Proc. 1991 Int. MicroProcess Conf. *J. Appl. Phys.* JJAP Series 5, p. 384, Japan, Tokyo. With permission.[39])

Chapter 5, photon absorption by a solid in the visible region of the spectrum depends on the atomic arrangement of the atoms and their bonding (for example, pure silicon is strongly absorbing, but silicon combined with oxygen is transparent). For the energetic photons in the x-ray regime, photon absorption is much easier to predict and is independent of the details of atomic arrangement. It depends primarily on the electron concentration per unit volume. Because the concentration of atoms per unit volume only differs by factors of two or three from each other, the electron concentration in two materials can be estimated from the atomic number, Z. Lead ($Z = 82$) absorbs x-rays much more efficiently than aluminum ($Z = 13$) and consequently is used in shielding around x-ray apparatus. X-ray absorption depends on the energy of the x-rays and decreases with increasing x-ray energy, E. Absorption decreases nearly proportionally to the cube of the energy (i.e., absorption proportional to $1/E^3$; Equation 10.4). Thus, the low energy portion of the synchrotron radiation is absorbed mainly in the top portion of the resist layer because absorption increases with increasing wavelength. The Kapton preabsorber filters out much of the low energy radiation to prevent overexposure of the top surface of the resist. The x-ray dose at which the resist gets damaged, D_{dm},

the dose required for development of the resist, D_{dv}, and the "threshold dose" at which the resist starts dissolving in a developer, D_{th}, are all indicated in Figure 10.18. In the areas under the absorber pattern of the x-ray mask, the absorbed dose must stay below the threshold dose, D_{th}. Otherwise, the structures partly dissolve, resulting in poor feature definition. From Figure 10.18, we can deduce that the height of the gold absorbers must exceed 6 µm to reduce the absorbed radiation dose of the resist under the gold pattern to below the threshold dose, D_{th}. In Figure 10.19, the necessary thickness of the gold absorber patterns of an x-ray mask is plotted as a function of the thickness of the resist to be patterned; the Au must be thicker for thicker resist layers and for shorter characteristic wavelengths, λ_c, of the x-ray radiation. To pattern a 500-mm-high structure with a λ_c of 0.225 nm, the gold absorber must be more than 11 µm high.

Exposure of yet more extreme photoresist thicknesses is possible if proper x-ray photon energies are used. At 3,000 eV, the absorption length in PMMA roughly measures 100 µm, which enables the above-mentioned 500-µm exposure depth.[40] Using 20,000-eV photons results in absorption lengths of 1 cm. PMMA structures up to 10 cm (!) thick have been exposed this way.[41] A high-energy mask used by Guckel for these high-energy exposures has an Au absorber 50 µm thick and a blank membrane of 400-µm-thick Si. Guckel obtained an absorption contrast of 400 when exposing a 1000-µm-thick PMMA sheet with this mask. An advantage of using such thick Si blank membranes is that larger resist areas can be exposed because one does not depend on a fragile membrane/absorber combination.[42]

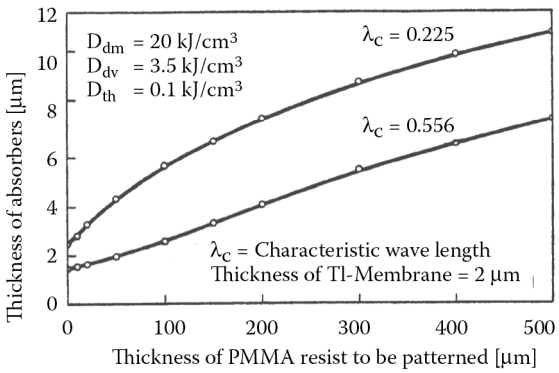


FIGURE 10.19 Necessary thickness of the gold absorbers of an x-ray mask. (From Bley, P., W. Menz, W. Bacher, K. Feit, M. Harmening, H. Hein, J. Mohr, W. K. Schomburg, and W. Stark. 1991. Proc. 1991 Int. MicroProcess Conf. *J. Appl. Phys*. JJAP Series 5, p. 384, Japan, Tokyo. With permission.[39])

Absorber Fabrication

Single-Layer Absorbers To make a mask with gold absorber structures of a height more than 10 µm, one must first succeed in structuring a resist of that thickness. The height of the resist should be a bit higher (e.g., 20%) than the absorber itself to accommodate the electrodeposited metal fully between the resist features. Currently, no means to structure a resist of that height with sufficient accuracy and perfect verticality of the walls exist unless x-rays are used. Different procedures for producing x-ray masks with thicker absorber layers using a two-stage lithography process have been developed.

The KfK solution calls for first making an intermediate mask with photo or e-beam lithography. This intermediate mask starts with a 3-µm-thick resist layer, in which case the needed line-width accuracy and photoresist wall steepness of printed features are achievable. After gold plating between the resist features and stripping the resist, this intermediate mask is used to write a pattern with x-rays in a thicker resist, e.g., 20 µm thick. After electrodepositing and resist stripping, the x-ray mask, that is, the master mask, is obtained.

Because hardly any accuracy is lost in the copying of the intermediate mask with x-rays to obtain the master mask, it is the intermediate mask quality that determines the ultimate quality of the LIGA-produced microstructures. The structuring of the resist in the intermediate mask is handled with optical techniques when the requirements on the LIGA structures are less stringent. The minimal lateral dimensions for optical lithography in a 3-µm-thick resist typically measure about 2.5 µm. Under optimum conditions, a wall angle of 88° is achievable. With e-beam lithography, a minimum lateral dimension of less than 1 µm is feasible. The most accurate pattern transfer is achieved through reactive ion etching of a trilevel resist system. In this approach, a 3–4-µm-thick polyimide resist is first coated onto the titanium or beryllium membrane, followed by a coat of 10–15-nm titanium deposited

with magnetron sputtering. The thin layer of titanium is an excellent etch mask for the polyimide; in optimized oxygen plasma, the titanium etches 300 times slower than the polyimide. To structure the thin titanium layer itself, a 0.1-µm-thick optical resist is used. Because this top resist layer is so thin, excellent lateral tolerances result. The thin Ti layer is patterned with optical photolithography and etched in argon plasma. After etching the thin titanium layer, exposing the polyimide locally, oxygen plasma helps to structure the polyimide down to the titanium or beryllium membrane. Lateral dimensions of 0.3 µm can be obtained in this fashion. Patterning the top resist layer with an e-beam increases the accuracy of the three-level resist method even further. Electrodeposition of gold on the titanium or beryllium membrane and stripping of the resist finish the process of making the intermediate LIGA mask. To make a master mask, this intermediate mask is printed by x-ray radiation onto a PMMA resist-coated master mask. The PMMA thickness corresponds to a bit more than the desired absorber thickness. Because the resist layer thickness is in the 10–20-µm range, a synchrotron x-ray wavelength of 10 Å is adequate for making the master mask. A further improvement in LIGA mask making is to fabricate intermediate and master mask on the same substrate, greatly reducing the risk for deviations in dimensions caused, for example, by temperature variations during printing.[43] The ultimate achievement would be to create a one-step process to make the master mask. Along this line, Hein et al.[38] investigated the direct patterning of 10-µm-high resist layers with a 100-kV e-beam.

Stepped Absorber In principle, stepped absorber structures may result in stepped LIGA structures by means of x-ray lithography with a single mask. In this manner, variable dose depositions can be achieved at the same resist heights. The variable dose results in different molecular weights and hence in a different developing behavior. This technique unfortunately leads to rounded features and poor step-height control. Further below under "Stepped and Slanted Microstructures" we will learn how to make better resolved stepped features.

CNC Machined Absorbers A rather unexpected approach to pattern absorber layers for x-ray masking is pursued by Friedrich et al. from the Michigan Technological University (http://www.me.mtu.edu/~microweb). Friedrich et al. are exploring x-ray mask fabrication by traditional machining methods, such as micromilling, micro-EDM, and lasers.[44] Using micromilling, this group succeeded in making mask features up to 62 µm deep and walls down to 4 µm thick and 10 µm high. The milling is carried out with a 22-µm end mill, fabricated itself by using a 20-keV gallium ion beam. The advantages of this approach are rapid turnaround (less than 1 day/mask), low cost, and flexibility (almost any type of material can be machined) because no intermediate steps interfere. Disadvantages are less dimensional edge acuity and nonsharp interior corners, as well as much lower absolute tolerance.

Alignment of X-Ray Mask to Substrate

The mask and resist-coated substrate must be properly registered to each other before they are put in an x-ray scanner. Alignment of an x-ray mask to the substrate is problematic because no visible light can pass through most x-ray membranes. To solve this problem, Schomburg et al.[37] etched windows in their Ti x-ray membrane. Diamond membranes have a potential advantage here because they are optically transparent and enable easy alignment for multiple irradiations without a need for etched holes.

Figure 10.20 illustrates an alternative, x-ray alignment system involving capacitive pickup between conductive metal fingers on the mask and ridges on a small substrate area; Si, in this case [US Patent 4,607,213 (1986) and 4,654,581 (1987)]. When using

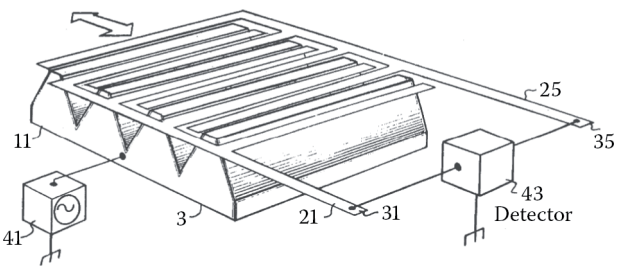


FIGURE 10.20 Mask alignment system in x-ray lithography. Conductive fingers on the mask and ridges on the Si are used for alignment. [From US Patents 4,654,581 (1987) and 4,607,213 (1986).]

multiple groups of ridges and fingers, two-axis lateral and rotational alignment become possible.

Another alternative may involve liquid nitrogen-cooled Si (Li) x-ray diodes as alignment detectors, eliminating the need for observation with visible light.[45]

Conformal, Transfer, or Self-Aligned Mask for High-Aspect-Ratio Microlithography

Vladimirsky et al.[46,47] developed a procedure to eliminate the need for an x-ray mask membrane. Unlike conventional masks, the so-called *x-ray transfer mask* does not treat a mask as an independent unit. The technique is based on forming an absorber pattern directly on the resist surface forming a conformal, self-aligned, or transfer mask. An example process is shown in Figure 10.21. In this sequence, a transfer mask plating base is first prepared on the PMMA substrate plate by evaporating 70 Å of chromium (as adhesion layer) followed by 500 Å of gold using an e-beam evaporator. A 3-µm-thick layer of standard Novolak-based AZ-type resist S1400-37 (Shipley Co., Marlborough, MA) is then applied over the plating base and exposed in contact mode through an optical mask using an UV exposure station. Three micrometers of electroplated gold on the exposed plating base further completes the transfer mask. A blanket exposure and subsequent development remove the remaining resist. The 500 Å of gold plating base is dissolved by a dip of 20–30 s in a solution of KI (5%) and I (1.25%) in water; the Cr adhesion layer is removed by a standard chromium etch (from KTI, Chemicals Inc., Sunnyvale, CA). Thus, fabrication of the transfer mask can be performed using standard lithography equipment available at almost any lithography shop. Depending on the resolution required, the x-ray transfer mask can be fabricated using known photon, e-beam, or x-ray lithography techniques. The patterning of the PMMA resist with a self-aligned mask is accomplished in multiple steps of exposure and development. An example of a cylindrical resonator made this way is shown in Figure 10.22. Each exposure/development step involves an exposure dose of about 8–12 J/cm^2. Subsequent 5-min development steps remove ~30 µm of PMMA. In seven steps, a self-supporting, 1.5-mm-thick PMMA resist is patterned to a depth

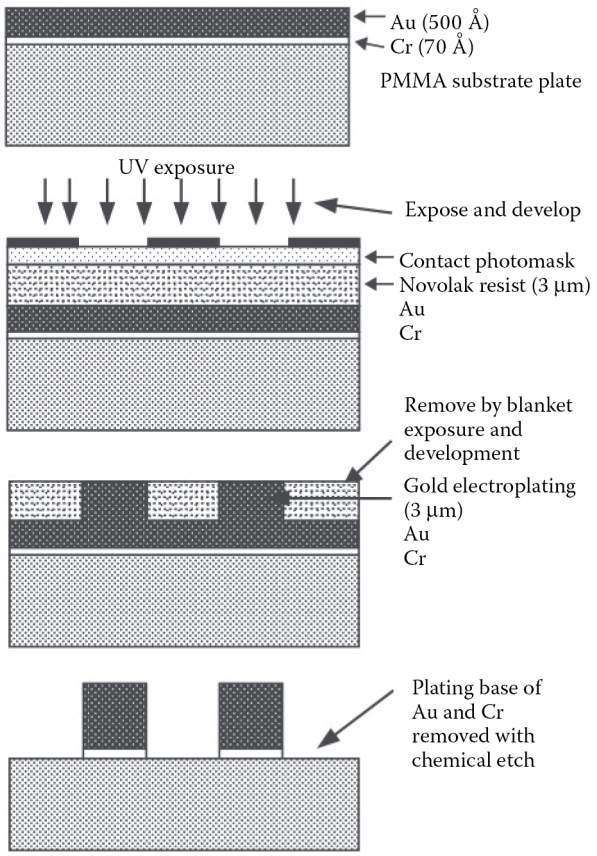


FIGURE 10.21 Sample transfer mask formation. [From Vladimirsky, Y., O. Vladimirsky, V. Saile, K. Morris, and J. M. Klopf. 1995. *Microlithography '95 (Proceedings of the SPIE).* Santa Clara, CA: SPIE.[46]]

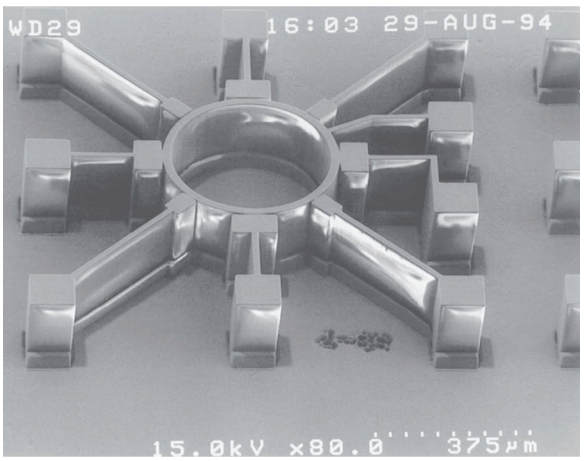


FIGURE 10.22 SEM micrograph of a cylindrical PMMA resonator made by the transfer mask method and multiple exposure/development steps. [From Vladimirsky, Y., O. Vladimirsky, V. Saile, K. Morris, and J. M. Klopf. 1995. *Microlithography '95 (Proceedings of the SPIE).* Santa Clara, CA: SPIE. With permission.[46]] (Courtesy of Dr. V. Saile.)

of more than 200 µm. The resist pattern shown in Figure 10.22 is 230 µm thick and exhibits a 2-µm gap between the inner cylinder and the pickup electrodes (aspect ratio is 100:1). The resonator pattern was produced using soft x-rays (=10 Å) and a 3-µm-thick Au absorber only.

Vladimirsky et al.[46] suggest that the forming of the transfer mask directly on the sample surface creates several additional new opportunities; besides in situ development, etching, and deposition, these include exposure of samples with curved surfaces and dynamic deformation of a sample surface during the exposure (hemispherical structures for lenses are possible this way). Elegant and cost-saving innovations like these could help mainstream LIGA.

Shih et al.[48] further developed and fine-tuned the conformal mask method for LIGA. In one very attractive embodiment of the technology, a PMMA layer on an Al substrate is coated with a Cu foil (17.5 µm) by cold pressing. The copper foil is further laminated with a dry photoresist foil (48-µm Hitachi H-N650). After exposure and development of the dry resist, gold or tin/lead absorber patterns are electroplated on the exposed Cu foil. After stripping the dry resist in a 3 wt% NaOH solution at around 50°C and the underlying copper foil in a $Cu(NH_3)_4Cl_2$ solution at 45°C, the conformal mask is ready for use. Using this approach, 1000-µm-high structures with an aspect ratio of 6.25 were fabricated with a double-exposure development cycle only. Using yet another transfer mask approach, this same research group made LIGA dies for spinnerets (see Example 10.1).[49] The authors summarize the advantages of the transfer mask method as follows:

- Alleviates the difficulty in fabricating fragile mask membranes
- Avoids alignment requirements during successive exposure steps
- Reduces exposure time and absorber thickness for the same exposure source
- Enhances pattern transfer fidelity because there is almost no proximity gap
- Avoids thermal deformation caused by exposure heat
- Increases photoresist development rate by stepwise elevated exposure dose

Choice of Primary Substrate

In the LIGA process, the primary substrate, or base plate, must be a conductor or an insulator coated with a conductive top layer. A conductor is required for subsequent electrodeposition. Some examples of primary substrates that have been used successfully are Al,[48] austenite steel plate, Si wafers with a thin Ti or Ag/Cr top layer,[50] and copper plated with gold, titanium, or nickel.[43] Other metal substrates and metal-plated ceramic, plastic, and glass plates have been used.[51] It is important that the plating base provide good adhesion for the resist. For that purpose, before applying the x-ray resist on copper or steel, the surface sometimes is mechanically roughened by microgrit blasting with corundum. Microgrit blasting may lead to an average roughness, R_a, of 0.5 µm, resulting in better physical anchoring of the microstructures to the substrate.[52] In the case of a polished metal base, chemical preconditioning may be used to improve adhesion of the resist microstructures. During chemical preconditioning, a titanium layer, sputter-deposited onto the polished metal base plate (e.g., a Cu plate), is oxidized for a few minutes in a solution of 0.5 M NaOH and 0.2 M H_2O_2 at 65°C. The oxide produced typically measures 30 nm thick and exhibits a microrough surface instrumental to securing resist to the base plate. The Ti adhesion layer may further be covered with a thin nickel seed layer (~150 Å) for electroless or electroplating of nickel. When using a highly polished Si surface, adhesion promoters need to be added to the resist resin (see "Resist Adhesion" further below). A substrate of special interest is a processed silicon wafer with integrated circuits (ICs). Integrating the LIGA process with IC circuitry on the same wafer will open up additional LIGA applications (see text below and Figure 10.39).

The back of electrodeposited microdevices is attached to the primary substrate but can be removed from the substrate if necessary. In the latter case, the substrate may be treated chemically or electrochemically to induce poor adhesion. Ideally, excellent adhesion exists between substrate and resist, and poor adhesion exists between the electroplated structure and plating base. Achieving these two contradictory demands is one of the main challenges in LIGA.

Thick resist plates can act as plastic substrates themselves. For example, using 20,000-eV rather than the more typical 3,000-eV radiation, Guckel et al.[40,41] exposed plates of PMMA up to 10 cm thick.

Resist Requirements

An x-ray resist ideally should have high sensitivity to x-rays, high resolution, resistance to dry and wet etching, thermal stability of greater than 140°C, and a matrix or resin absorption of less than 0.35 μm^{-1} at the wavelength of interest.[53] These requirements are only those for IC production with x-ray lithography.[54] To produce high-aspect-ratio microstructures with very tight lateral tolerances demands an additional set of requirements. The unexposed resist must be absolutely insoluble during development. This means that a high contrast (γ) is required. The resist must also exhibit very good adhesion to the substrate and be compatible with the electroforming process. The latter imposes a resist glass transition temperature (T_g) greater than the temperature of the electrolyte bath used to electrodeposit metals between the resist features remaining after development (e.g., at 60°C). To avoid mechanical damage to the microstructures induced by stress during development, the resist layers should exhibit low internal stresses.[55] If the resist structure is the end product of the fabrication process, further specifications depend on the application itself, for example, optical transparency and refractive index for optical components or large mechanical yield strength for load-bearing applications. For example, PMMA exhibits good optical properties in the visible and near-infrared range and lends itself to the making of all types of optical components.[56] Because of excellent contrast and good process stability known from e-beam lithography, PMMA is the preferred resist for deep-etch synchrotron radiation lithography. Two major concerns with PMMA as a LIGA resist are a rather low lithographic sensitivity of about 2 J/cm^2 at a wavelength λ_c of 8.34 Å and susceptibility to stress cracking. For example, even at shorter wavelengths, λ_c = 5 Å, more than 90 min of irradiation are required to structure a 500-µm-thick resist layer with an average ring storage current of 40 mA and a power consumption of 2 MW at the 2.3-GeV ELSA synchrotron (Bonn, Germany) (see Figure 10.18).[57] The internal stress arising from the combination of a polymer and a metallic substrate can cause cracking in the microstructures during development, a phenomenon PMMA is especially prone to, as illustrated in the scanning electron microscope (SEM) in Figure 10.23.

To make throughput for deep-etch lithography more acceptable to industry, several avenues to more sensitive x-ray resists have been pursued. For example, copolymers of PMMA were investigated: methylmethacrylate (MMA) combined with methacrylates with longer ester side chains show sensitivity increases of up to 32% (with tertiary butylmethacrylate). Unfortunately, a deterioration in structure quality was observed.[59] Among the other possible approaches for making PMMA more x-ray sensitive, we can count on the incorporation of x-ray-absorbing high-atomic-number atoms or the use of chemically amplified photoresists. X-ray resists explored more recently for LIGA applications are poly(lactides), for example, poly(lactide-co-glycolide) (PLG), polymethacrylimide (PMI), polyoxymethylene (POM), and polyalkensulfone (PAS). PLG is a new positive resist developed by BASF AG, more sensitive to x-rays by a factor of two to three compared with PMMA. From the comparison of different resists for deep x-ray lithography in Table 10.12, PLG emerges as the most promising LIGA

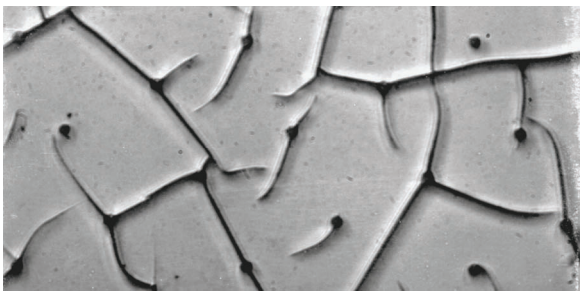


FIGURE 10.23 Cracking of PMMA resist. Method to test stress in thick resist layers. The onset of cracks in a pattern of holes with varying size (e.g., 1–4 µm) in a resist is shifted toward smaller hole diameter the lower the stress in the film. The SEM picture displays extensive cracking incurred during development of the image in a 5-µm-thick PMMA layer on an Au-covered Si wafer. The 5-µm-thick PMMA layer resulted from five separate spin coats. Annealing pushed the onset of cracking toward smaller holes until the right cycle was reached and no more cracks were visible. (From Madou, M. J., and M. Murphy. 1995. Unpublished results.[58])

Table 10.12 Properties of Resists for Deep X-Ray Lithography

	PMMA	POM	PAS	PMI	PLG
Sensitivity	–	+	++	0	0
Resolution	++	0	––	+	++
Sidewall smoothness	++	––	––	+	++
Stress corrosion	–	++	+	––	++
Adhesion on substrate	+	+	+	–	+

Source: Ehrfeld, W. 1994. Notes from handouts. The Commercialization of Microsystems. September 11–16, Banff Conference Center, Banff, Canada.[60]

++, excellent; +, good; 0, reasonable; –, bad; ––, very bad.

resist. POM, a promising mechanical material, may also be suited for medical applications given its biocompatibility. All the resists shown in Table 10.12 exhibit significantly enhanced sensitivity compared with PMMA, and most exhibit a reduced stress corrosion.[57] Negative x-ray resists have inherently higher sensitivities compared with positive x-ray resists, although their resolution is limited by swelling. Poly(glycidyl methacrylate-co-ethyl acrylate) (PGMA), a negative e-beam resist (not shown in Table 10.12), has also been used in x-ray lithography. In general, resist materials sensitive to e-beam exposure also display sensitivity to x-rays and function in the same fashion; materials positive in tone for e-beam radiation typically are also positive in tone for x-ray radiation. A strong correlation exists between the resist sensitivities observed with these two radiation sources, suggesting that the reaction mechanisms might be similar for both types of irradiation. IMM, in Germany, started developing a negative x-ray resist 20 times more sensitive than PMMA, but the exact chemistry has not yet been disclosed.[2] More common x-ray resists from the IC industry are reviewed in Table 10.13.

Methods of X-Ray Resist Application

Multiple Spin Coats

Different methods to apply ultrathick layers of x-ray resists such as PMMA have been studied. In the case of multilayer spin coating, high interfacial stresses between the layers can lead to extensive crack propagation on developing the exposed resist. For example, in Figure 10.23, we present an SEM picture of a

Table 10.13 Few Common Resists for E-Beam and X-Ray Lithography

Resist	Tone	EBL Sens ($\mu C/cm^2$)	EBL Contrast	XRL Sens (mJ/cm^2)	XRL Contrast
PMMA	+	100	2.0	6500	2.0
PBS	+	1	2.0	170	1.3
EBR-9	+	1.2	3.0		
Ray-PF	+			125	*
COP	–	0.5	0.8	100	1.1
GMCIA		7.0	1.7		
DCOPA	–			14	1.0
Novakak based	*	200–500	2–3	750–2000	~2

Source: Campbell, S. A. 1996. *The science and engineering of microelectronic fabrication.* New York: Oxford University Press.[61]

*Value is process-dependent.

5-µm-thick PMMA layer deposited in five sequential spin coatings. Development results in the cracked riverbed mud appearance, with the most intensive cracking propagating from the smallest resist features. The test pattern used to expose the resist consisted of arrays of holes ranging in size from 1–4 µm. Annealing the PMMA films shifted the cracking toward holes with smaller diameter compared with the unannealed film shown in Figure 10.23.[58] CAMD (see Table 10.8) has demonstrated that multiple spin coating of PMMA can be used for up to 15-µm-thick resist layers and that applying the appropriate annealing and developer (see below) eliminates cracking.[62] Below, we will learn that a prerequisite for low stress and small lateral tolerances in PMMA films is a high mean molecular weight. The spin-coated resist films in Figure 10.23 do not have a high-enough molecular weight to lead to good enough selectivity between radiated and nonradiated PMMA during a long development process.

Commercial PMMA Sheets

High-molecular-weight PMMA is commercially available as prefabricated plate (e.g., GS 233, Rohm GmbH, Darmstadt, Germany), and several groups have used freestanding or bonded PMMA resist sheets for producing LIGA structures.[42,52] After overcoming the initial problems encountered when attempting to glue PMMA foils to a metallic base plate with adhesives, this has become the preferred method in several labs.[52] Guckel used commercially available thick PMMA sheets (thickness >3 mm), XY-sized and

solvent-bonded them to a substrate, and, after milling the sheet to the desired thickness, exposed the resist plate without cracking problems.[42]

Casting of PMMA

PMMA also can be purchased in the form of a casting resin, for example, Plexit 60 (PMMA without added cross-linker) and Plexit 74 (PMMA with cross-linker added) from Rohm GmbH. In a typical procedure, PMMA is in situ polymerized from a solution of 35 wt% PMMA of a mean molecular weight of anywhere from 100,000 g/mol up to 106 g/mol in methylmethacrylate (MMA). Polymerization at room temperature takes place with benzoyl peroxide (BPO) catalyst as the hardener (radical builder) and dimethylaniline (DMA) as the starter or initiator.[52,59] The oxygen content in the resin, inhibiting polymerization, and gas bubbles, inducing mechanical defects, are reduced by degassing while mixing the components in a vacuum chamber at room temperature and at a pressure of 100 mbar for 2–3 min.

In a practical application, resin is dispensed on a base plate provided with shims to define pattern and thickness and subsequently covered with a glass plate to avoid oxygen absorption. Because of the hardener, polymerization starts within a few minutes after mixing the components and comes to an end within 5 min. The glass cover plate is coated with an adhesion-preventing layer (e.g., Lusin L39, Firma Lange u. Seidel, Nuremberg, Germany). After polymerization, the antiadhesion material is removed by diamond milling, and a highly polished surface results. In situ polymerization and commercial-cast PMMA sheets top the list of thick resist options in LIGA today.

Plasma polymerization of PMMA in layers (>100 μm) was discussed in Chapters 7 and 8 (see Figure 8.9).

Resist Adhesion

Adhesion promotion by mechanically or chemically modifying the primary substrate was introduced above, under "Choice of Primary Substrate." Smooth surfaces such as Si wafers with an average roughness, R_a, smaller than 20 nm pose additional adhesion challenges often solved by modifying the resist itself. To promote adhesion of resist to polished untreated surfaces, such as metal-coated Si wafers, coupling agents must be used to chemically attach the resist to the substrate. An example of such a coupling agent is methacryloxypropyl trimethoxy silane (MEMO). With 1 wt% of MEMO added to the casting resin, excellent adhesion results. The adherence is brought about by a siloxane bond between the silane and the hydrolyzed oxide layer of the metal. As illustrated in

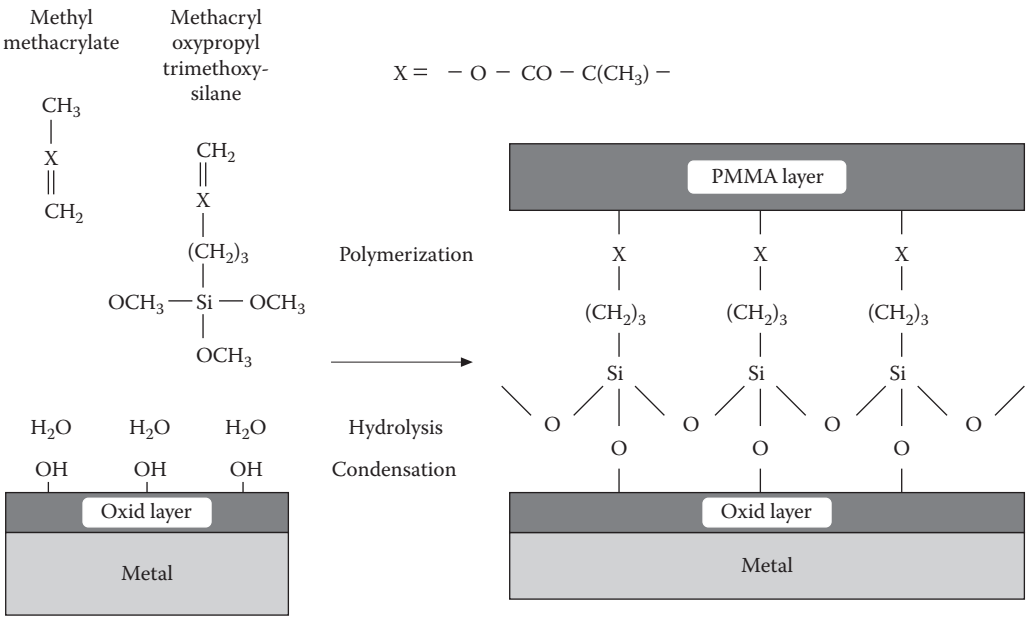


FIGURE 10.24 Schematic presentation of the adherence mechanism of methacryloxypropyl trimethoxy silane (MEMO). (From Mohr, J., W. Ehrfeld, and D. Munchmeyer. 1988. Requirements on resist layers in deep-etch synchrotron radiation lithography. *J Vac Sci Technol* B6:2264–67. With permission.[59])

Figure 10.24, the integration of this coupling agent in the polymer matrix is achieved via the double bond of the methacryl group of MEMO.[59]

Hydroxyethyl methacrylate (HEMA) can improve PMMA adhesion to smooth surfaces, but higher concentrations are needed to obtain the same adhesion improvement. Silanization of polished surfaces before PMMA casting, instead of adding adhesion promoters to the resin, did not seem to improve the PMMA adhesion. In the case of PMMA sheets, as mentioned before, one option is solvent bonding of the layers to a substrate. In another approach, Galhotra et al.[63] simply mechanically clamped the exposed and developed self-supporting PMMA sheet onto a 1.0-mm-thick Ni sheet for subsequent Ni plating. Rogers et al.[64] have shown that cyanoacrylate can be used to bond PMMA resist sheets to a Ni substrate and that it can be lithographically patterned using the same process sequence used to pattern PMMA. For a 300-µm-thick PMMA sheet on a sputtered Ni coating on a silicon wafer, a 10-µm-thick cyanoacrylate bonding layer was used. Such a thick cyanoacrylate layer caused some problems for subsequent uniform electrodeposition of metal. The dissolution rate of the cyanoacrylate is faster than the PMMA resist, resulting in metal posts with a wide profile at the base.

Stress-Induced Cracks in PMMA

The internal stress arising from the combination of a polymer on a metallic substrate can cause cracking in the microstructures during development. To reduce the number of stress-induced cracks (see Figure 10.23), both the PMMA resist and the development process must be optimized. Detailed measurements of the heat of reaction, the thermomechanical properties, the residual monomer content, and the molecular weight distribution during polymerization and soft baking have shown the necessity to produce resist layers with a high molecular weight and with only a very small residual monomer content.[52,59] Figure 10.25 compares the molecular weight distribution determined by gel permeation chromatography of a polymerized PMMA resist (two hardener concentrations were used) with the molecular weight distribution of the casting resin. The casting resin is unimodal, whereas the polymerized resist layer typically shows a bimodal distribution, with peak molecular weights centered around 90,000 and 300,000 g/mol. The first low-molecular-weight peak belongs to the PMMA oligomer dissolved in the casting resin, and the second molecular-weight peak results from the polymerization of the monomer. The molecular weight distribution is constant across the total resist thickness, except for the boundary layer at the base plate, where the average molecular weight can be significantly higher (~450,000 g/mol).[55]

The amount of the high-molecular-weight portion in the polymerized resist depends on the concentration of the hardener. A low hardener content leads to a high-molecular-weight dominance and vice versa (see Figure 10.25).[59] Because high molecular weight is required for low stress, a hardener concentration of

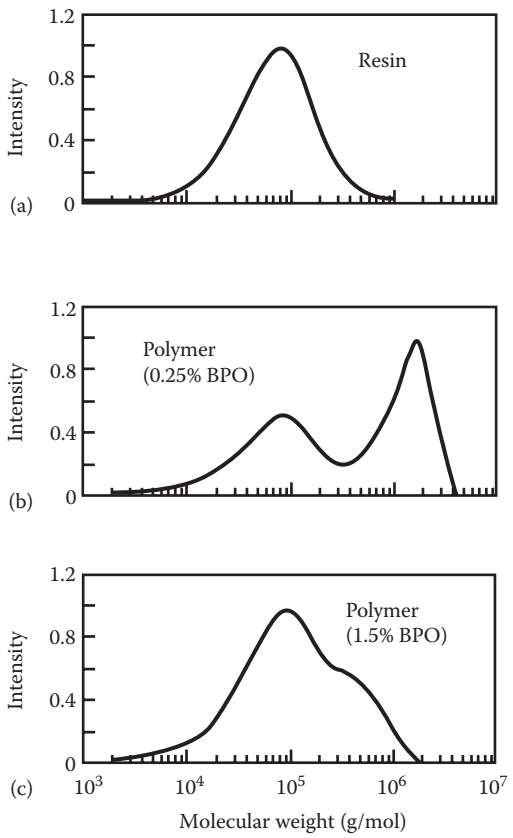


FIGURE 10.25 Molecular weight distribution of the casting resin (a), a resist layer polymerized at low hardener content (b), and a resist layer polymerized at high hardener content (c) determined by gel permeation chromatography. (From Mohr, J., W. Ehrfeld, and D. Munchmeyer. 1988. Requirements on resist layers in deep-etch synchrotron radiation lithography. *J Vac Sci Technol* B6:2264–67. With permission.[59])

less than 1% BPO must be used. Ideally, for a low stress resist, the residual monomer content should be less than 0.5%. The residual monomer content decreases with increasing hardener content, and >1% BPO is needed to reduce the residual monomer content to less than 0.5%. The problem resulting from these opposite needs can be overcome by the addition of 1% of a cross-linking dimethacrylate (triethylene glycol dimethacrylate) to the resin. In such cross-linked PMMA, a smaller amount of BPO suffices to suppress the residual monomer content; crack-free PMMA can be obtained with 0.8% of BPO.[52]

For solvent removal, and to further minimize the defects caused by stress, the polymerized resin is cured at 110°C for 1 h (soft bake). The measurement of the reaction enthalpy shows that postpolymerization reactions occur at room temperature and during heating to the glass transition temperature.[59] The rate of heating up to that temperature is 20°C/h; after curing, the samples are cooled down from 110°C to room temperature at a very low rate of 5–10°C/h.[52, 55] The soft-bake temperature is slightly below the glass transition temperature measured to be 115°C.

Another important factor that reduces stress in thick PMMA resist layers is the optimization of the developer. Stress-induced cracking can be minimized with solvent mixtures whose dissolution parameters lie near the boundary of the PMMA solubility range; that is, a nonaggressive solvent is preferred. This is discussed in more detail below under "Development." Small amounts of additives such as described above for reducing stress or to promote adhesion do not influence the mechanical stability of the microstructures or the sensitivity of the resist.

Exposure

Optimum Wavelength

For a given polymer, the lateral dimension variation in a LIGA microstructure could, in principle, result from the combined influence of several mechanisms. These include Fresnel diffraction, the range of high-energy photoelectrons generated by the x-rays, the finite divergence of synchrotron radiation, and the time evolution of the resist profiles during the development process. The theoretical manufacturing precision obtainable by deep x-ray lithography was investigated by means of computer simulation of both the irradiation step and the development step by Becker et al.[65] and Munchmeyer.[66] Results were further tested experimentally and confirmed by Mohr et al.[52] The theoretical results demonstrate that the effect of Fresnel diffraction (edge diffraction), which increases as the wavelength increases, and the effect of secondary electrons in PMMA, which increases as the wavelength decreases, lead to minimal structural deviations when the characteristic wavelength ranges between 0.2 and 0.3 nm (assuming an ideal development process and no x-ray divergence). To fully use the accuracy potential of a 0.2–0.3-nm wavelength, the local divergence of the synchrotron radiation at the sample site should be less than 0.1 mrad. Under these conditions, the variation in critical lateral dimensions likely to occur between the ends of a 500-µm-high structure as a result of diffraction and secondary electrons is estimated to be 0.2 µm. The estimated Fresnel diffraction and secondary electron scattering effects are shown as a function of characteristic wavelength in Figure 10.26.

Using cross-linked PMMA, or linear PMMA with a unimodal and extremely high-molecular-weight distribution (peak molecular weight greater than 1 million g/mol), the experimentally determined lateral tolerances on a test structure, as shown in Figure 7.27, are 0.055 µm/100-µm resist thickness, in good agreement with the 0.2 µm over 500 µm expected

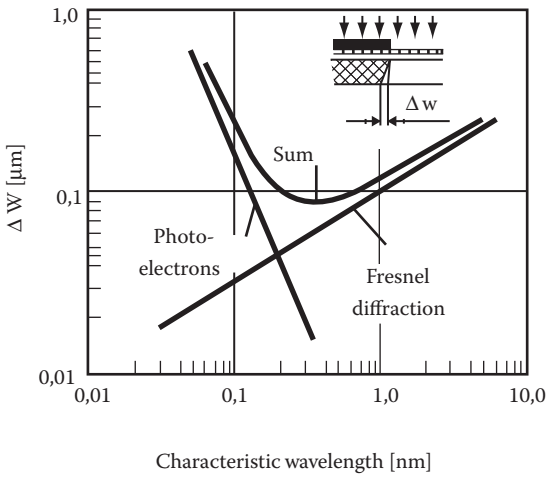


FIGURE 10.26 Fresnel diffraction and photoelectron generation as a function of characteristic wavelength, λ_c, and the resulting lateral dimension variation (ΔW). (From Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure*. Weinheim, Germany: VCH Publishers.[67])

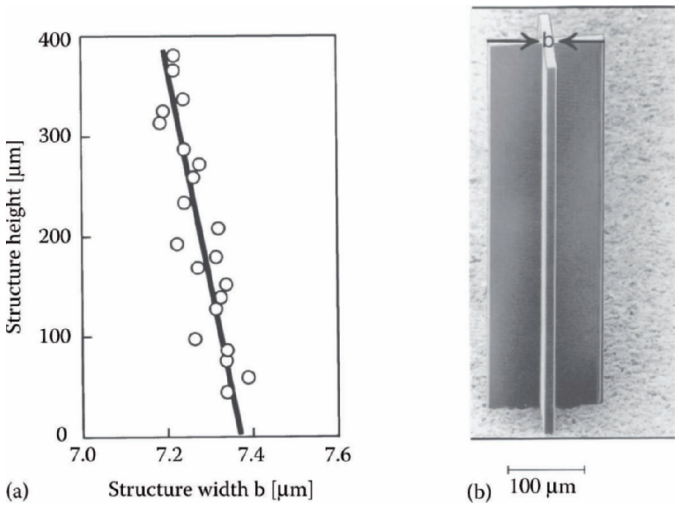


FIGURE 10.27 Structural tolerances. (a) SEM micrograph of a test structure to determine conical shape. (b) Structural dimensions as a function of structure height. The tolerances of the dimensions are within 0.2 mm over the total structure height of 400 mm. (From Mohr, J., W. Ehrfeld, and D. Munchmeyer. 1988. Requirements on resist layers in deep-etch synchrotron radiation lithography. *J Vac Sci Technol* B6:2264–67.[59]) (Courtesy of the Karlsruhe Nuclear Research Center.)

on a theoretical basis.[52] These results are obtained only when a resist/developer system with a ratio of the dissolution rates in the exposed and unexposed areas of approximately 1000 is used.

The use of resist layers, not cross-linked and displaying a relatively low bimodal molecular-weight distribution, as well as the application of excessively strong solvents such as used to develop thin PMMA resist layers in the IC industry, leads to a more pronounced conical shape in the test structure of Figure 10.23. An illustration of the effect of molecular weight distribution on lateral geometric tolerances is that linear PMMA with a peak molecular weight less than 300,000 g/mol shows structure tolerances of up to 0.15 µm/100 µm.[52] To obtain the best tolerances requires a PMMA with a very high molecular weight, also a prerequisite for low stress in the developed resist. Finally, if the synchrotron beam is not parallel to the absorber wall but at an angle greater than 50 mrad, greater coning angles may also result.[52]

Deposited Dose

As shown in Figure 10.28, depicting the average molecular weight of PMMA as a function of radiation

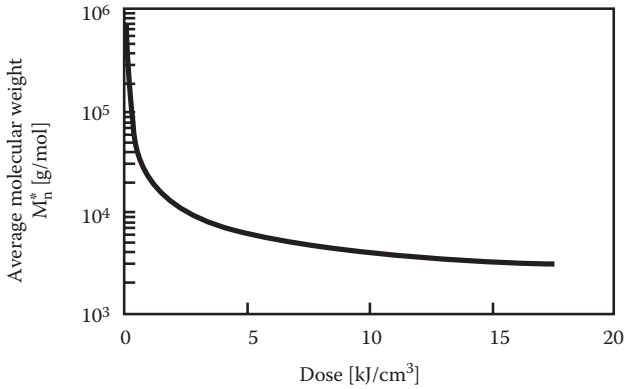


FIGURE 10.28 Average molecular weight M_n^* versus x-ray radiation dose. (From Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure.* Weinheim, Germany: VCH Publishers. With permission.[67])

dose, the x-ray irradiation of PMMA reduces the average molecular weight (M_n', see Chapter 1, Equation 1.10).[67] For one-component positive resists, this decrease of the average molecular weight causes the solubility of the resist in the developer to increase dramatically. The average molecular weight making dissolution possible is a sensitive function of the type of developer used and the development temperature. It can be observed from Figure 10.28 that above a certain dose (15–20 kJ/cm^3), the average molecular weight does not decrease any further.

The molecular weight distribution, measured after resist exposure, is unimodal with peak molecular weights ranging from 3,000–18,000 g/mol, dependent on the dose deposited during irradiation. The peak molecular weight increases nearly linearly with increasing resist depth; that is, decrease of the absorbed dose.[55] At the bottom of the resist layer, the absorbed dose must be higher than the development dose, D_{dv}, whereas at the top of the resist, the absorbed dose must be lower than the damaging dose, D_{dm}. In Figure 10.18, where the absorption of x-rays along the path from source to sample was illustrated, the exposure time and the preabsorber were chosen so that the bottom of a 500-µm-thick PMMA layer received the necessary development dose D_{dv}, whereas the dose at the top of the layer stayed well below D_{dm}. Exposure of PMMA with longer wavelengths results in correspondingly longer exposure times and can lead to an overexposure of the top surface, where the lower energy radiation is mainly absorbed.

Menz and Bley[67] describe the influence of the radiation dose on the quality of the resulting LIGA structures in a slightly different manner. Following their approach, Figure 10.29a illustrates a typical bimodal molecular weight distribution of PMMA before radiation, exhibiting an average molecular weight of 600,000. The gray region in this figure indicates the molecular weight region where PMMA readily dissolves; that is, less than the 20,000 g/mol level for the temperature and developer used. Because the fraction of PMMA with a 20,000 molecular weight is very small in nonirradiated PMMA, the developer hardly attacks the resist. After irradiation with a dose D_{dv} of 4 kJ/cm^3, the average molecular weight becomes low enough to dissolve almost all the resist (Figure 10.29b). With a dose D_{dm} of 20 kJ/cm^3, all the PMMA dissolves swiftly (Figure 10.29c). At a dose more than D_{dm}, the microstructures are destroyed by the formation of bubbles. It follows that to dissolve PMMA completely and to make defect-free microstructures, the radiation dose for the specific type of PMMA used must lie between 4 and 20 kJ/cm^3. These two numbers also lock in a maximum value of five for the ratio of the radiation dose at the top and bottom of a PMMA structure. To make this ratio as small as possible, the soft portion of the synchrotron radiation spectrum is usually filtered out by a preabsorber [for example, a 100-µm-thick polyimide foil (Kapton)] to reduce differences in dose deposition in the resist.

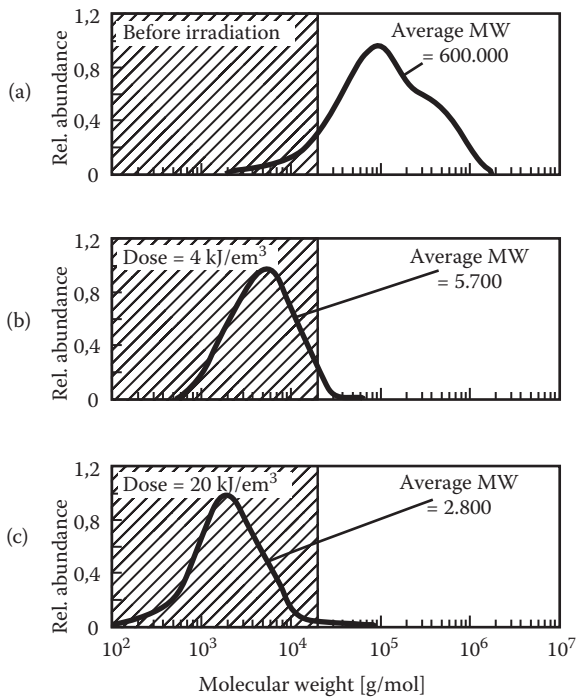


FIGURE 10.29 Molecular weight distribution of PMMA before (a) and after irradiation with 4 (b) to 20 kJ/cm^3 (c). The shaded areas indicate the domain in which PMMA is minimally 50% dissolved (at 38°C in the LIGA developer described in the text). (From Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure*. Weinheim, Germany: VCH Publishers. With permission.[67])

X-ray exposure equipment developed by KfK has vibration-free bedding; the exposure chamber is under thermostatic control (±0.2°C) and includes a precision scanner for the periodic movement of the sample through the irradiation plane. The polyimide window isolates the vacuum of the accelerator from the helium atmosphere (200 mbar), which serves as coolant for substrate and mask in the irradiation chamber.[43] IMM, in collaboration with Jenoptik GmbH, developed an x-ray scanner for deep lithography, enabling irradiation of up to 1000-µm-thick resists. A mask-to-resist registration within ±0.3 µm is claimed.[68]

Stepped and Slanted Microstructures

For many applications, stepped or inclined resist sidewalls are very useful—consider, for example, the fabrication of multilevel devices or prisms or, more basic yet, angled resist walls to facilitate the release of molded part.* Using stepped absorber layers on a single mask to make stepped multilevel microstructures was discussed above under "Absorber Fabrication" (see also the discussion in Chapter 2 and Figure 2.11). We mentioned that structures made this way are not always very well resolved. To make better-resolved stepped features one can first relief print a PMMA layer, for example, by using a Ni mold insert made from a first x-ray mask. Subsequently, the relief structure may be exposed to synchrotron radiation to further pattern the polymer layer through a precisely adjusted second x-ray mask. To carry out this process, a two-layer resist system needs to be developed consisting of a top PMMA layer, which

* In using LIGA or LIGA-like techniques for plastic micromolding it is important to be able to control the resist wall inclination to facilitate the release of molded devices. Even a 2° angle can make the difference between form locking or easy release.

fulfills the requirement of the relief printing process, and a bottom layer, which fulfills the requirements of the x-ray lithography.[69] The bottom resist layer promotes high molecular weight and adhesion, whereas the top PMMA layer is of lower molecular weight and contains an internal mold-release agent. This process sequence combining plastic impression molding with x-ray lithography is illustrated in Figure 10.30. The two-step resist then facilitates the fabrication of a mold insert by electroforming, which can be used for the molding of two-step plastic structures. Extremely large structural heights can be obtained from the additive nature of the individual microstructure levels.

There are several options to achieve miniaturized features with slanted walls. It is possible to modulate the exposure/development times of the resist, fabricate an inclined absorber, angle the radiation, or, following Tabata et al.,[71] move the mask during exposure in so-called *moving mask deep x-ray lithography* (M^2DXL) (Example 10.2).

To make a slanted absorber, a slab of material can be etched into a wedge by pulling it at a linear rate out of an etchant bath. Changing the angle at which synchrotron radiation is incident on the resist, usually 90°, also enables the fabrication of microstructures with inclined sidewalls.[72] This way, slanted microstructures may be produced by a single oblique irradiation or by a swivel irradiation (see Figure 2.# for an example). One potentially very important application of microstructures incorporating inclined sidewalls is the vertical coupling of light into waveguide structures using a 45° prism.[73] Such optical devices must have a wall roughness of less than 50 nm, making LIGA a preferred technique for this application. The sharp decrease of the dose in the resist underneath the edge of the inclined absorber, and the resulting sharp decrease of the dissolution of the resist as a function of the molecular weight in the developer, results in little or no deviation of the inclination of the resist sidewall over the total height of the microstructure.

In Chapter 8, we described additional means to fabricate slanted and curved metal shapes using electrochemical and electroless plating.

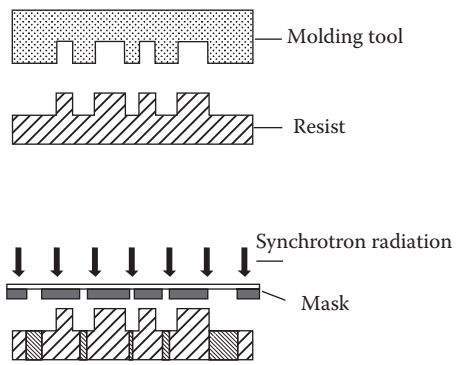


FIGURE 10.30 Stepped microstructures. (From Mohr, J., C. Burbaum, P. Bley, W. Menz, and U. Wallrabe. 1990. *Microsystem technologies 90.* Ed. H. Reichl. Berlin: Springer. With permission.[70])

Development

X-ray radiation changes the polymer in the unmasked areas, and chemicals etch away the regions that have been exposed during development. To fully use the accuracy potential of synchrotron radiation lithography, it is essential that the resist/developer system have a ratio of dissolution rate in the exposed and unexposed areas of approximately 1000 (see above). The developer empirically arrived at by KfK consists of a mixture of 20 vol% tetrahydro-1,4-oxazine (an azine), 5 vol% 2-aminoethanol-1 (a primary amine), 60 vol% 2-(2-butoxyethoxy) ethanol (a glycolic ether), and 15 vol% water.[52,55,74] This developer causes an infinitely small dissolution of unexposed, high-molecular-weight, cross-linked PMMA and achieves a sufficient dissolution rate in the exposed area. It also exhibits much less stress-induced cracking than developers conventionally used for thin PMMA resists.

Systematic investigation of different organic solvents and mixtures of the above developer systems shows that solvents with a solubility parameter at the periphery of the solubility range of PMMA dissolve exposed PMMA slowly but selectively and without stress-induced cracking or swelling of unexposed areas. Solvents with a solubility parameter close to those of MMA show a much higher dissolution rate but cause serious problems related to cracking and swelling. As we have already seen, the application of excessively strong solvents also leads to more pronounced conical shapes in the test structures shown in Figure 10.27. An improved developer found in the above systematic investigation is a mixture of

tetrahydro-1,4-oxazine and 2-aminoethanol-1. Its sensitivity is 30% higher, but the process latitude is much narrower compared with the developer described above.[55]

KfK has built a machine for the development process, enabling the continuous and homogeneous transport of developing and rinsing agents into deep structure elements and the removal of the dissolved resist from these structures. Several substrates are arranged vertically on a rotor, with each structure surface facing out. During development, the developing agent flows toward the resist surface being developed, circulates, is filtered continuously, and the temperature remains controlled at 35°C. To stop development, less-concentrated developer solutions are applied to prevent the precipitation of already-dissolved resist. Three independent medium circuits are available for immersion and spraying processes.[43]

After development, the microstructures are rinsed with deionized water and dried in a vacuum. Alternatively, drying may be done by spinning and blasting with dry nitrogen. At this stage, the devices can be the final product, for example, as microoptical components, or they can be used for subsequent metal deposition.

Metal Deposition in LIGA and Pseudo-LIGA

Introduction

Electroplating, which is sometimes also called *electrodeposition*, is a process in which a metal is coated on a conductive surface through electrochemical reactions that are facilitated by an applied electrical potential. In this process, the surface to be coated is immersed into a solution of one or more metal salts. LIGA and pseudo-LIGA depend critically on metal deposition to replicate the master photoresist mold. The additive metal deposition process was detailed in Volume I, Chapter 7, where we discussed the electroplating of metals in high-aspect-ratio crevices, and in Chapter 8, this volume, where we compared electroless with electrodeposition and the crafting of slanted and curved plated metal shapes. In this section we add some modifications to the metal process when applied to high-aspect-ratio resist molds encountered in LIGA and pseudo-LIGA.

Nickel Electrodeposition

Physical Properties of Electrodeposited Nickel

Nickel, together with copper, is one of the most widely used metals in electroplating LIGA and pseudo-LIGA structures. A nickel sulfamate bath composition, optimized for Ni electrodeposition, is presented in Table 10.14. In addition to nickel sulfamate and boric acid as a buffer, a small quantity of an anion-active wetting agent is usually added. Sulfur-depolarized nickel pellets are used as anode materials (sulfur-depolarized nickel has sulfur cast into it to aid in anode corrosion and keep the anode from going passive) and are held in a titanium basket.[75]

Table 10.14 also lists operational parameters. Nickel sulfamate baths produce low internal stress deposits without the need for additional agents, thus avoiding a cause for more defects. The bath is operated at 50–62°C and at a pH value between 3.5 and 4.0. Metal deposition is carried out at current densities up to 10 A/dm^2. Growth rates vary from 12 (at 1 A/dm^2) to 120 µm/h (at 10 A/dm^2).

The hardness of the Ni deposits can be adjusted from 200–350 Vickers by varying the operating conditions. Hardness decreases with increasing current density. To reach a high hardness of 350 Vickers, the electroforming must proceed at a reduced current of 1 A/dm^2. Also, for low compressive stress of 20 N/mm^2 or less, a reduced current density must be used. Internal stress in the Ni deposits is not only influenced by current density but also by the layer thickness, pH, temperature, and solution agitation. In the case of pulse plating (see below under Pulse Plating), pulse frequency has a distinct influence as well.[76] From Figure 10.31 we derive that for thin Ni deposits the stress is high and decreases very fast as a function of thickness. For thick Ni deposits (>30 µm),

Table 10.14 Composition and Operating Conditions of Nickel Sulfamate Bath

Parameter	Value
Nickel metal (as sulfamate)	76–90 g/L
Boric acid	40 g/L
Wetting agent	2–3 mL/L
Current density	1–10 A/dm^2
Temperature	50–62°C
pH	3.5–4.0
Anodes	Sulfur depolarized

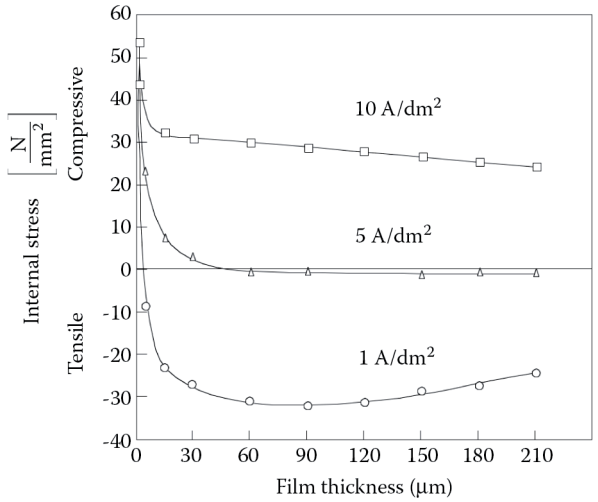


FIGURE 10.31 Influence of nickel layer thickness on internal stress. The electrolyte used is described in Table 10.14 (pH = 4; bath temperature = 52°C). (From Harsch, S., W. Ehrfeld, and A. Maner. 1988. *KfK report no. 4455.* Karlsruhe, Germany: KfK.[76]) (Courtesy of the Karlsruhe Nuclear Research Center.)

the stress as a function of thickness reaches a plateau. At a current density of 10 A/dm², these thick Ni films are under compressive stress; at 1 A/dm², they are under tensile stress; and at 5 A/dm², the internal stress reduces to practically zero. Stirring of the plating solution reduces stress dramatically, indicating that mass transport to the cathode is an important factor in determining the ultimate internal stress. Consequently, because high-aspect-ratio features do not experience the same agitation of a stirred solution as bigger features, this results in higher stress concentration in the smallest features of the electroplated structure. Because the stress is most severe in the thinnest Ni films, Harsch et al.[76] undertook a separate study of internal stresses in 5-µm-thick Ni films. For three plating temperatures investigated (42°C, 52°C, and 62°C), 5-µm-thick films were found to exhibit minimal or no stress at a current density of 2 A/dm². At 62°C, the 5-µm-thick films show no stress at 2 A/dm² and remain at a low compressive stress value of about 20 N/mm² for the whole current range (1–10 A/dm²). The internal stress at 2 A/dm² is minimal at a pH value between 3.5 and 4.5 of a sulfamate electrolyte (Table 10.14). Ni concentration (between 76 and 100 g/L) and wetting agent concentration do not seem to influence the internal stress. The higher the frequency of the pulse in pulse plating (see next section), the smaller the internal stress.

The long-term mechanical stability of Ni LIGA structures was investigated by electromagnetic activation of Ni cantilever beams by Mohr et al.[77] The number of stress cycles, N, necessary to destroy a mechanical structure depends on the stress amplitude, S, and is determined from fatigue curves, or S-N curves (applied stress on the x-axis and number of cycles necessary to cause breakage on the y-axis). Experimental results show that for Ni cantilevers produced by LIGA the long-term stability reaches the range of comparable literature data for bulk annealed and hardened nickel specimens. Usually, stress leads to crack initiation, which often starts at the surface of the structure. Because microstructures have a higher surface-to-volume ratio, one might have expected the S-N curves to differ from macroscopic structures, but so far this has not been observed. To the contrary, it seems that the smaller structures are more stable.

Pulse Plating

For the fabrication of microdevices with high deposition rates exceeding 120 µm/h, a reduction of the internal Ni stress can occur only by increasing the temperature of the bath or by using alternative electrodeposition methods. Increasing the bath temperature is not an attractive option, but using alternative electrochemical deposition techniques deserves further exploration. For example, using pulsed (≥500 Hz) galvanic deposition instead of a DC method can influence several important properties of the Ni deposit. Properties such as grain size, purity, and porosity can be manipulated this way without the addition of organic additives.[76] In pulse plating, the current pulse is characterized by three parameters: pulse current density, i_p; pulse duration, t_d; and pulse pause, t_p. These three independent variables determine the average current density, i_a, which is the important parameter influencing the deposit quality and is given by:

$$i_a = \frac{t_d}{t_d + t_p} i_p \qquad (10.5)$$

Pulse plating leads to smaller metal grain size and smaller porosity as a result of a higher deposition potential. Because each pulse pause allows some

time for metal cation replenishment at the cathode (e.g., Ni^{2+} enrichment) and for diffusion away of undesirable reaction products that might otherwise get entrapped in the metal deposit, a cleaner metal deposit results. The higher the frequency of the pulse, the smaller the internal stress in the resulting metal deposit.[76]

Pulse plating represents only one of many emerging electrochemical plating techniques considered for microfabrication. For more background on techniques such as laser-enhanced plating, jet plating, laser-enhanced jet plating, and ultrasonically enhanced plating, refer to the review of Romankiw et al.[78] and references therein.

LIGA and Pseudo-LIGA Plating Issues

Electrodeposition in High-Aspect-Ratio Resist Crevices In Volume I, Chapter 7 we studied how a microelectrode of the same size or less than the diffusion layer thickness δ could be expected to plate faster than a larger electrode because of the extra current resulting from the nonlinear diffusion (Volume I, Equation 7.131). This turns out to be incorrect: as derived from Volume I, Equation 7.132, describing the current to a metal electrode recessed in a resist layer, the nonlinear diffusion contribution increases with decreasing radius r, but it also decreases with increasing resist thickness, L. High-aspect-ratio features consequently plate slower than low-aspect-ratio features. Moreover, the hydrogen consumption accompanying cathodic reduction of the depositing metal in the high-aspect-ratio features causes the pH to locally increase. Because agitation is impossible in these deep crevices, an isolating layer might form (e.g., nickel hydroxide, in the case of Ni deposition), preventing further metal deposition. This all contributes to make the metal deposition rate, important for an economical production, much smaller than the rates expected from the linear diffusion model of current density in large, low-aspect-ratio structures.

A major cause of defects in electroplating in high-aspect-ratio features is an incomplete wetting of the microchannels in the resist structure. For example, the contact angle between PMMA and the plating electrolyte at 50°C lies between 70 and 80°. Thus, a wetting agent is indispensable for wetting the surface of the plastic structures for the electrolyte to penetrate into the microchannels. With a wetting agent, the contact angle between PMMA and the plating electrolyte can be reduced from 80 to 5°.[76] In the electroforming of microdevices, a much higher concentration of wetting agent is necessary than in conventional electroplating. A dramatic illustration of this wetting effect can be seen in the case of plating of Ni posts, as shown in Figure 10.32. In the case illustrated in Figure 10.32a, where only 2.5 mL/L of a wetting agent is added to the sulfamate nickel deposition solution (see Table 10.14), nickel posts with a diameter of 50 μm often fail to form. In the same experiment, no posts with a diameter of 5 and 10 μm form. Increasing the wetting agent concentration to 10 mL/L results in perfectly formed nickel posts with a diameter of 5 μm, as shown in Figure 10.32b.

Plating Compatibility Issues Two major sources of difficulty associated with plating of any metal structure in photoresist molds, not only the high-aspect ones described above, are chemical and mechanical incompatibility. Chemical incompatibility means that the photoresist mask may be attacked by the plating solution; mechanical incompatibility is film stress in the plated layer, which causes the plated structure to lose adhesion to the substrate.

If the plating solution attacks the photoresist even slightly, considerable damage to the photoresist layer may occur by the time a 200-μm-thick structure has been created. Therefore, the limiting thickness of a plated structure resulting from a chemical interaction is dependent on both the photoresist chemistry and the plating bath itself. In general, the plating bath must have a pH in the range of ~3–9.5 (acidic to mildly alkaline), a fairly wide range that can accommodate many of the commercially available photoresists. A surprising number of plating baths do not fall into that range, however. Baths, either very strongly acidic or more than mildly alkaline, tend to attack and destroy photoresist.

The plated structure must have extremely low stress to avoid cracking or peeling during the plating process. In addition, if the structure contains narrow features, the stress must be tensile because any compressive stress would result in buckling of the structure. The primary concern with regard to

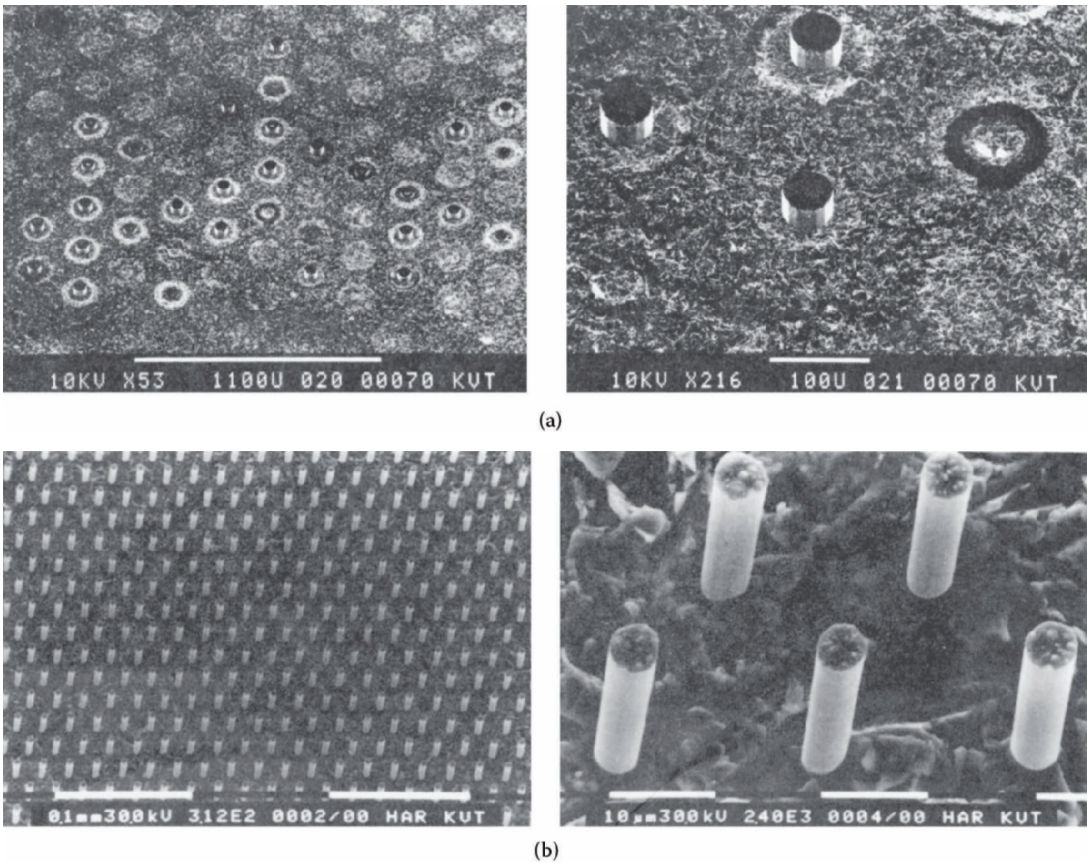


FIGURE 10.32 Electrodeposition of nickel posts of varying diameter. (a) Nickel posts with diameter of 50 µm. Only 2.5 mL/L wetting agent was added to the nickel sulfamate solution. Many 50-µm posts are missing, and posts with a diameter of 10 or 5 µm do not even form. (b) Nickel posts with diameter of 5 µm; 10 mL/L wetting agent was added, and all posts developed perfectly. (Courtesy of the Karlsruhe Nuclear Research Center.)

stresses is the incorporation of "brightening agents" into the plating solution. These can be selenium, arsenic, thallium, and ammonium ions, as well as others. The brightening materials often exhibit different atomic configurations and sizes compared with most materials that we would like to plate, such as Cu, Au, Ni, FeNi, Pt, Ag, and NiCo. This represents the source of a good deal of the stress intrinsic to the plating process. Ag and FeNi are notorious for their high stresses. The Ni-sulfamate bath, presently used extensively in LIGA, is used without brighteners, leading to Ni deposits with extremely low stress. Electroless plating baths (see Chapter 8) also are a good option: these *autocatalytic* baths coat indiscriminately but cause very low stress. They should be compatible with the LIGA process for certain applications, such as mold insert formation where significant overplating is required.

A third difficulty in finding plating baths compatible with LIGA is that most plating baths are not intended for use in the semiconductor industry. Information (including normal deposition parameters, uniformity, stress, compatibility with various semiconductor processes, and particulate contamination) normally supplied by a vendor is not available. Because the majority of the semiconductor industry does not incorporate these kinds of "thin film" processes, it may be difficult to determine whether a plating bath will be suitable simply from conversations with the vendor. In many instances, trial and error decide.

The treatise of electroplating in micropatterned resists is of importance for both micromachining and the IC industry. Additional reading on this topic is suggested (e.g., Masuko et al.,[79] Dukovic,[80] and Leyendecker et al.[81]).

Finishing Metal Parts and Metal Mold Inserts

Slight differences in final metal layer thickness cannot be avoided in the electroforming process. Finish

grinding of the metal samples with diamond paste is used to even out microroughness and slight variations in structural height. After finish grinding an electroplated LIGA or pseudo-LIGA work piece, a primary metal shape results from removing the photoresist by ashing in oxygen plasma or stripping in solvent. In the case of cross-linked PMMA, the resist may be exposed again to synchrotron radiation, guaranteeing sufficient solubility before stripping.

If the metal part needs to function as a mold insert, metal is plated several millimeters beyond the front faces of the resist structures to produce a monolithic micromold (see Figure 10.33). In the latter case, to avoid damage to the mold insert when separating it from the plating base, an intermediate layer sometimes is deposited on the base plate, ensuring adhesion of the resist structures while facilitating the separation of the electroformed mold insert from its plating base. In addition, it helps to prevent burrs from forming at the front face of the mold insert as a result of underplating of the resist structures. Underplating can occur because the resist does not adhere well to the substrate, allowing electrolyte solution to penetrate between the two, or the plating solution might attack the substrate/resist interface. Finally, microcracks at the interface might contribute to underplating. Burrs are easily eliminated then by dissolving the thin auxiliary metal layer with a selective etchant, removing the mold without the need for a mechanical load. In view of the observed underplating problems, it is surprising that Galhotra et al.[63] obtained good plating results by simply mechanically clamping the exposed and developed PMMA sheet to a Ni plating box.

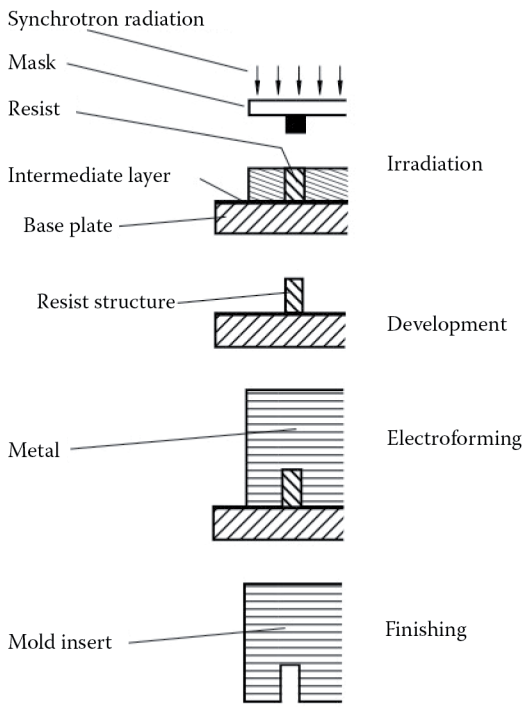


FIGURE 10.33 Fabrication of a LIGA mold insert. (From Hagmann, P., W. Ehrfeld, and H. Vollmer. 1987. Fabrication of microstructures with extreme structural heights by reaction injection molding. *First meeting of the European Polymer Federation, European symposium on polymeric materials*, pp. 241–51. Lyon, France. With permission.[28])

Plating Automation

An automated galvanoforming facility used by KfK is shown in Figure 10.34.[82] This setup includes provisions for on-line measurement of each electrolyte constituent in flow-through cells and concentration corrections when tolerance limits are exceeded. A computer-controlled transport system moves the individual plating racks, holding the microdevice substrates, through the process stages whereby the substrates are degreased, rinsed, pickled, electroplated, and dried and are then returned to a magazine, which can accommodate up to seven racks. The facility is designed as clean-room equipment because contamination of the microstructures must be prevented.[75] An instrument as described here can be bought, for example, from Reinhard Kissler (the µGLAV 750).[82] Although this type of automation is a must for the eventual commercialization of the LIGA technique, it should be recognized that at this explanatory stage overautomatization may be counterproductive.

Demolding

Introduction

Even high-aspect-ratio metal structures can be molded and demolded with polymers easily, as long as a polymer with a small adhesive power and rubber-elastic properties, such as silicone rubber, is used. However, rubber-like plastics have low shape stability and would not be adequate. Shape-preserving polymers, on the other hand, after hardening, require a mold with extremely smooth inner surfaces to prevent form-locking between the

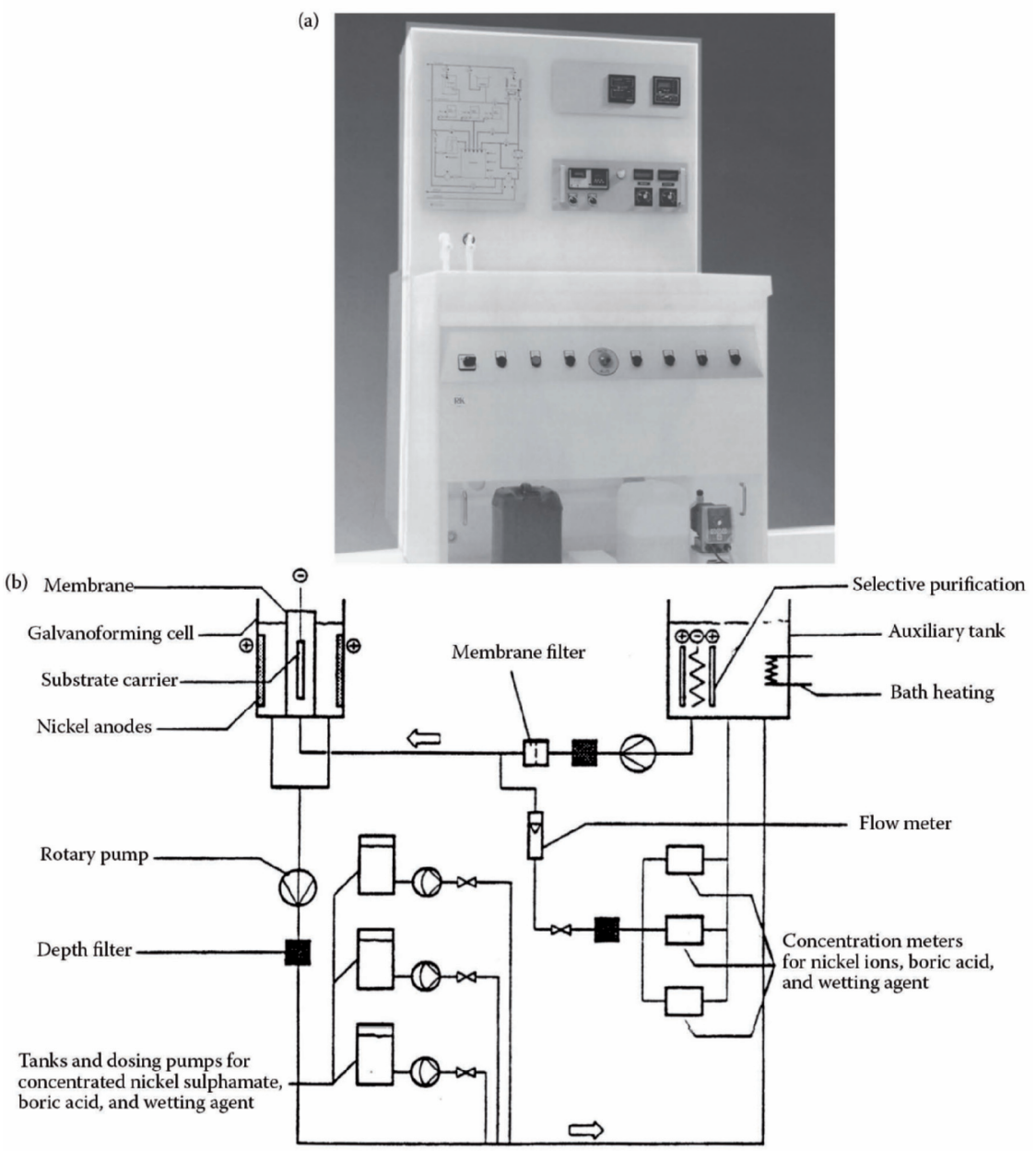


FIGURE 10.34 (a) The μGALV 750 comprises the galvanoforming cell with nickel anodes and substrate carrier and an auxiliary tank incorporating heating and purification auxiliary equipment. Three concentration meters measure nickel ions, boric acid, and the concentration of the wetting agent. The concentrations are adjusted automatically by adding via-metering pumps. (From Reinhard Kissler, GmbH, Speyer, Germany. 1994. With permission.) (b) Schematic drawing of a galvanoforming unit. (From Maner, A., S. Harsch, and W. Ehrfeld. 1988. Mass production of microdevices with extreme aspect ratios by electroforming. *Plating Surface Finish* 75:60–65. With permission.[75])

mold and the hardened polymer. Mold release has a geometry aspect (i.e., smooth and slightly inclined mold walls) and a chemical aspect (i.e., mold release agents). A mold release agent may be required for demolding. Using external mold release agents is difficult because of the small dimensions of the microstructures to be molded because typically they are sprayed onto the mold. Consequently, internal mold release compounds need to be mixed with the polymer, hopefully without significantly changing the polymer characteristics.

Demolding in LIGA and LIGA-Like Applications

The demolding step in LIGA is often carried out by means of a clamping unit at preset temperatures and rates. Demolding is facilitated by slightly inclined

mold walls and an internal mold release agent such as PAT 665 (Wurtz GmbH, Bingen, Germany), normally used with polyester resins.[83] The optimum yield with this agent and Plexit M60 PMMA occurs at 3–6 wt%. The yield decreases very quickly to less than 3 wt% as adhesion between the molded piece and the mold insert becomes stronger. These adhesive forces are estimated by qualitatively determining the forces necessary to remove the plastic structures from the mold insert. An upper limit is 10 wt%, where the MMA does not polymerize anymore. At more than 5 wt%, the Young's modulus and hence the mechanical stability decrease, and at more than 6 wt%, pores start forming in the microstructures. The internal mold release agent also has a marked influence on the optimum demolding temperature. The demolding yield decreases quickly at more than 60°C for a 4 wt% PAT internal mold release agent. With 6 wt%, one can only obtain good yields at 20°C.[28,83] The molding process in LIGA initially led to a production cycle time of 120 min. For a commercial process, much faster cycle times are needed, for example, by optimizing the mold release agent. With that in mind, the KfK group started to work with a special salt of an organic acid, leading to a 100% yield at a release agent content of 0.2 wt% only and a temperature of 40°C (at 0.05 wt%, a 95% yield is still achieved). At 80°C, a 100% demolding yield was obtained, and significantly a cycle time of 11.5 min was reached. During these experiments the mold was filled at 80°C and heated to 110°C within 7.5 min. As the curing occurs at 110°C, the material needs to be cooled down to 80°C for demolding. Moreover, the 0.2 wt% of the "magic release agent" did not impact the Young's modulus and the glass transition temperature of Plexit M60.[83]

Use of Sacrificial Layers in LIGA

As in surface micromachining, sacrificial layers make it possible to fabricate partially attached and freed metal structures in the primary mold process.[84] The ability to implement these features leads to assembled micromechanisms with submicrometer dimension accuracies, opening up many additional applications for LIGA, especially in the field of actuators. The sacrificial layer may be polyimide, silicon dioxide, polysilicon, or some metal.[85] The sacrificial layer is patterned with photolithography and wet etching before polymerizing the resist layer over it. At KfK, a several-micrometers-thick titanium layer often is used as the sacrificial layer because it provides good adhesion of the polymer, and it can be etched selectively against several other metals used in the process. If before exposure the x-ray mask is adjusted to the sacrificial layer pattern, some parts of the microstructures will lie above the openings in the sacrificial layer, whereas other parts will be built up on it. These latter parts will be able to move after removal of the sacrificial layer.[70] The fabrication of a movable LIGA structure is illustrated in Figure 10.35.

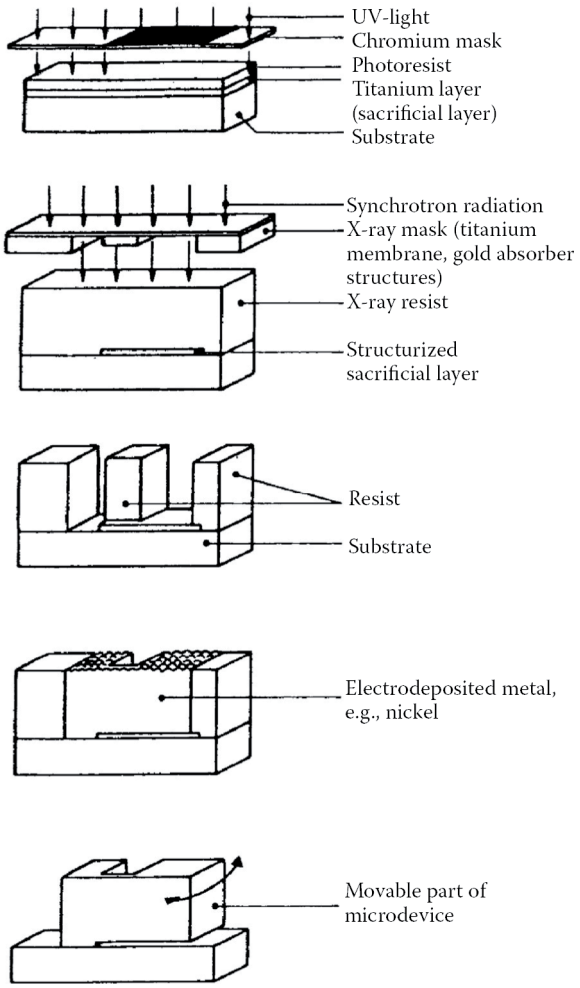


FIGURE 10.35 Movable LIGA microstructures. (From Mohr, J., C. Burbaum, P. Bley, W. Menz, and U. Wallrabe. 1990. *Microsystem technologies 90.* Ed. H. Reichl. Berlin: Springer. With permission.[70])

Molding Technology Applied to LIGA

Application of Reaction Injection Molding to LIGA

A laboratory-sized reaction injection molding (RIM) setup, as shown in Figure 10.36, was used by the KfK group for their LIGA work.[83] The setup consists of a container for mixing the various reactants, a vacuum chamber for evacuation of the mold cavity, the molding tool, and a hydraulic clamping unit to open and close the vacuum chamber and the molding tool. After the vacuum chamber has been closed and evacuated, the tool is closed by the clamping unit. The mixed reagents are degassed in the materials container and, under a gas overpressure of up to 3 MPa, pushed through the opened inlet valve into the tool holding the evacuated insert mold. To compensate for shrinkage caused by polymerization, an overpressure of up to 30 MPa must be applied during hardening of the casting resin. If the holding overpressure is too low, sunken spots appear in the plastic because of shrinkage of the polymer. For PMMA, the volume shrinkage is about 21%. If the mold material is not degassed and the mold cavity is not evacuated, bubbles develop in the molded piece, resulting in defects and possible partial filling of the mold. To harden the mold material and anneal material stress, the RIM machine is operated at temperatures up to 150°C.

To fabricate RIM-based microproducts, a variety of resins were tried, including epoxy resins, silicone resins, and acrylic resins. The most promising results are obtained with resins on an MMA base to which an internal mold release agent is added to reduce adhesion of the molded piece to the walls of the metal mold. The mold insert in the evacuated tool is covered by means of an electrically conductive perforated plate, called the *gate plate* (Figure 10.37a). For filling the mold, injection holes are positioned above large, free spaces in the metal structure, and low-viscosity reactants fill the smaller sections laterally. After hardening of the molded polymer, a form-locking connection between the produced part and the gate plate is established at the injection holes, permitting the demolding of the part from the insert (Figure 10.37b). No damage to the mold insert is observed after up to 100 mold/demold cycles, even at the level of an SEM picture. The secondary plastic structures formed are exact replicas of the original PMMA structures obtained after x-ray irradiation and development. If the final product is plastic, the molding process can be simplified considerably by omitting the gate plate. The mold insert is then cast over with the molding polymer in a vacuum vessel, building a stable gate block over the mold insert after hardening. The polymeric gate block is used to demold the part from the insert.

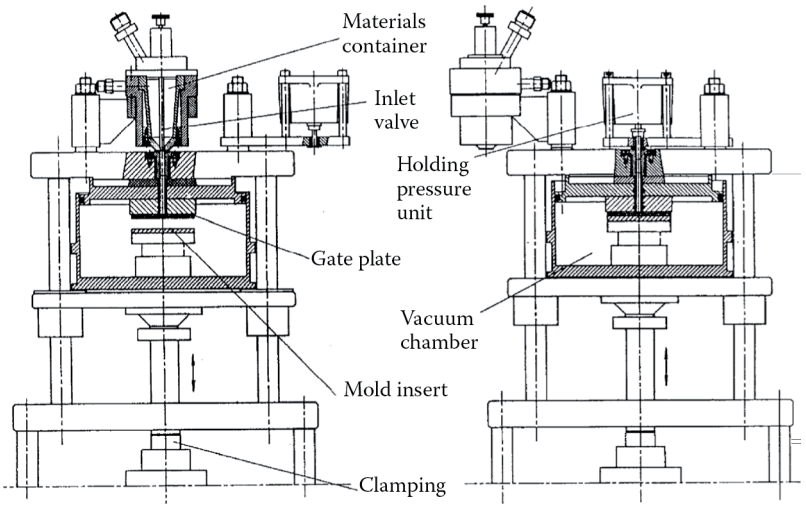


FIGURE 10.36 Schematic presentation of a vacuum molding setup. With minor changes, this setup can be used for reaction injection molding, thermoplastic injection molding, and compression molding. (From Hagmann, P., and W. Ehrfeld. 1988. Fabrication of microstructures of extreme structural heights by reaction injection molding. *J Polymer Process Soc* 4:188–95. With permission.[83])

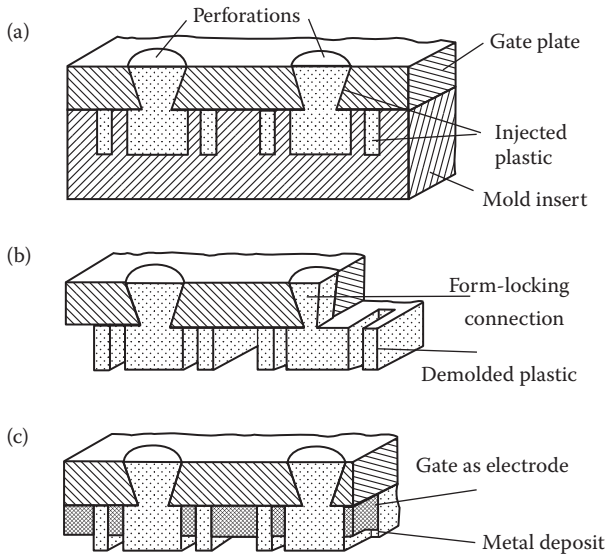


FIGURE 10.37 Molding process with a perforated gate plate. (a) Filling of the insert mold with plastic material. (b) Demolding. (c) Electrodeposition with the gate plate as plating base. (From Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure*. Weinheim, Germany: VCH Publishers.[67])

The secondary plastic templates may not be the final product and can be filled with a metal by electrodeposition as with the primary molds. In this case, the secondary plastic templates must be provided with an electrode or plating base. The gate plate can be used directly as the electrode for the deposition of metal in the secondary plastic templates (see Figure 10.37c). It is important that the mold insert is pressed tightly against the gate plate so that no uniform isolating plastic film forms over the whole interface between the gate plate and the mold insert; otherwise, electrodeposition might be impossible. Safe sealing between mold insert and gate plate is achieved by use of soft-annealed aluminum gate plates into which the insert is pressed when closing the tool. It was found that the condition to be met for perfect deposition of metal does not require that the gate plate be entirely free of plastic.[83] It suffices that a number of electrical conducting points emerge from the plastic film. Transverse growth of the metal layer produced in electrodeposition allows a fault-free, continuous layer to be produced. The secondary metal microstructures are replicas of the primary metal structures, and a comparison of, for example, a master separation nozzle mold insert with a secondary separation nozzle structure did not reveal any differences in quality. After the metal forming, the top surface must be polished and the plastic structures must be removed. Depending on the intended use of the secondary metal shape, one can leave it connected to the gate plate, or the gate plate can be selectively dissolved or mechanically removed.

Gate plates can be applied only if the microstructures are interconnected by relatively large openings. The injection holes in the gate plate are produced by mechanical means with openings of about 1 mm in diameter and must be aligned with large openings in the mold insert. Many desirable microstructures cannot be built this way. For example, in Figure 10.38, we show a Ni honeycomb structure with a cell diameter of 80 µm, a wall thickness of 8 µm, and a height of 70 µm.[86] The holes in a gate plate are too large to inject plastic into this structure, and the interconnections in the honeycomb are too small to provide a good transverse movement of the polymerizing resin. In this case an electrically conductive plastic is used instead. The Ni plating occurs on the conductive plastic carrier plate, which is fused with the insulating plastic of the microstructure.

Application of Compression Molding to LIGA

Compression molding or relief printing is the method of choice for incorporating LIGA structures on a processed Si wafer. In a first step, the mold material is applied by polymerization onto the processed wafer with an electrically isolating layer and overlaid with a metal plating base (Figure 10.39a). At the glass transition temperature, the plastic is in a viscoelastic condition suitable for the impression molding step. At more than this temperature (about 160°C for PMMA), the plastic is patterned in vacuum by impression with the molding tool. To avoid damaging the electronic circuits by contact with the molding tool, the mold material is not completely displaced during molding. A thin, electrically insulating residual layer is left between the molding tool and the plating base on the processed wafer (Figure 10.39b). The residual layer is removed by reactive ion etching (RIE) in an oxygen plasma etch (Figure 10.39c), freeing the plating base for subsequent electrodeposition. The oxygen RIE process is as anisotropic as possible so

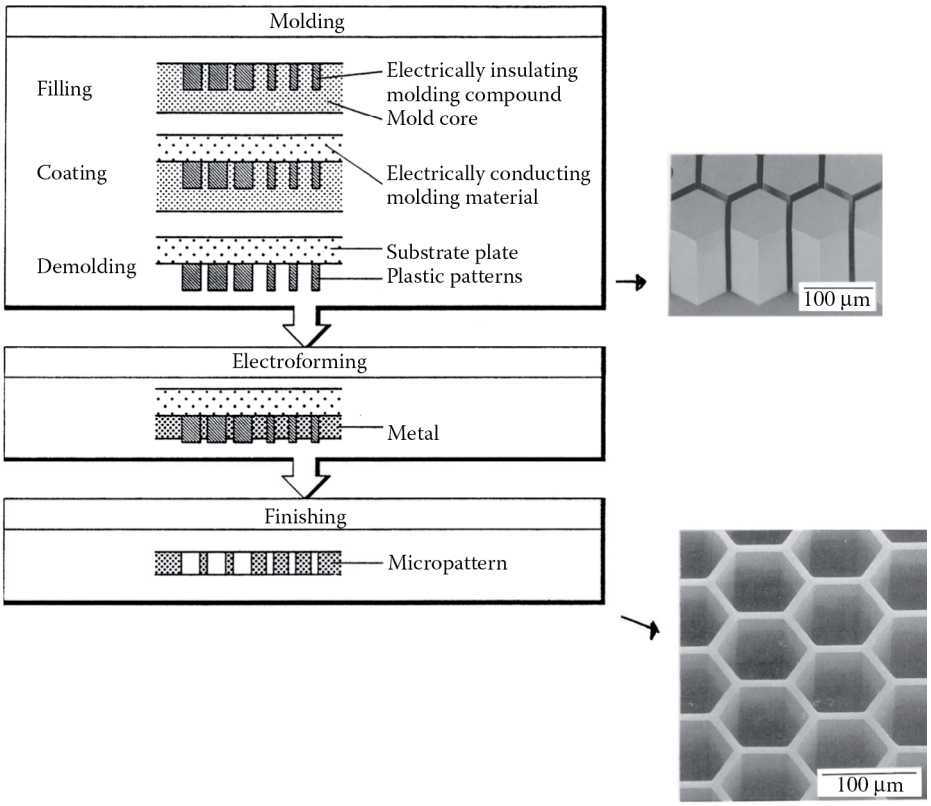


FIGURE 10.38 Ni honeycomb structure with cell diameter of 80 mm, wall thickness of 8 mm, and wall height of 70 mm. A structure like this could not be produced using a gate plate; a conductive plastic is used instead. [From Harmening, M., W. Bacher, and W. Menz. 1991. Molding plateable micropatterns of electrically insulating and electrically conducting poly(methyl methacrylate)s by the LIGA technique. *Makromol Chem Macromol Symp* 50:277–84. With permission.[86]]

that the sidewalls of the structures do not deteriorate, and the amount of material removed from the top of the plastic microstructures is small compared with the total height. After electrodeposition on the plating base, the metal microstructures are laid bare by dissolution of the mold material (Figure 10.39d). Finally, using argon sputtering or chemical etching, the plating base between the metal microstructures is removed so that the metal parts do not short circuit (Figure 10.39e). In the case of chemical etching, a very fast etch should be used to avoid etching of the plating base from underneath the electrodeposited structures.[5,50]

Comparison of Master Micromold Fabrication Methods

The high cost of x-ray lithography has caused many miniaturization engineers to search for alternate means for fabricating high-aspect-ratio metal, silicon, or polymer micromasters, or mold inserts.

Polymer micromold inserts are widely used in lost mold micromolding of metals and ceramics where the molds are removed by dissolving or thermal degradation after molding. However, this is not always the case; certain polymer micromold inserts, such as elastomeric mold inserts (e.g., polydimethylsiloxane), can be used repeatedly in soft lithography (see Chapter 2). The polymer mold inserts used for lost micromolding can be fabricated using photolithography, microstereolithography, and so on.

Most of the micromold inserts in micromolding are metal micromold inserts. Nickel is most commonly used because of the well-known nickel electroplating process and its low internal stress. However, the hardness of nickel is relatively low compared with metals such as iron and stainless steel, resulting in a limited lifetime for nickel mold inserts. Other materials such as nickel-iron or tungsten-cobalt alloys with higher hardness are considered for longer-lived metal micromold inserts. Metal micromold inserts (or micromasters) can be

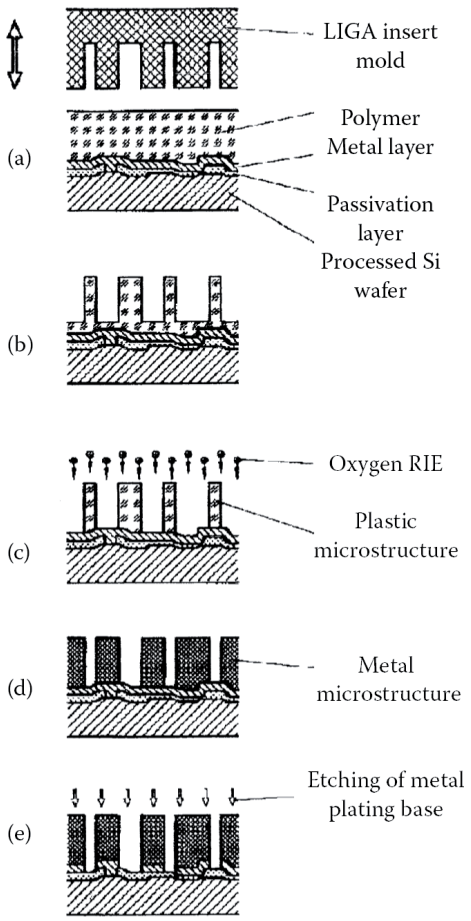


FIGURE 10.39 Fabrication of microstructures on a processed Si wafer. (a) Patterned isolation layer, conductive plating base, and in situ polymerization of PMMA. (b) Impression molding using a LIGA primary insert mold. (c) Removal of remaining plastic from the plating base by a highly directional oxygen etch. (d) Electrodeposition of metal shape. (e) Removal of plating base from between the metal structures by argon sputtering. (From Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure.* Weinheim, Germany: VCH Publishers.[67])

fabricated by a variety of alternate techniques, such as CNC machining, e-beam or ion-beam milling, precision EDM or excimer laser ablation of a piece of metal, thick deep UV resist photolithography, resist DRIE, silicon wet bulk micromachining, polymer excimer laser ablation, and polymer e-beam or ion-beam writing, all followed by electroplating (see Figure 10.1).

Both silicon wet and dry etching can be used for silicon mold insert fabrication. Silicon mold inserts have been applied to, for example, the hot embossing of polymer structures and for casting of ceramics. In Table 10.15, LIGA metal molds are compared with metal masters fabricated by other means. For example, comparing metal mold inserts made by spark erosive cutting and x-ray lithography, the latter proves far superior.[83] LIGA PMMA features as small as 0.1 μm are replicated in the metal shape with almost no defects. The electroformed structures have a superior surface quality with a surface roughness, R_{max}, of less than 0.02 μm.[28]

DRIE and thick deep UV-sensitive resists (DUV) such as polyimides, AZ-4000, and SU-8 are recent contenders for micromaster mold fabrication. With respect to dry etching, higher and higher aspect-ratio features are being achieved; especially when using highly anisotropic etching conditions as in cryogenic DRIE and in the Bosch process, remarkable results are obtained. Wall roughness, causing form locking, remains a problem with DRIE; the dry etching process was optimized for speed, not for demolding. For small-quantity production, where the lifetime of mold inserts is not crucial, a silicon

TABLE 10.15 Comparison of Micromolds

Parameters	LIGA	DUV	DRIE	LASER	CNC	EDM
Aspect ratio	100	22	10–25	<10	14 (hole drilling)	Up to 100
Wall roughness	<20 nm	~1 μm	~2 μm	1 μm–100 nm	Several micrometers	0.3–1 μm
Accuracy	<1 μm	2–3 μm	<1 μm	A few micrometers	See Figure 6.3	Some micrometers
Mask needed?	Yes	Yes	Yes	No	No	No
Maximum height	A few 100 μm up to 1 cm	A few 100 μm	A few 100 μm	A few 100 μm	Unlimited	Micrometers to millimeters

Source: Weber, L., W. Ehrfeld, H. Freimuth, M. Lacher, H. Lehr, and B. Pech. *SPIE: micromachining and microfabrication process technology II.* Austin, TX: SPIE. With permission.[87]

wafer etched by DRIE can be used directly as a mold insert for anywhere from 5–30 molding cycles.[88,89] Figure 10.40a shows such a Si mold for building a two-point calibration fluidic device on a compact disc (CD). Wet etching of Si leads to much smoother surfaces than DRIE (see, for example, the wet anisotropically etched Si nozzle in Figure 4.45b and compare that with the sidewall scalloping caused by Si DRIE in Figure 3.38) and is therefore the preferred method for making master molds out of Si. For much longer-lasting molds, metallizing the Si structure and using the metal as mold is preferred. Photoresist structures on a silicon substrate have also been tested as mold insert in plastic molding because of the simplicity and low cost of the process. Figure 10.40b shows an SU-8 photoresist mold for the same lab CD platform shown in Figure 10.40a. In low-pressure molding processes, such mold inserts work for a limited number of runs (applying a thin metal layer over the top of the resist may further extend the lifetime of the mold), but their applicability in high-pressure processes needs to be further verified. A better approach is to use DUV-photosensitive resists for electroplating to yield a metal tool, usually nickel or nickel-cobalt. Figure 10.40c shows a nickel mold insert made using SU-8 in our lab for the two-point calibration microfluidic platform. Both DUV and DRIE are more accessible than LIGA and will continue to improve, taking more opportunities away from LIGA. Like LIGA, both alternative techniques

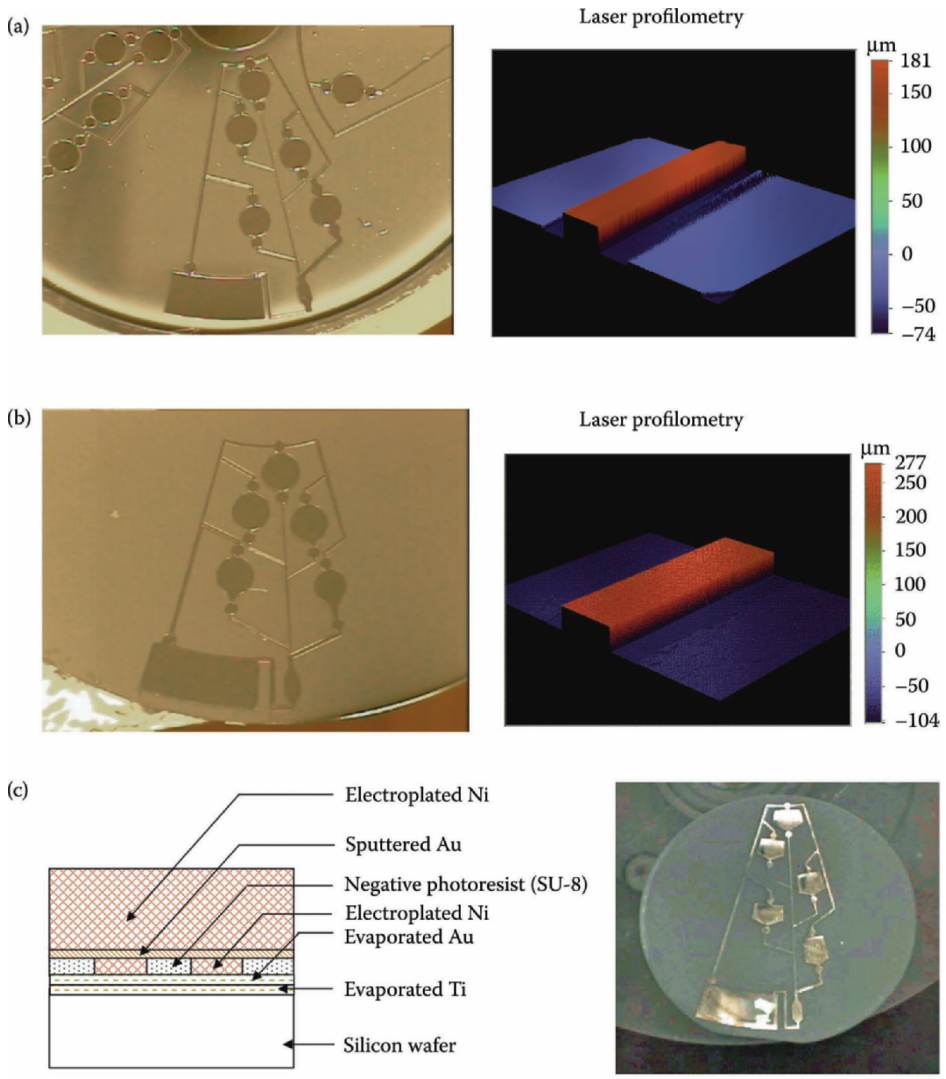


FIGURE 10.40 (a) Silicon mold insert made by UV photolithography and DRIE at Burstein Technologies/UCLA/OSU. (b) Photoresist (SU-8) mold insert made by UV lithography at Ohio State University by the author. (c) Nickel mold insert made by UV photolithography and electroplating (OSU). The depth of the etched channel is measured by a laser profilometer.

can be coupled with plating, but neither technique can yet achieve the extremely low surface roughness and vertical walls of LIGA.

Other competing technologies for making metal masters are laser ablation methods and ultraprecision CNC machining. The latter two methods are serial processes and rather slow, but because we are considering the production of a master only these technologies are competitive for certain applications.

Laser microablation produces minimum features of about 10 µm width and aspect ratios of 1/10. Challenges include taper and surface finish control. Recast layers around the laser-drilled features cause form locking and infidelity in the replication. Femtosecond pulse lasers promise thinner or even the absence of recast layers and excellent resolution and should be investigated further for the manufacture of micromold inserts.[90]

For large features (>50 µm) with tolerances and repeatability in the range of about 10 µm, traditional CNC machining of materials like tool steel and stainless steel are often accurate enough for making metal mold inserts. The advantage of this technique is that the tool materials used are the same as those in conventional polymer molding; therefore, their design, strength, and service life are well established. Complicated 3D structures can also be machined easily. The main drawbacks are that it is difficult to make sharp corners or right angles, and the surface quality is usually poor (surface roughness is around several micrometers).[8] In contrast, lithographic methods can produce molds with excellent surface quality (surface quality <0.1 µm) and sharp corners or right angles. However, they cannot be used on conventional tool materials like steel. Diamond-based micromilling and microdrilling[91] reduce the surface roughness to 1 µm or less.[92] Although diamond-based machining methods can achieve features smaller than 10 µm, they are only applicable to "soft" metals such as nickel, aluminum, and copper.

From Table 10.15, a significant potential application of LIGA remains the fabrication of those metal molds that cannot be accomplished with other techniques because of the tight wall roughness tolerances,s small size, and high aspect ratios. From the table it is obvious that LIGA micromolds excel both in very low surface roughness and excellent accuracy.

Summarizing, the requirements for an optimal micromold insert fabrication technique are[93]:

- The master has to be removed from the molded structure, so the ease of release through wall inclination control is crucial (undercuts, for example, cannot be tolerated as they cause form locking)
- The most important parameter, including master lifetime and achievable aspect ratios, depends very strongly on the surface quality of the master
- The interface chemistry between master and polymer is a critical factor and must be controlled

Alternative Molding Materials in LIGA and Pseudo-LIGA

In this last section we discuss alternative molding materials for LIGA and pseudo-LIGA. Besides PMMA and POM (see above), used in a commercially available form, semicrystalline polyvinylidene fluoride (PVDF), a piezoelectric material, has been used to make polymeric microstructures.[69] The optimum molding temperature of PVDF is 180°C, and PVDF structures can be molded without using mold release agents. Fluorinated polymers such as PVDF will enable higher temperature applications than PMMA. PC, well known for molding CDs, is under development for LIGA and LIGA-like processes and should perform well.[51]

Other materials that have the ability to flow or be sintered, such as glasses and ceramics, can be incorporated in the LIGA and pseudo-LIGA processes as well. In the case of ceramics, for example, plastic microstructures are used as disposable or lost molds. The molds are filled with slurry in a slip-casting process. Before sintering at high temperatures, the plastic mold and the organic slurry components are removed completely in a burnout process at lower temperatures. First results have been obtained for zirconium oxide and aluminum oxide. An important application for ceramic microstructures is

the fabrication of arrays of piezoceramic columns embedded in a plastic matrix.

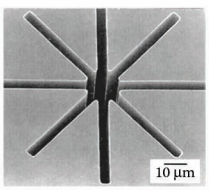
An alternative method to make LIGA and pseudo-LIGA ceramic or glass structures is based on sol-gel technology (see Chapter 8). This technique involves relatively low temperatures and may enable LIGA and pseudo-LIGA products such as glass capillaries for gas chromatography, piezoelectric arrays, and even high-T_C superconductor actuators. Sol-gel techniques should work well with LIGA and pseudo-LIGA-type molds if they are filled under a vacuum to eliminate trapped gas bubbles. In the sol-gel technique, a solution is spun on a substrate that is then given an initial firing at around 200°C, driving off the solvent in the film. Subsequently, the substrate is given a high-temperature firing at 800–900°C to drive out the remaining solvents and crystallize the film (see Figure 8.2). A major issue is the large shrinkage of sol-gel films during the initial firing. For example, the maximum thickness of high-T_C superconductor films currently achievable measures approximately 2 µm as a result of the high stress in the film caused during shrinkage. The latter is related to the ceramic yield or the amount of ceramic in the sol-gel compared with the amount of solvent. The sol-gel technique, in general, suits the production of thicker films well, as long as high-ceramic-yield sol-gels are used. Because LIGA-style films tend to be thick, a reduction in the amount of solvents to allow a high ceramic yield after firing would be to make sol-gel compatible with LIGA. This would require some development. Combining sol-gel technology with LIGA and pseudo-LIGA, lead-zirconium-titanate (PZT) piezoelectric devices can be made. In this case it is better to use the metal-plated structure rather than the direct PMMA template for the creation of the device because PZT must be processed in several elevated temperature steps. PZT sol-gel contains no particles that would prevent flow into small channels in a plated mold. The sol-gel contains high-molecular-weight polymer chains that hold the constituent metal salts, which are further processed into the final ceramic film. Thus, the sol-gel could be formed into the mold in a process much like reaction injection molding (RIM), with the mold filled with the chemistry under vacuum. Because the performance of piezoelectric actuator arrays is linked to the height and width ratio of the individual ceramic columns, as well as the distance between the columns in the array, LIGA, with its tremendous capacity for tall, dense, and high-aspect-ratio features, is an ideal fabrication tool.[94,95]

Examples

Example 10.1: LIGA Spinneret Nozzles

Profiled capillaries (nozzles) in a spinneret plate for spinning synthetic fibers from a molten or dissolved polymer, as shown in Figure 10.41, normally are produced by micro-EDM (see Chapter 5). This process establishes a practical lower limit of 20–50 µm for the minimal characteristic dimension, and the method cannot satisfy the requirements to produce complexly shaped, high-aspect-ratio spinnerets at a low cost. Smaller, more precise, and high-aspect-ratio nozzles can be produced by LIGA.[96]

Spinneret nozzles lend themselves to LIGA from the technical point of view. Compared with fabrication by micro-EDM, the minimum characteristic dimensions with LIGA can be reduced by an order of magnitude, and a high capillary length with excellent surface finish can easily be obtained. Moreover, the LIGA process makes all the nozzles in parallel, whereas micro-EDM operates as a serial technique. The market might focus on niche applications such as specialty, multilumen catheters. Also, medical use



(a)

(b)

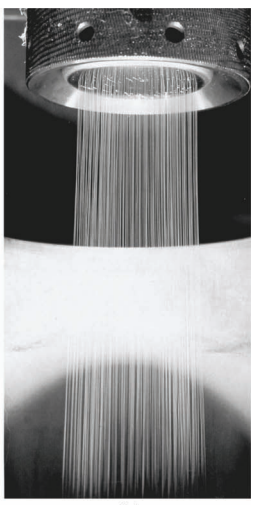

(c)

FIGURE 10.41 LIGA spinneret nozzles. (a) Spinneret plate. (b) Profiled spinneret nozzles. (c) Spinning synthetic fiber. (Courtesy of IMM, Germany.)

of atomizers for dispensing drugs is projected to increase dramatically in the coming years; this might cause a run for a wide variety of precise, inexpensive micronozzles.

Shew et al.[97] used a conformal mask to fabricate a LIGA die with capillaries 2 mm deep and 70 μm wide for the mass production of electroplated spinnerets for the spinning of polyester fibers. The conformal mask eliminates the need of alignment steps accompanying the multiple exposure processes required in the deep x-ray lithography used by these authors. As illustrated in Figure 10.42A, a thin sputtered copper layer covers a thick PMMA layer on an Al substrate (a). A resist layer (JSR 137N, a negative photoresist from Japan Synthesis Rubber Co.) is then patterned on the copper layer by UV lithography, followed by plating gold absorbers (b). After stripping the resist and the Cu layer (c), the conformal mask is ready for multiple x-ray exposure and development steps (d) and die electroforming (e). The die (f) is made by electroforming of a NiCo alloy on the Al substrate from a Ni/Co sulfamate bath at current densities between 1 and 5 A/dm^2. The addition of increasing amounts of cobalt sulfamate increases the hardness of the deposit to about 420 Vickers. At a concentration higher than 20 g/L cobalt sulfamate, the hardness saturates. With increasing hardness also comes increased residual stress, which needs to be carefully controlled. Polypropylene (PP) was used to duplicate the NiCo microstructures by injection molding (IM) through a stainless steel tool as shown in Figure 10.42B. After removing sthe LIGA die, the result is a field of polymer microstructures standing up from the stainless steel tool. The stainless steel substrate is subsequently used for plating the spinneret structures.

Polyester fibers typically melt at 260°C and are extruded at speeds of more than 3,000 m/min. During this fiber extrusion, the spinneret suffers high-temperature excursions of up to 280°C, and high-temperature wear resistance becomes a problem when using a NiCo alloy. To improve the high-temperature hardness of Ni spinnerets, Shew et al.[97] plated a Ni/SiC composite. With the addition of SiC nanopowder to the sulfamate plating electrolyte, stable composite hardness of 500 Hv (Vickers hardness) at temperatures as high as 500°C was demonstrated. The authors suggest that these new types of spinnerets will enable the production of a new generation of fibers with ultrafine sizes and new functionalities at a low cost.

Example 10.2: Inclined LIGA Walls

Using LIGA, microstructures as high as hundreds or even thousands of micrometers can be achieved with an aspect ratio of 200, lateral

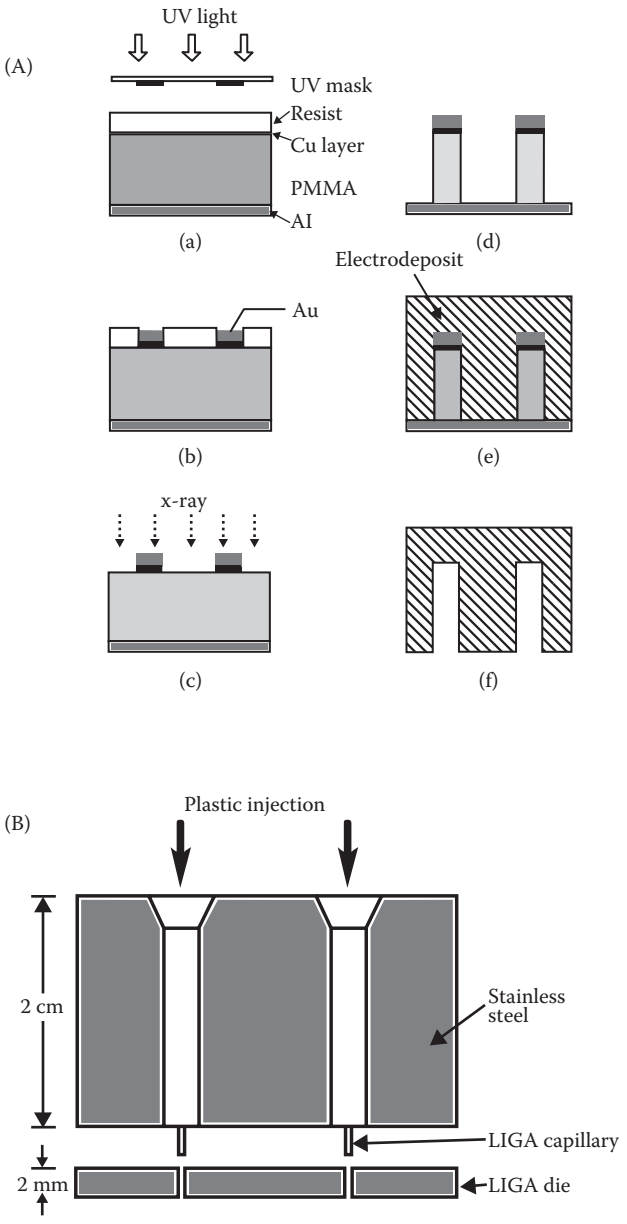


FIGURE 10.42 (A) Schematic diagram of the modified LIGA process (dimensions in this figure not to scale). (B) Replication process of the LIGA spinneret. (From Cheng, Y., B.-Y. Shew, C.-Y. Lin, D.-H. Wei, and M. K. Chyu. 1999. Ultra-deep LIGA process. *J Micromech Microeng* 9:58–63.[49])

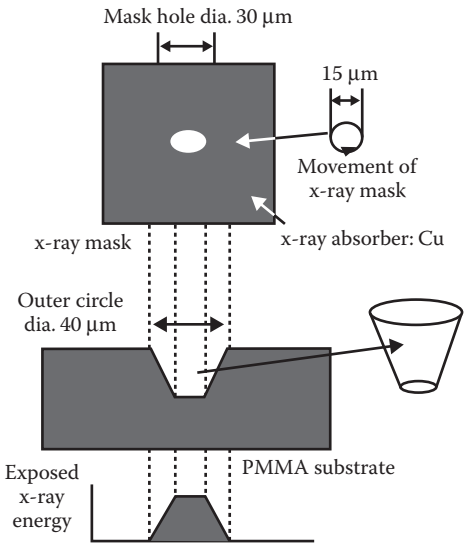


FIGURE 10.43 Principle of M²DXL. (Based on Tabata, O., H. You, H. Shiraishi, H. Nakanishi, T. Nishimoto, K. Yamamoto, and Y. Baba. 2000. *Micro total analysis systems 2000: proceedings of the μTAS 2000 symposium.* Enschede, the Netherlands: Kluwer Academic.[71])

dimensions smaller than 0.5 μm, accuracy of 0.1 μm, and wall perpendicularity of 90°. For many applications, inclined sidewalls would be a very useful addition to the LIGA tool box; for example, for making lenses or for easier release of molded parts. This second LIGA example concerns both these important applications. When using LIGA for plastic molding, such as through hot embossing, it is important to be able to control the resist wall inclination to facilitate the release of molded devices. To achieve this, one could make an inclined absorber or, following Tabata et al.,[71] one can also move the mask during exposure in M²DXL. The movement of the mask is realized by a precision *x-y* stage with a 10-nm step resolution. The principle of the technique is illustrated in Figure 10.43. Here, the x-ray mask with a circular hole of 30 μm is moved in a circle with a 15-μm diameter. The x-ray dose distributes according to the movement of the mask during exposure. As a result, the energy delivered to the PMMA photoresist shows a high constant value within the inner circle of the trajectory of the hole's movement and decreases with increasing trajectory diameter, becoming zero at the outside region. A truncated conical PMMA structure results with an inclination of the wall controlled by the diameters of the mask hole and its circular movement and the total x-ray dose.

For making inclined microfluidic channels whose depth is 40 μm, a mask with a channel width of 50 μm and mask movement of 10 μm was used. A good inclination for the sidewalls of microfluidic devices for easy release of molded parts is 80°. Tabata et al.[71] demonstrated that their technique may also be used to make microlens arrays. In the latter case, the mask hole diameter was 210 μm, and the mask movement 100 μm.

Example 10.3: LIGA Fiber-Chip Coupling

LIGA elements for coupling monomode fibers with integrated optical chips have been developed. These prealignment arrays may use fixed nickel guiding structures in combination with

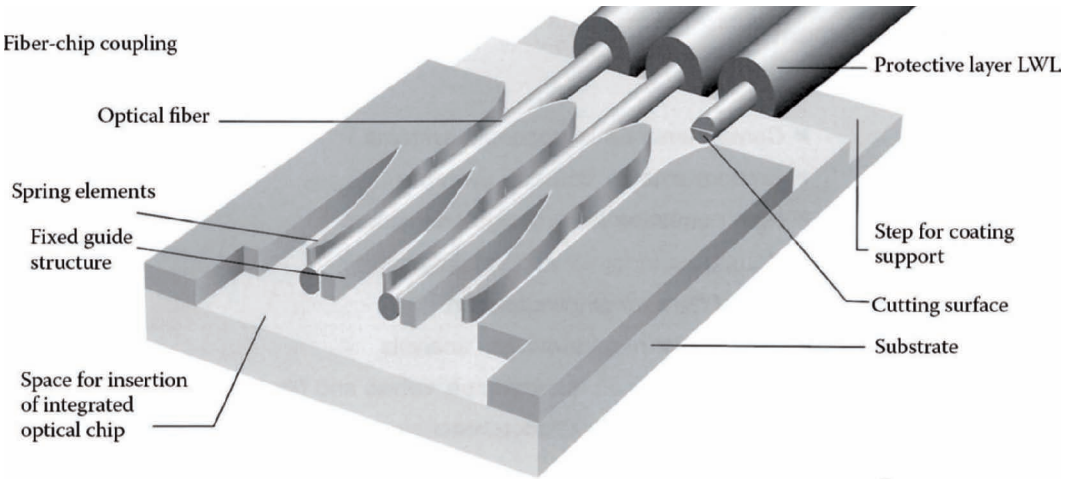


FIGURE 10.44 Coupling for monomode fibers with nickel spring elements. (Courtesy of IMM, Germany, http://www.imm-mainz.de/.)

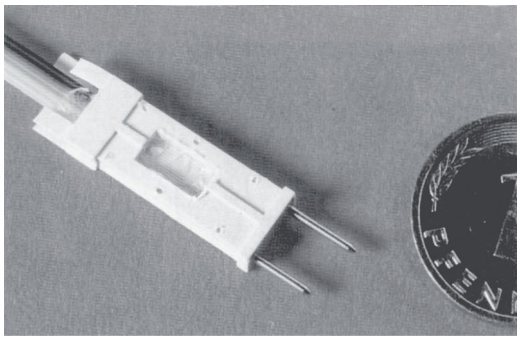


FIGURE 10.45 Pull-push connector for multimode fiber. (Courtesy of IMM, Germany, http://www.imm-mainz.de/.)

leaf springs to ensure a precise alignment of the optical fibers relative to the optical chip with accuracy in the submicrometer range, as shown in Figure 10.44. The thermal expansion coefficient of the substrate material is matched to the optical chip material—for example, a glass substrate is used for coupling fibers to glass chips—and the use of spring elements simplifies the handling of the fibers.[78] Alternatives involve the use of optical adhesives and silicon V-groove arrays (see Chapter 4). Both alternative technologies are more labor intensive and suffer from thermal expansion problems.

The use of microoptical LIGA components is especially attractive for the coupling of multimode fibers; here, the capability of exact lithographic positioning of mechanical mounting supports is used advantageously to position the multimode fibers very precisely with respect to the microoptical components without need for any additional adjusting operations.[73,98]

A pull-push LIGA connector for single-mode fiber ribbons, shown in Figure 10.45, incorporates a set of precision microsprings coupling up to 12 fibers spaced at 250 μm. It is now on the market and is one of the LIGA products with mass market appeal. To obtain good coupling efficiency, the fibers are positioned horizontally with a precision of 1 μm. LIGA is an excellent method to provide that type of precision. These connectors were developed at IMM.[99] In line with the micromanufacturing philosophy presented in this book (i.e., to optimize the use of each micromachining technique for optimal cost/performance application), those parts of the mold insert that do not require such high accuracy are fabricated using other methods of precision machining such as EDM (see Chapter 5).

Acknowledgments

Special thanks to Dr. Nahui Kim, formerly at UCI, now at Samsung.

Questions

10.1: Describe why a synchrotron is the best tool for producing the x-rays required to make the most accurate LIGA parts and make a table comparing x-ray lithography with UV lithography.

10.2: An electroplating solution containing nickel sulphamate is used to electroplate nickel LIGA structures. The formula incorporates 450 g of nickel sulphamate [$Ni(NH_2SO_3)2$] per liter of water (plus some additives). The solution is heated to 50°C. Assuming the plating is occurring on a nickel cathode.

(a) Calculate the concentration of the Ni^{++} ions in the solution.

(b) Calculate the equilibrium voltage across the nickel-solution interface at the cathode prior to electroplating.

(c) If the plating is occurring inside high-aspect-ratio PMMA cavities with a depth of 500 micrometers, and the diffusivity of Ni^{++} ions in water is 5×10^{-6} cm^2/s, calculate the limiting current density (mA/cm^2) assuming the boundary layer thickness is governed by the height of the PMMA.

(d) Plot the current density versus potential curve at the cathode using the values provided from part (c).

(e) If the current density is specified at 75 mA/cm^2, what is the likely result (qualitatively)?

(f) If the pH of the solution is 4 and the hydrogen ion diffusivity is 3×10^{-5} cm^2/s, calculate the limiting current associated with the reaction $H^+ + e \rightarrow 0.5\ H_2$.

10.3: What is the difference between a thermoplastic, a thermoset, and an elastomer? Give three examples of each.

10.4: Assume a polymer is under a tensile load at a variety of temperatures from –20°C to its

breakdown temperature at 150°C. Outline the potential stress-strain diagrams that may be found for a thermoplastic polymer under very short loading times.

10.5: What is viscoelastic polymer? Plot the stress-strain curve for a viscoelastic polymer.

10.6: Consider a laboratory setup (Galvanic cell) with a bar of copper (Cu) and a bar of silver (Ag) immersed in an electrolyte solution. The following are the redox half reactions and electrochemical potentials. The Gibbs free energy = (–) number of electron exchanged × electrochemical potential × Faraday constant. For a reaction to spontaneously occur, the Gibbs free energy must be negative.

$$Ag^{+} + e^{-} \rightarrow Ag \qquad E° = 0.80\ V$$
$$Cu^{++} + 2e^{-} \rightarrow Cu \qquad E° = 0.34\ V$$

(a) Will Ag react with Cu++ (i.e., will we be able to plate copper on the silver bar)?

(b) Will Cu react with Ag+ (i.e., will we be able to plate silver on the copper bar)?

(c) If you place a sheet of copper into an $AgNO_3$ solution, what would happen? What is the reaction?

10.7: Electrodeposition:

(a) In an electrochemical cell, we deposit Zn from a Zn^{2+} solution. After the deposition, we monitor the voltage of the Zn deposit with respect to a reference electrode (in this case we are using a standard hydrogen electrode, i.e., SHE). First we measure the potential in a $ZnSO_4$ solution at 10^{-2}M and then in a ZnSO4 solution of 10^{-3}M. What will be the change in potential of the Zn electrode between those two solutions? What is the potentsial in the 10^{-2}M solution? (E_o vs. SHE for Zn is –0.763V.)

$$Zn^{2+} + 2e^{-} \rightarrow Zn$$

(b) Explain the difference between electroless deposition and electro-deposition.

10.8: What issues do traditional plating and polymer molding companies face when trying to plate and mold LIGA-type structures?

10.9: Compare the pros and cons for injection molding, reaction injection molding, and hot embossing in LIGA. What are the differences in a micro part made with LIGA vs. pseudo-LIGA? How would you find out which technology was used, LIGA or pseudo-LIGA?

10.10: Describe two MEMS applications in which the use of LIGA may be justified.

10.11: Fabricate an x-ray mask detailing all materials and processes required. The mask should be adequate for the fabrication of Au parts with an aspect ratio of 50 (height over width).

10.12: For the following designs, decide which standard MEMS process (CMOS, LIGA, MUMPs) you would use and give a brief justification.

- You want to build a very cheap impact sensor.
- You want to build a rotating disk for a microoptical stroboscope.
- You want to build a strong microgear train.
- You want to build an array of structures with different resonance frequencies to test new theories on viscous damping.

10.13: Describe some techniques that may be employed for the fabrication of MEMS structures with slanted walls.

References

1. Greener, J., and R. Wimberger-Friedl, eds. 2006. *Precision injection molding-process, materals, and applications*. Munich, Germany: Hanser.
2. IMM. 1995. Commercial brochure. IMM, http://www.imm-mainz.de/.
3. Lee, L. J., C.-H. Shih, Y.-J. Juang, J. Garcia, M. J. Madou, and K. W. Koelling. 1999. *Novel microfabrication options for BioMEMS conference*. San Francisco: The Knowledge Foundation.
4. Hanemann, T., R. Ruprecht, and J. H. Haubelt. 1998. Photomolding in microsystem technology. *Polym Preprints* 39:657–58.
5. Eicher, J., R. P. Peters, and A. Rogner. 1992. VDI-Verlag. Report no. *VDI-Bericht 960*, Dusseldorf, Germany.
6. Bifano, T. G., H. E. Fawcett, and P. A. Bierden. 1997. Precision manufacture of optical disc master stampers. *Precision Eng* 20:53–62.
7. Yu, L., Y.-J. Juang, K. W. Koelling, and L. J. Lee. 2000. *Proceedings: Society of Plastics Engineers ANTEC*. Orlando, FL.
8. Madou, M. J., Y. Lu, S. Lai, Y.-J. Juang, L. J. Lee, and S. Daunert. 2000. A novel design on a CD disc for two-point calibration measurement. *Proceedings of the IEEE: solid-state sensor and actuator workshop*. Hilton Head Island, SC.

9. Dunke, K., H.-D. Bauer, W. Ehrfeld, J. Hobfeld, L. Weber, G. Horcher, and G. Muller. 1998. Injection-molded fiber ribbon connectors for parallel optical links fabricated by the LIGA technique. *J Micromech Microeng* 8:301–6.
10. Fahrenberg, J., W. Bier, D. Mass, W. Menz, R. Ruprecht, and W. K. Schomburg. 1995. A microvalve system fabricated by thermoplasting molding. *J Micromech Microeng* 5:169–71.
11. Ramos, B. L., S. J. Choquette, and F. F. Nicholas. 1996. Embossable grating couplers for planar waveguide optical sensors. *Anal Chem* 68:1245–49.
12. Becker, H., and W. Dietz. 1998. Microfluidic device for μ-TAS applications for fabrication by polymer hot embossing. *Proc SPIE* 3515:177–82.
13. Kopp, M. U., H. J. Crabtree, and A. Manz. 1997. Developments in technology and applications in microsystems. *Curr Opin Chem Biol* 1:410–19.
14. Chou, S. Y., P. R. Krauss, and P. J. Renstrom. 1996. Imprint lithography with 25-nanometer resolution. *Science* 272:85–87.
15. Schift, H., R. W. Jaszewski, C. David, and J. Gobrecht. 1999. Nanostructuring of polymers and fabrication of integrated electrodes by hot embossing lithography. *Microelectron Eng* 46:121–24.
16. Jaszewski, R. W., H. Schift, J. Gobrecht, and P. Smith. 1998. Hot embossing in polymers as a direct way to pattern resist. *Microelectron Eng* 41/42:575–78.
17. Casey, B. G., W. Monaghan, and C. D. W. Wilkinson. 1997. Embossing of nanoscale features and environment. *Microelectron Eng* 35:393–96.
18. Gottschalch, F., T. Hoffman, C. M. Sotomayor Torres, H. Schulz, and H.-C. Scheer. 1999. Polymer issues in nanoimprinting technique. *Solid State Electron* 43:1079–83.
19. Jaszewski, R. W., H. Schift, P. Groning, and G. Margaritondo. 1997. Properties of thin anti-adhesive films used for the replication of microstructures in polymers. *Microelectron Eng* 35:381–84.
20. Jaszewski, R. W., H. Schift, B. Schnyder, A. Schneuwly, and P. Groning. 1999. The deposition of anti-adhesive ultrathin Teflon-like films and their interaction with polymers during hot embossing. *Appl Surface Sci* 143:301–8.
21. Becker, H., and U. Heim. 1999. Silicon as tool material for polymer hot embossing. *12th IEEE International Conference on MEMS '99: Micro Electro Mechanical Systems.* Orlando, FL.
22. Lin, L.-W., C.-J. Chiu, W. Bacher, and M. Heckele. 1996. Microfabrication using silicon mold inserts and hot embossing. *Seventh international symposium on micro machine and human science.* Nagoya, Japan.
23. Casey, B. G., D. R. S. Cumming, I. I. Khandaker, A. G. S. Curtis, and C. D. W. Wilkinson. 1999. Nanoscale embossing of polymers using a thermoplastic die. *Microelectron Eng* 46:125–28.
24. Ruprecht, R., W. Bacher, J. Haubelt, and V. Piotter. 1995. Injection molding of LIGA and LIGA-similar microstructures using filled and unfilled thermoplastics. *Proc SPIE* 2639:146–57.
25. Spiller, E., R. Feder, J. Topalian, E. Castellani, L. Romankiw, and M. Heritage. 1976. X-ray lithography for bubble devices. *Solid State Technol* April:62–68.
26. Romankiw, L. T., I. M. Croll, and M. Hatzakis. 1970. Batch-fabricated thin-film magnetic recording heads. *IEEE Trans Magn* 6:597–601.
27. Becker, E. W., W. Ehrfeld, D. Munchmeyer, H. Betz, A. Heuberger, S. Pongratz, W. Glashauser, H. J. Michel, and V. R. Siemens. 1982. Production of separation nozzle systems for uranium enrichment by a combination of x-ray lithography and galvanoplastics. *Naturwissenschaften* 69:520–23.
28. Hagmann, P., W. Ehrfeld, and H. Vollmer. 1987. Fabrication of microstructures with extreme structural heights by reaction injection molding. *First meeting of the European Polymer Federation, European symposium on polymeric materials.* Lyon, France.
29. Waldo, W. G., and A. W. Yanof. 1991. 0.25 micron imaging by SOR x-ray lithography. *Solid State Technol* 34:29–31.
30. Clemens, J., and R. Hill, ed. 1991. *JTEC Panel Report on X-ray Lithography in Japan*. PB92-100205, Springfield, VA:NTIS.
31. Urisu, T., and H. Kyuragi. 1987. Synchrotron radiation-excited chemical-vapor deposition and etching. *J Vac Sci Technol* B5:1436–40.
32. Goedtkindt, P., J. M. Salome, X. Artru, P. Dhez, N. Maene, F. Poortmans, and L. Wartski. 1991. X-ray lithography with a transition radiation source. *Microelectron Eng* 13:327–30.
33. Muray, J. J., and I. Brodie. 1991. *Report no*. 2019. Menlo Park, CA: SRI International.
34. Bley, P., D. Einfeld, W. Menz, and H. Schweickert. 1992. A dedicated synchrotron light source for micromechanics. *Proceedings of the EPAC 92*. Berlin, Germany.
35. Ehrfeld, W., W. Glashauer, D. Munchmeyer, and W. Schelb. 1986. Mask making for synchrotron radiation lithography. *Microelectron Eng* 5:463–70.
36. Lawes, R. A. 1989. Sub-micron lithography techniques. *Appl Surf Sci* 36:485–99.
37. Schomburg, W. K., H. J. Baving, and P. Bley. 1991. Ti- and Be-x-ray masks with alignment windows for the LIGA process. *Microelectron Eng* 13:323–26.
38. Hein, H., P. Bley, J. Gottert, and U. Klein. 1992. Elektronenstrahllithographie zur Herstellung von Rontgen-masken fur das LIGA-Verfahren. *Feinwtech Messtech* 100:387–89.
39. Bley, P., W. Menz, W. Bacher, K. Feit, M. Harmening, H. Hein, J. Mohr, W. K. Schomburg, and W. Stark. 1991. Application of the LIGA process in fabrication of three-dimensional mechanical microstructures. *Microprocess 91, 1991 International Microprocess Conference.* Kanazawa, Japan.
40. Guckel, H., T. R. Christenson, T. Earles, J. Klein, J. D. Zook, T. Ohnstein, and M. Karnowski. 1994. Laterally driven electromagnetic actuators. *Technical digest: 1994 solid state sensor and actuator workshop*. Hilton Head Island, SC.
41. Siddons, D. P., and E. D. Johnson. 1994. Precision machining using hard x-rays. *Synchrotron Radiation News* 7:16–18.
42. Guckel, H. 1994. Notes from handouts. The Commercialization of Microsystems. September 11–16, Banff Conference Center, Banff, Canada.
43. Becker, E. W., W. Ehrfeld, P. Hagmann, A. Maner, and D. Munchmeyer. 1986. Fabrication of microstructures with high aspect ratios and great structural heights by synchrotron radiation lithography, galvanoforming, and plastic molding (LIGA process). *Microelectron Eng* 4:35–56.
44. Friedrich, C. 1994. Notes from handouts. The Commercialization of Microsystems. September 11–16, Banff Conference Center, Banff, Canada.
45. Henck, R. 1984. Detecteurs au silicium pour electrons et rayons X, principes de fonctionnement, fabrication et performance. *J Microsc Spectrosc Electron* 9:131–33.

46. Vladimirsky, Y., O. Vladimirsky, V. Saile, K. Morris, and J. M. Klopf. 1995. *Microlithography '95 (Proceedings of the SPIE)*. Santa Clara, CA: SPIE.
47. Khan Malek, C., Y. Vladimirsky, O. Vladimirsky, J. Scott, B. Craft, and V. Saile. 1996. X-ray microfabrication activities at the center for advanced microstructures and devices (CAMD). *Rev Sci Instrum* 67:1–6.
48. Shih, W.-P., Y. Cheng, C.-Y. Lin, and G.-J. Hwang. Low-cost x-ray conformal mask using x-ray dry FILM resist. *Microelectron Eng* 40:43–50.
49. Cheng, Y., B.-Y. Shew, C.-Y. Lin, D.-H. Wei, and M. K. Chyu. 1999. Ultra-deep LIGA process. *J Micromech Microeng* 9:58–63.
50. Michel, A., R. Ruprecht, and W. Bacher. 1993. *Abformung von. mikrostrukturen auf prozessierten Wafern. KfK report no. 5171*. Karlsruhe, Germany: Kernforschungszentrum Karlsruhe.
51. Rogner, A., J. Eichner, D. Munchmeyer, R.-P. Peters, and J. Mohr. 1992. The LIGA technique: what are the new opportunities? *J Micromech Microeng* 2:133–40.
52. Mohr, J., W. Ehrfeld, and D. Munchmeyer. 1988. *Analyse der defectursachen und der genauigkeit der structuru-bertragung bei der rontgentiefenlithographie mit synchrotronstrahlung. KfK report no. 4414*. Karlsruhe, Germany: Kernforschungszentrum Karlsruhe.
53. Moreau, W. M. 1988. *Semiconductor lithography*. New York: Plenum Press.
54. Lingnau, J., R. Dammel, and J. Theis. 1989. Recent trends in x-ray resists: part I. *Solid State Technol* 32:105–12.
55. Mohr, J., W. Ehrfeld, D. Munchmeyer, and A. Stutz. 1989. Resist technology for deep-etch synchrotron radiation lithography. *Makromol Chem Macromol Symp* 24:231–51.
56. Gottert, J., J. Mohr, and C. Muller. 1991. Mikrooptische Komponenten aus PMMA, Hergestellt Durch Roentgentiefenlithographie Werkstoffe der Mikrotechnik-Bais fur neue Producte. *VDI Berichte* Düsseldorf, Germany: Deutscher Ingenieur-Verlag, pp. 249–63.
57. Wollersheim, O., H. Zumaque, J. Hormes, J. Langen, P. Hoessel, L. Haussling, and G. Hoffman. 1994. Radiation chemistry of poly(lactides) as new polymer resists for the LIGA process. *J Micromech Microeng* 4:84–93.
58. Madou, M. J., and M. Murphy. 1995. Unpublished results.
59. Mohr, J., W. Ehrfeld, and D. Munchmeyer. 1988. Requirements on resist layers in deep-etch synchrotron radiation lithography. *J Vac Sci Technol* B6:2264–67.
60. Ehrfeld, W. 1994. Notes from handouts. The Commercialization of Microsystems. September 11–16, Banff Conference Center, Banff, Canada.
61. Campbell, S. A. 1996. *The science and engineering of microelectronic fabrication*. New York: Oxford University Press.
62. Vladimirsky, O. 1995. Private communication with the author, LSU.
63. Galhotra, V., C. Marques, Y. Desta, K. Kelly, M. Despa, A. Pendse, and J. Collier. 1996. Fabrication of mold inserts using a modified LIGA procedure. *SPIE: micromachining and microfabrication process technology II*. Austin, TX.
64. Rogers, J., C. Marques, and K. Kelly. 1996. Cyanoacrylate bonding of thick resists for LIGA. *SPIE: microlithography and metrology in micromachining II*. Austin, TX.
65. Becker, E. W., W. Ehrfeld, and D. Munchmeyer. 1984. *Untersuchungen zur abbildungsgenauigkeit der rontgentiefenlithographie mit synchrotronstrahlung. KfK report no. 3732*. Karlsruhe, Germany: KfK. Kernforschungszentrum Karlsruhe.
66. Munchmeyer, D. 1984. PhD thesis. University of Karhsruhe, Germany.
67. Menz, W., and P. Bley. 1993. *Mikrosystemtechnik fur Ingenieure*. Weinheim, Germany: VCH Publishers.
68. Editorial. 1994. Commercial Brochure.
69. Harmening, M., W. Bacher, P. Bley, A. El-Kholi, H. Kalb, B. Kowanz, W. Menz, A. Michel, and J. Mohr. 1992. *Proceedings: IEEE micro electro mechanical systems (MEMS '92)*. Travemunde, Germany: IEEE.
70. Mohr, J., C. Burbaum, P. Bley, W. Menz, and U. Wallrabe. 1990. *Microsystem technologies 90*. Ed. H. Reichl. Berlin: Springer.
71. Tabata, O., H. You, H. Shiraishi, H. Nakanishi, T. Nishimoto, K. Yamamoto, and Y. Baba. 2000. *Micro total analysis systems 2000: proceedings of the μTAS 2000 symposium*. Enschede, the Netherlands: Kluwer Academic.
72. Bley, P., J. Gottert, M. Harmening, M. Himmelhaus, W. Menz, J. Mohr, C. Muller, and U. Wallrabe. 1991. *Microsystem technologies '91*. Eds. R. Krahn and H. Reichl. Berlin: VDE-Verlag.
73. Gottert, J., J. Mohr, and C. Muller. 1992. *Integrierte Optik und Mikrooptik mit Polymeren*. Mainz, Germany: Max-Planck-Institut für Polymerforschung.
74. Ghica, V., and W. Glashauser. 1982. Verfahren fur die Spannungsrissfreie Entwicklung von Bestrahlthen Polymethylmethacrylate-Schichten, in Deutsche Offenlegungsschrift. Patent No. 3039110, Germany.
75. Maner, A., S. Harsch, and W. Ehrfeld. 1988. Mass production of microdevices with extreme aspect ratios by electroforming. *Plating Surface Finish* 75:60–65.
76. Harsch, S., W. Ehrfeld, and A. Maner. 1988. *Untersuchungen zur herstellung von mikrostrukturen grosser strukturhöhe durch galvanoformung im nickelsilfamatelektrolyten. KfK report no*. 4455. Karlsruhe, Germany: Kernforschungszentrum Karlsruhe.
77. Mohr, J., and M. Strohmann. 1992. Examination of long-term stability of metallic LIGA microstructures by electromagnetic activation. *J Micromech Microeng* 2:193–95.
78. Romankiw, L. T., and T. A. Palumbo. 1987. *Proceedings: symposium on electrodeposition technology, theory and practice*. San Diego.
79. Masuko, N., T. Osaka, and Y. Ito, eds. 1996. *Electrochemical technology: innovation and new developments*. Tokyo and Amsterdam: Kodansha and Gordon and Breach Science Publishers, copublishers.
80. Dukovic, J. O. 1993. Feature-scale simulation of resist patterned electrodeposition. *IBM J Res Dev* 37:125–40.
81. Leyendecker, K., W. Bacher, W. Stark, and A. Thommes. 1994. New microelectrodes for the investigation of the electroforming of LIGA microstructures. *Electrochim Acta* 39:1139–43.
82. Sales Brochure, 1994. RK Kissler, Speyer, Germany.
83. Hagmann, P., and W. Ehrfeld. 1988. Fabrication of microstructures of extreme structural heights by reaction injection molding. *J Polymer Process Soc* 4:188–95.
84. Burbaum, C., J. Mohr, P. Bley, and W. Ehrfeld. 1991. Fabrication of capacitive acceleration sensors by the LIGA technique. *Sensors Actuators A* A25:559–63.
85. Guckel, H., K. J. Skrobis, T. R. Christenson, J. Klein, S. Han, B. Choi, and E. G. Lovell. 1991. *Proceedings: IEEE micro electro mechanical systems (MEMS '91)*. Nara, Japan.
86. Harmening, M., W. Bacher, and W. Menz. 1991. Molding plateable micropatterns of electrically insulating and electrically conducting poly(methyl methacrylate)s by the LIGA technique. *Makromol Chem Macromol Symp* 50:277–84.

87. Weber, L., W. Ehrfeld, H. Freimuth, M. Lacher, H. Lehr, and B. Pech. SPIE: *micromachining and microfabrication process technology II*. Austin, TX.
88. Madou, M. 1997. *Fundamentals of microfabrication*. Boca Raton, FL: CRC Press.
89. Wimberger-Friedl, R. 1999. Injection molding of sub-µm grating optical elements. *Proc SPE ANTEC* 57:476–80.
90. Momma, C., S. Nolte, N. Chichkov, B. V. Alvensleben, and F. A. Tunermann. 1997. Precise laser ablation with ultrashort pulses. *Appl Surf Sci* 109/110:15–19.
91. Warrington, R. O. 1999. *Novel microfabrication options for BioMEMS conference (proceedings)*. San Francisco: The Knowledge Foundation.
92. Roberts, M. A., J. S. Rossier, P. Bercier, and H. Girault. 1997. UV laser machined polymer substrates for the development of microdiagnostic systems. *Anal Chem* 69:2035–42.
93. Becker, H., and C. Gärtner. 2000. Polymer microfabrication methods for microfluidic analytical applications. *Electrophoresis* 21:12–26.
94. Preu, G., A. Wolff, D. Cramer, and U. Bast. 1991. *Proceedings: second European ceramic society conference (2nd ECerS '91)*. Augsburg, Germany: European Ceramic Society.
95. Lubitz, K. 1989. *Mikrostrukturierung von Piezokeramik*. Düsseldorf, Germany: VDI-Tagungsbericht.
96. Maner, A., S. Harsch, and W. Ehrfeld. 1987. *Proceedings: 74th AESF annual technical conference (SUR/FIN '87)*. Chicago: AESF.
97. Shew, B.-Y., Y. Cheng, C.-H. Lin, W.-P. Ma, G.-J. Huang, C.-L. Kuo, S.-C. Tseng, D.-S. Lee, and G.-L. Chang. 1999. Manufacturing process for LIGA spinnerets. *Sensors Materials* 11:329–37.
98. Gottert, J., J. Mohr, and C. Muller. 1992. *Proceedings: micro system technologies '92*. Berlin: VDE-Verlag.
99. Editorial. 1994. Commercial brochure. http://www.imm-mainz.de/.

Index

A

Abrasive gas jet machining (AJM)
  abrasive powders, 375–376
  accelerometer in glass, 376–377
  erosion efficiency, 376
  erosion rate, 376
  material characteristics, 376
  microerosion process, 375–376
  powder blasting, 375–377
Abrasive machining
  categories of, 365
  conventional synthetic abrasives, 365
  electrolytic in-process dressing, nanofinishing of, 367–368
  grinding, 366–367
  jet machining-powder blasting, 375–377
  Knoop Hardness for, 366
  ultrasonic machining, 368–372
  water jet machining, 372–375
Abrasive water jet (AWJ) machining, 374–375
Additive (forming) processes, 386–389
Adhesion, of thin films, 445–447
Aerogel, 513–514
Aerosol decomposition synthesis, *see* Spray pyrolysis
Aerosol deposition (AD)
  characteristics, 412
  piezoelectrics in, 412–414
  schematic illustration, 412
  zirconate ceramics, 413
Aluminum oxide, 365
Analog devices accelerometer
  ADXL-50, 496–497
  IC chip preparation, for polysilicon, 498
  mechanical structure, 497
  polysilicon deposition and IC metallization, 498
  prerelease preparation and release, 499
Analog optical modulator, 470
Angle-dependent sputter rate, 175–176
Anisotropic etching
  ammonium hydroxide-water/tetramethyl ammonium hydroxide-water, 254
  Arrhenius plots, 255
  description, 250
  disadvantages, 251
  doping dependence
    etch stop region, 273
    mechanisms, 272
    moderate dopant concentration, 272–273
  etchants variety, 251
  ethylenediamine pyrocatechol, 253–254
  hydrazine (H_2N_4), 254–255
  masking for, 256–257
  potassium hydroxide, 252–253
  principal characteristics, 251–252
  procedure, 251
  semiconductor bias
    anodic dissolution, 280–281
    cyclic voltammogram, 280
    electroless etching, 280
    electropolishing, 281–282
    silicon, 282–296
  silicon wafer protection, 257
  Si surface roughness, 255–256
Anisotropy
  dopant-driven, 189–190
  energy-driven, 185–188
  inhibitor-driven, 188–189
Anodization, 556
Apertureless near-field scanning optical microscopy (ANSOM), 135–136
Aqueous development, 25
Arc-free plasma etching, 184–185
Arsine-doped epitaxial silicon, 427
Atmospheric plasma etching
  arc-free plasma etching, 184–185
  downstream plasma, 183–184
  voltage current plot, 182–183
Atmospheric pressure CVD (APCVD)
  carbon nanotubes, 428–429
  cold wall reactor type, 425
  continuous-feed, 427
  epitaxial Si, 425–427
  tetraethylorthosilicate (TEOS), with oxide, 427–428
Atomic layer deposition (ALD)
  Al_2O_3 process, 436–437
  applications, 435
  comparison of, 438
  flow of precursor, 436–437
  growth cycles, 435
  purge cycle, 436–437
  setup, 436
Atomistic model, Gosalvez–Kelly
  chemical oxidation, 267
  electrochemical mechanisms, 266–267
  nucleophilic attack representation, 265–266
  origin, 265
  Si etching process, 267
  surface microstructure, 268
  surface morphology, 267–268
Atomization, 568
Atom lithography/mechanosynthesis
  assembly sequence and schematic for, 134
  Au chains, STM topographic images, 132–133
  manipulation, for nanomatching, 132
5-Axis-HP-Milling Center Gamma High Performance 303 by WISSNER GmbH (Göttingen), 363

B

Backscattering, 177
Batch spray development, 25
Bis(aryl)azide rubber resists, 28–29
Black silicon, 181–182
Block copolymer lithography
  applications, 121–122
  diblock copolymer phase diagram, 118
  flash memory, 120
  investigation, 117
  patterned wettability, 120–121
  vs. traditional lithography, 119
Blocking mechanism, *see* Inhibitor-driven anisotropy
Blood gas sensor, excimer laser manufacture, 345
Boron nitride, 365
Borophosphosilicate glass (BPSG), 469
Bosch process, 197
Bulk matter, 405
Bulk micromachining, 392

C

Carbon nanotubes, APCVD, 428–429
Cathode dark space, 161
Charged-particle beam lithography
  DIB, 111–112
  direct write e-beam lithography, 104–105
  EBL, 104
  electron-and ion-beam applications, 103
  electron-beam resists, 106–107
  electron emission sources, 105–106
  electron projection lithography, 107–109
  FIB, 111
  IPL, 112–113
  maximum dose, 103
  micromachined electron emission sources, 109–111

Chemical forming, 510
organic layers, in bio-MEMS
glow discharge (plasma) polymerization, 518–520
Langmuir-Blodgett deposition and self-assembled monolayers, 523–524
organic film spin coating, 515
polymer casting, 517
polymer dip coating, 516
polymer doctor's blade/knife coating, 517–518
polymer dry film lamination, 516
polymer spraying, 516–517
silk screening/screen printing, 520–522
organic materials, patterning of
conductive polymer patterning, 534–535
digital mirror array patterning, 528–529
ink-jet printing, 528–532
mechanical microspotting, 532–534
microcontact printing, 534
photolithography, of hydrogels, gas-permeable membranes and ISEs, 524–525
proteins patterning, with lithography, 527–528
VLSIPS, 525–527
sol-gel technique
aerogel, 513–514
definitions, 512
gelation, 513
nanoparticle products and manufacture, 512
processes and products in, 513
silica sol formation, 514
xerogel, 514
Chemically assisted ion-beam etching (CAIBE), 185–186
Chemical machining, 149
Chemical-mechanical polishing (CMP) process, 70
Chemical milling, 218–219
Chemical vapor deposition (CVD), 69
advantages, 416
applications, 415
atmospheric pressure CVD (APCVD)
carbon nanotubes, 428–429
cold wall reactor type, 425
continuous-feed, 427
epitaxial Si, 425–427
tetraethylorthosilicate (TEOS), with oxide, 427–428
atomic layer deposition (ALD)
Al_2O_3 process, 436–437
applications, 435
comparison of, 438
flow of precursor, 436–437
growth cycles, 435
purge cycle, 436–437
setup, 436
cluster tools, 440
disadvantages, 416
epitaxy
epilayer thickness, 440
liquid phase epitaxy (LPE), 439
selective epi, 439–440
silicon, 438
solid-phase epitaxy, 439
equipment, 440
high-density plasma CVD (HDPCVD), 433
low-pressure CVD (LPCVD)
hot wall reactor, 429
polysilicon surface micromachining, 429–430
metalorganic CVD (MOCVD), 433–434
plasma-enhanced CVD (PECVD)
polysilicon, 432–433
reactor configurations, 431
silicon nitride, 432
vs. PVD method, 416
reaction mechanisms and reactor
atmospheric pressure CVD (APCVD), 421
boundary layer development, in gas phase, 418
configurations, 419
high-density plasma CVD (HDPCVD), 418
laminar development length, 419
low-pressure CVD (LPCVD), 421
mass transport, 420
material flux, 418–419
photon-and laser-assisted, 418
Reynolds number, 418
schematic of, 417
silicon growth rate, 420
temperature, 421
thermal entrance length, 419
transport and reaction processes, 416–417
reactions, 415
silane (SiH_4) with oxygen, 414
step coverage
arrival angle, 423
conformal coating, 422
deposited film, 422
mean free path, in various gases, 422
reactive sticking coefficient (RSC), 423–424
thin films, profile of, 421
transport of reactants, low pressures, 423
thin films, 414
Cold field emission (CFE), 105–106
Complementary metal oxide semiconductor (CMOS) transistors, 91
Compression molding
advantages, 599
application to LIGA, 630–631
cyclic and continuous process, 599–600
hot embossing, 600–601
principle of, 598
pros and cons, 601
Computer-controlled fabrication (CCF), 364
Computer numerical control (CNC) machining, 632, 634
Conductive polymer patterning, 534–535
Contact printing, 54–55
Conventional injection molding, 595–596
Copper electroplating
in printed circuit boards (PCB), 553–554
for thin film read-write heads, 551–553
Critical dimension (CD), 44
Critical modulation transfer function (CMTF), 62–63
Cross-linking resists, 26–27, 29
Curtain coating, 18, 20

D

Damage mechanism, *see* Energy-driven anisotropy
Damascene processes, 553–556
Deep ion-beam lithography (DIB), 111–112
Deep reactive ion etching (DRIE)
Bosch process, 197
characteristics, 197
electron cyclotron resonance, 198–199
inductively coupled plasma, 196–198
micromold, 632–633
MIT 59 method, 198
scalloping effects, 197–198
Deep UV-sensitive resists (DUV), 632–634
Deep x-ray lithography (DXRL), *see* X-ray lithography
Defocus tolerance, *see* Depth of focus (DOF)
Deposition method, 387–388
Depth of focus (DOF), 58–59
Descumming
definition, 25
dry resist
dry film roll, 38–39
high-aspect-ratio devices, 40
vs. liquid resist, 38–39
potential benefits, 40
dry stripping reactors, 43
glass transition temperature (T_g), 40–41
HMDS reaction mechanism, 41–42
negative resists
bis(aryl)azide rubber resists, 28–29
cross-linking, 26–27, 29

development, 26, 31–32
exposure, 26
lithographic sensitivity, 31
optical proximity effect, 30
pattern transfer, 26
vs. positive photoresists, 30
printed wiring boards (PWBs), 28
ozone-water method, 42–43
permanent resists
dry etching, 33
EPON resist, 34–35
flexible polyimide waveguide, 33
glycidyl ether, 34–35
imidization process, 32–33
LIGA, 37–38
photochemical acid generators (PAGs), 35
photosensitive polyimide process, 33–34
polyimides, 31–33
SU-8 resist, 34–36, 38
T-topping, 36–37
ultrahigh-aspect-ratio structures, 37
photoresist stripping, 42
positive resists
development, 26
diazonaphthoquinone ester (DNQ), 27–28
exposure, 26
pattern transfer, 26
PMMA, 27
polymer chain scission, 26–27
resist stripping
Dry, 43–44
wet, 42–43
resist tone, 26
solid-gas resist stripping, 43
tailoring resist sidewalls, 26
wafer priming, 41–42
Desktop factories (DTFs), 351, 355, 382
conventional and mesoscale machines, volumetric utilization, 380
die punching, 381
size of, 380
Diamond machining, 360–361, 365
Diazonaphthoquinone ester (DNQ) resist, 27–28
Digital micromirror device (DMD™), 461, 499–501
Digital mirror array patterning, chemical forming, 528–529
Dip coating, 20
Dip pen nanolithography (DPN)
principle of, 133–134
tDPN, 135
topographic image, AFM cantilever tip, 135
Direct write e-beam lithography, 104–106
Doctor blading, *see* Tape casting
DOF, *see* Depth of focus (DOF)
Dopant-driven anisotropy, 189–190
DRIE, *see* Deep reactive ion etching (DRIE)
Dry etching, 33
DC
glow discharge, 164–165
plasmas, 160–164
deep reactive ion etching (DRIE)
Bosch process, 197
characteristics, 197
electron cyclotron resonance, 198–199
inductively coupled plasma, 196–198
MIT 59 method, 198
scalloping effects, 197–198
definition, 156–159
dry *vs.*wet isotropic etching, 156
etch endpoint wavelength, 201–202
FTIR sensor capabilities, 202
isotropic etching, 159
methods, 157–158
models, in situ monitoring, 201–202
Paschen's law, 165–167
physical/chemical etching
carbon-containing additives, 195–196
chemically assisted ion-beam etching (CAIBE), 185–186
dopant-driven anisotropy, 189–190
energy-driven anisotropy, 185–188
etch ratios, 192, 194
F/C ratio, 193–194
features of, 190–191
gas compositions, 190–192
gas mixtures evolution, 192–193
hexode reactor, 185
III–V compounds, 195
inhibitor-driven anisotropy, 188–189
mask materials, 192, 195
metal etching, 195
organic films, 195
plasma etchants, 192, 194
reactive plasma gases, 192
selective *vs.* unselective, 194–195
substrate bias, 195
physical etching
energy requirements, 171
etching profiles, 175–177
focused ion beam (FIB), 174
ion-beam milling (IBM), 172–174
sputter deposition, 171
sputtering, 171–172
temperature control, 177–178
plasma-assisted, 157
plasmas, 159
polymeric materials, 208–209
post-CMOS processing, 207–208
profiles, 157–158
radical etching
atmospheric plasma etching, 182–185
macroscopic loading effects, 180–181
microloading effects, 181
reaction mechanism, 178–180
reactor configurations, 178
silicon grass/black silicon, 181–182
RF plasma reactors, 167–171
short dry etching synopsis, 157
silicon (Si), 203
single crystal reactive etching, 203–207
temperature control, 177–178
vapor-phase etching, 199–201
vs. wet etching, 209–211
Dry etching of silylated image resist (DESIRE) process, 75–76
Dry resist
dry film roll, 38–39
high-aspect-ratio devices, 40
vs. liquid resist, 38–39
potential benefits, 40
Dry stripping reactors, 43
Dynamic random access memory (DRAM), 91–92

E

EBL, *see* Electron beam lithography (EBL)
ECR, *see* Electron cyclotron resonance (ECR)
Edge-scattered radiation profile, 48
Electrical discharge machining (EDM), 347
characteristics, 321
μ-EDM, 323–327
sinker/cavity-type, 320–322
wire cutting, 322–323
Electrical discharge machining, wire cutting (EDM-WC), 322–323
Electrochemical fabrication (EFAB), instant masking, 557–559
Electrochemical forming, 511
applications of, 545
electroless metal deposition (*see* Electroless metal deposition)
IC industry and micromachining, 546
Electrochemical grinding, 225–226
Electrochemical isotropic silicon etching
apparatus, 246, 248
critical current density, 249
I-V curve, 248–249
reaction mechanisms, 249–250
two-electrode system, 248
Electrochemical jet etching, 229
Electrochemical machining (ECM)
advantages, 227
cathode tool characteristics, 227
schematic illustration, 226
service providers, 227–228
Electrochemical removal techniques
definition, 225
electrochemical jet etching, 229
grinding, 225–226
machining, 226–228
mask EMM, 230–231

metal removal rate determination, 231–232
microelectrochemical machining, 228–229
scanning electrochemical microscope (SECM), 230
Electroless metal deposition
electrodeposition
anodization, 556
copper electroplating for, 551–554
instant masking, in electrochemical fabrication, 557–559
localized electrochemical deposition, 560
microjet plating and laser-enhanced jet plating, 559–560
polymer masks, 556–557
principles, 549
single and dual damascene processes, 553–556
slanted and curved plated metal shapes, 560–561
through-mask electroplating, 550–551
plating
buried Cu conductor process, 548
in IC Industry, 548
in microfabrication, 549–550
principles
electroless and immersion deposition, 546
exaltants, 547
hydrogen reduction, 547
Ni deposition, 547
plating baths, 548
Electrolytic in-process dressing (ELID)
desktop machine, 369
grinding method, 368
setup for, 367
Electron beam lithography (EBL), 104–107
Electron beam machining, 348
Electron cyclotron resonance (ECR), 198–199
Electron cyclotron resonance CVD (ECR-CVD), 69–70, 433–434
Electronic ink, 521–522
Electronic stringency control, 564
Electron projection lithography (EPL), 107–109
Electron scanning probe lithography, bias effect, 130–131
Electrophoretic photoresist deposition (ED), 19–20
Electroplating, 622
Electrospray ionization (ESI), 574
Electrostatic spraying, 19
Elwenspoek et al.'s model
activation energy, 274
free energy change, 275–276
isotropic etching condition, 276–277
kinetic roughening, 276–277
roughening transition temperature T_R, 276
Energy-driven anisotropy, 185–188
Epitaxy
epilayer thickness, 440
liquid phase epitaxy (LPE), 439
selective epi, 439–440
Si, 425–427, 438
solid-phase epitaxy, 439
EPON resist, 34–35
Etch-back process, 69
Etch-stop techniques
boron, 258–260
electrochemical, 260–263
parameters, 257
photoassisted electrochemical, 263
photoinduced preferential anodization, 263–264
thin insoluble films, 264–265
wider mask openings design, 258
Excimer laser, 342–344
blood gas sensor, 345
stripping of wires, 345
Exposure technology
immersion lithography, 79–81
Köhler illumination, 79
OAI, 78–79
Extreme ultraviolet lithography (EUVL)
depth of focus (DOF) equation, 95
reflective masks, 96
resolution (R), 95
setup for, 95–96
Extrusion coating, 18, 20–21

F

Fast atom beam (FAB) machining, 379–380
Femtosecond lasers, 346
Flame spray pyrolysis, 572–573
Flexible manufacturing system (FMS), 364
Flexible polyimide waveguide, 33
Focused ion beam (FIB)
lithography, 111
machining
etching process, 379
tungsten structures, 378–379
uses of, 378
Four-jet lathe burner, 567
Fused deposition modeling (FDM) method, 538–539

G

Gaussian laser beam, 331–332
Glass transition temperature (T_g), 40–41
Glow discharge, 164–165
Glycidyl ether, 34–35
Gray-tone masks (GTMs)
3D components fabrication, 12
definition, 10–11
high-energy beam-sensitive (HEBS) glass, 13
magnetron sputtering, 12–13
Rohm and Haas dual-tone resist, 11–12
Green tape technology, 579–580
Grinding wheels, 366–367

H

Hexagonal honeycomb polysilicon (HEXSIL) process
schematic illustration of, 489
tweezers, SEM micrograph of, 491
Hexamethyldisilazane (HMDS) reaction mechanism, 41–42
Hexode reactor, 185
High-density plasma CVD (HDPCVD), 418, 433
High-energy beam-sensitive (HEBS) glass, 13
Hinged polysilicon, 487–489
Holographic/interference lithography
arrangement, 138
hologram recording, 137
Hydrofluoric acid (HF), 494
Hydroxyethyl methacrylate (HEMA), 617

I

IBM, *see* Ion-beam milling (IBM)
ICP, *see* Inductively coupled plasma (ICP)
IH process, 540–542
Imidization process, 32–33
Immersion lithography, 79–81
Imprint lithography
application, 115–116
NIL, 113–114
SFIL, 114–115
Inductively coupled plasma (ICP), 196–198
Inhibitor-driven anisotropy, 188–189
Injection molding
conditions for, 596
conventional IM, 595–596
equipment, 597–598
setup, 595
stamper substrates, material selection for, 595–596
thermostatting mold parts, 594
thin-wall IM, 596–597
Ink-jet printing
advantage of, 531–532
bubble jet, 529
deposition system, 530
Hewlett Packard's (HP) thermal ink jet (TIJ) technology, 531
piezoelectric, 530
Integrated circuits (ICs), 387–388
compatibility, surface micromachining, 477–481
subtractive processes in, 151–152
Intelligent Micro Patterning's SF-100 lithography system, 15
Interference lithography, *see* Holographic/interference lithography
Intrinsic resist sensitivity, 45–46

Ion-beam etching (IBE), *see* Ion-beam milling (IBM)
Ion-beam lithography, 111–112
Ion-beam milling (IBM)
  cyclotron resonance frequency, 174
  DC triode setup, 172
  electron path, 173
  magnetically enhanced ion etching (MIE), 172–173
  sputter yield, 174
Ion cluster beam (ICB) deposition
  cluster beam technology, 406
  film formation, with atom clusters, 406
  nanoparticles, 405–406
  setup for, 406
Ion plating, PVD, 405
Ion projection lithography (IPL), 112–113
Iso-etch curves, 241–243
Isotropic etching, 159
  Arrhenius plot, 243
  dopant dependence, 245–246
  electrochemical, 246–250
  iso-etch curves, 241–243
  loading effects, 243–244
  masking materials, 244–245
  risk of, 246
  semiconductor bias
    anodic dissolution, 280–281
    cyclic voltammogram, 280
    electroless etching, 280
    electropolishing, 281–282
    silicon, 282–296
  simplified reaction scheme, 241
  uses, 240

K

Knudsen number, 423
Köhler illumination, 79
Krypton uoride (KrF) laser, in polyimide, 344

L

Laminated object manufacture (LOM) process, 538
Langmuir–Blodgett (LB)
  deposition, 523–524
  resists, 116
Laser ablation
  for carbon nanotube production, 409
  chemical etching, 339–341, 347
  depth of focus/field
    numerical aperture (NA), 333
    ranges, 333
    transverse electromagnetic (TEM) modes, 334
  dilemma, 338
  heat-affected zone (HAZ), 330
  laser power, 334–335
  parameters, 329–330
  pulse length, 330
    long pulses, 335–336
    short pulses, 336–338
  PVD, 407–410
  shot-to-shot repeatability, 338
  spot size
    Gaussian beams, 331–332
    hole cut, size of, 333
    mass, 331
    Raleigh range, 332
    waist spot size and collimation distances, 332
  water jet-guided laser machining, 338–339
  wavelength, 330–331
Laser-assisted chemical etching (LACE), 339–340
Laser-assisted chemical vapor deposition (LCVD), 586–587
Laser beam machining (LBM)
  applications, 327–328
    advantages of, 340–341
    blood gas sensor, 345
    disadvantages of, 341–342
    EDM, mechanical machining, and chemical etching, comparisons of, 347
    excimer laser, 342–344
    hole-drilling capabilities, 342
    ink-jet printers, microvias, catheters, and DVDs, 344–346
    krypton uoride (KrF) laser, in polyimide, 344
    Laser Centrum in Hannover (LZH), 346
    long laser pulse machining, 343
    silicon capacitive pressure sensor, 343
    transparent materials, 346
    wires, stripping of, 345
  CO_2 laser, 327
  definition, 327
  excimer, 327
  laser ablation (*see* Laser ablation)
  manufacturing, 329
  Nd:YAG, 327
Laser Centrum in Hannover (LZH), 346
Laser microablation, 632, 634
Laser spoiling, 239
Liftoff profile, 49–50
LIGA, *see* Lithographie, Galvanoformung, Abformung (LIGA)
Line width metrology, 44
Liquid crystal display (LCD), 543–544
Liquid resin molding techniques
  pros and cons, 595
  reaction injection molding (RIM), 594
  transfer molding (TM), 594
Lithographic sensitivity
  definition, 52
  intrinsic resist, 44
  response curves, 52
  size *vs*. dose, 53
Lithographie, Galvanoformung, Abformung (LIGA), 37–38
  common x-ray resists, IC industry, 615
  demolding, 626–628
  development, 621–622
  exposure
    deposited dose, 619–620
    optimum wavelength, 618–619
    slanted microstructures, 621
    stepped microstructures, 620–621
  fiber-chip coupling, 637–638
  history of, 601–602
  inclined LIGA walls, 636–637
  metal deposition
    electrodeposition in high-aspect-ratio resist crevices, 624
    electroplating, 622
    finishing metal parts and metal mold inserts, 625–626
    nickel electrodeposition, 622–623
    plating automation, 626
    plating compatibility issues, 624–625
    pulse plating, 623–624
  molding technology
    ceramics, 634
    compression molding, 630–631
    master micromold fabrication methods, comparison, 631–634
    polyvinylidene fluoride (PVDF), 634
    reaction injection molding (RIM), 629–630
    sol-gel techniques, 635
  plastic microfabrication, 592
  primary substrate, 613–614
  resist requirements, 614–615
  resists properties, 615
  sacrificial layers, 628
  spinneret nozzles, 635–636
  synchrotron orbital radiation (SOR)
    access to, 605–606
    applications, 603
    electron energy, 604–605
    fabrication of x-ray masks, 603
    pros and cons, 602
    schematic of x-ray exposure station, 604
    total radiated power, 605
  x-ray lithography, 98–100
  x-ray masks
    absorber fabrication, 610–611
    absorber materials, 609–610
    in IC industry, 607
    membrane materials, 608–609
    plastic molding and second electroforming, 606
    substrate assembly in x-ray scanner, 607
    transfer mask, 612–613
    x-ray deep-etch lithography, 606

x-ray resist application methods
commercial PMMA sheets, 615–616
multiple spin coats, 615
resist adhesion, 616–617
stress-induced cracks, PMMA, 617–618
Lithography research
block copolymer lithography
applications, 121–122
diblock copolymer phase diagram, 118
flash memory, 120
investigation, 117
patterned wettability, 120–121
vs. traditional lithography, 119
holographic/interference lithography
arrangement, 138
hologram recording, 137
plasmon lithography, 127–128
quantum lithography, 124–125
scanning proximal probe lithography
ANSOM, 135–136
atom lithography/ mechanosynthesis, 131–133
dip pen lithography, 133–135
electrochemical effect, 129–130
electron scanning probe lithography, bias effect, 130–131
NSOM, 135–136
scratch lithography-mechanical effect, 129
STM, 128–129
soft lithography
advantages, 141
curved glassy carbon structure, 140
disadvantages, 141
microcontact printing (μ-CP), 139–140
micromolding in capillaries (MIMIC), 141
microreplica molding, 141
microtransfer molding (μ-TM), 140
PDMS, 138–139
thin resist layers
Langmuir–Blodgett film, 116
self-assembled monolayers, 116–117
ultrathin film (UTF), 117
two-photon 3D lithography, 125–127
zone plate array lithography (ZPAL)
coherent light expression, 122–123
Fresnel zone plate, 122–123
maskless optical projection lithography (MOPL), 124
resolution, 124
Low-pressure CVD (LPCVD)
hot wall reactor, 429
polysilicon surface micromachining, 429–430
LUMEX 25C, 567, 588

M

Madou model
band model, 277–278
focus on, 279–280
n-type Si band diagram, 278
reactive redox species, 278–279
surface bonds density, 279
Magnetically enhanced ion etching (MIE), 172–173
Magnetron sputtering, 12–13
Manufacturing techniques
categories, 148–150
etch parameters, 150
removing and forming processes, 148
subtractive processes, 151–152
Maskless array synthesizer (MAS), 528–529
Maskless optical projection lithography (MOPL)
Intelligent Micro Patterning's SF-100 lithography system, 15
laser printer resolution, 13–14
Maskless zone-plate-array lithography (ZPAL), 15
Masks
fiducial marks, projection printing
critical alignment process, 63
errors, in projection printing, 63–64
standard optical vernier test pattern, 63
gray-tone lithography
3D components fabrication, 12
definition, 10–11
high-energy beam-sensitive (HEBS) glass, 13
magnetron sputtering, 12–13
Rohm and Haas dual-tone resist, 11–12
MEMS alignment, projection printing
SUSS MA-150 operation, 64
two-sided alignment jig, 65
optical maskless technology
Intelligent Micro Patterning's SF-100 lithography system, 15
laser printer resolution, 13–14
standard photolithography, 9–10
technology
conventional transmission mask comparison, 77
OPC, 78
phase-shifting masks, 76–78
Mechanical machining, 148, 347
desktop factory, 382
conventional and mesoscale machines, volumetric utilization, 380
die punching, 381
size of, 380
history of
desktop factories, 355
Egyptian bas relief, 352
progress of accuracy in, 354
Ptolemy's astrolabe, 353
manufacturing, absolute and relative tolerances in
human-engineered objects, 356
mesoscale machining, 357
nontraditional and nonconventional, 358
precision machining, 357–358
Taniguchi and Moore plots, 355
precision mechanical machining *vs*. IC-based micromachining, 381–382
ultraprecision
abrasive machining, 365–377
single-point, 360–365
surface finish, 377–380
Mechanical microspotting
arraying methods, 534
spotted cDNA arrays, 532–533
synthesized oligo array, 533
Meniscus coating, 18–19
Mesoscale machining, 357
Metal deposition, in LIGA and pseudo-LIGA
electrodeposition in high-aspect-ratio resist crevices, 624
electroplating, 622
finishing metal parts and metal mold inserts, 625–626
nickel electrodeposition
influence on internal stress, 622–623
operational parameters of nickel sulfamate baths, 622
physical properties, 622
plating automation, 626
plating compatibility issues, 624–625
pulse plating, 623–624
Metal etching, 195
Metalorganic CVD (MOCVD), 433–434
Methacryloxypropyl trimethoxy silane (MEMO), 616–617
Microabrasive blasting, *see* Abrasive gas jet machining (AJM)
Microcontact printing (μ-CP), 139–140, 534
Microelectrical discharge machining (μ-EDM)
and ECM, machining characteristics of, 327
handheld at Panasonic, 323
high-speed micromilling, diamond crystal, 326
hybrid desk-top machine, 325
LIGA, 325
micronozzle fabrication process, 324
pagoda, 319, 325–326
vertical and horizontal, 326
wire cutting, 323–324
Microelectrical discharge machining, wire cutting (μ-EDM-WC), 323–324

Microelectrochemical machining, 228–229
Microelectromechanical systems (MEMS), 387–388, 483–484
Microfabrication
  electroless plating in
    bevel plating, 549–550
    three-electrode system, formation of, 551
  hinges, 487
Microlithographic DOF, 58
Micromachining, 362
  subtractive processes, 151–152
Micromolding in capillaries (MIMIC), 141
Micronozzle fabrication process, 324
Microphotoforming process, *see* Stereolithography/microphotoforming process
Microstereolithography
  projection, 543–544
  vs. rapid prototyping, 540
  and two-photon polymerization, scanning and parallel, 541–543
Microtransfer molding (μ-TM), 140
Milling techniques
  chemical, 218–219
  photochemical, 219–221
Modulation transfer function (MTF)
  coherent and incoherent light, 62
  critical, 62–63
  curves, 61–62
  cutoff frequency, 62
  exposure system, 61
  vs. feature size, 61
  modulation index, 61
  optical imaging quality, 61
Molecular beam epitaxy (MBE)
  advantages, 410
  heteroepitaxy, 410
  schematic of, 410
Moore Nanotechnology Systems LLC, 364
Moore's law, 90
  CMOS transistors, 91
  cooling and interconnections, 92
  description, 90–91
  DRAM bits, 91–92
  Kurzweil's optimistic predictions, 93–94
  read/write head, 94
  second law, 92–93
  SIA road map, 91
MOPL, *see* Maskless optical projection lithography (MOPL)
MTF, *see* Modulation transfer function (MTF)

N

Nanoelectromechanical systems (NEMS), 387–388
Nanogen chip, 562–564
Nanoimprint lithography (NIL), 113
Nanomachining, subtractive processes, 151–152
Nanotech 220UPL, Moore Nanotechnology Systems LLC, 364
Near-field optical scanning microscope (NSOM), 135–136
Next-generation lithographies (NGL)
  charged-particle beam lithography
    deep ion-beam lithography (DIB), 111–112
    EBL, 104–105
    electron-and ion-beam applications, 103
    electron-beam resists, 106–107
    electron emission sources, 105–106
    electron projection lithography, 107–109
    FIB, 111
    IPL, 112–113
    maximum dose, 103
    micromachined electron emission sources, 109–111
  comparison of, 97
  EUVL
    depth of focus (DOF) equation, 95
    reflective masks, 96
    resolution (R), 95
    setup for, 95–96
  imprint lithography
    application, 115–116
    NIL, 113–114
    SFIL, 114–115
  x-ray lithography
    blurring (δ) expression, 102
    electron-beam x-ray source, 102
    features of, 98
    LIGA, 98–100
    light sources, 97
    masks, 100–101
    resists, 99–100
    synchrotron, 97–98, 102–103
NSOM, *see* Near-field optical scanning microscope (NSOM)

O

Off-axis illumination (OAI), 78–79
Oligo microarray, 527
OPC, *see* Optical proximity correction (OPC)
Optical maskless technology
  Intelligent Micro Patterning's SF-100 lithography system, 15
  laser printer resolution, 13–14
Optical proximity correction (OPC), 78
Optical proximity effect, 30
Organic light-emitting diodes (OLEDs), 521–522
Ozone-water method, 42–43

P

Paschen's law, 165–167
Phase-shifting masks, 76–78
Phosphosilicate glass (PSG), 466, 468–469
Photochemical acid generators (PAGs), 35
Photochemical forming, 509, 511
  rapid prototyping (RP)
    3D-printer, Z Corporation, 538
    FDM method, 538–539
    LOM process, 538
    techniques, 536–537
  stereolithography/microphotoforming process
    microstereolithography and two-photon polymerization, 541–543
    principles, 539–541
    projection microstereolithography, 543–544
Photochemical milling (PCM)
  definition, 219–220
  of metals, 220–221
Photochemical quantum efficiency, 45–46
Photofabrication
  definition, 221
  industrial application, 221–222
  photosensitive glasses/ceramics
    composition, 222
    development of, 224
    fluidic components, 222–223
    FotoForm glass application, 222
    Foturan, 223–224
    heat treatment, 222
    microfluidic structures, 223
  photosensitive plastics, 224–225
Photolithography
  aqueous development, 25
  batch spray development, 25
  coating methods comparison, 21
  critical dimension (CD), 44
  definition, 4–5
  descumming and postbaking
    definition, 25
    dry resist, 38–40
    dry stripping reactors, 43
    glass transition temperature (T_g), 40–41
    HMDS reaction mechanism, 41–42
    negative resists, 28–31
    ozone-water method, 42–43
    permanent resists, 31–38
    photoresist stripping, 42
    positive resists, 26–28
    resist stripping, 42–44
    resist tone, 26
    solid-gas resist stripping, 43
    tailoring resist sidewalls, 26
    wafer priming, 41–42

development, 24–25
dry development, 25
electron-beam, 10
exposure treatment
bleaches, 23–24
exposure station, 22
pilot and production lines development, 23
resist-coated wafer, 22
resolution-enhancing techniques (RETs), 23
of hydrogels, gas-permeable membranes and ISEs, 524–525
IC production processes, 8
intrinsic resist sensitivity, 45–46
line width metrology, 44
lithographic sensitivity, 44
definition, 52
response curves, 52
size *vs.* dose, 53
maskless zone-plate-array lithography (ZPAL), 15
mask polarity, 9
masks
gray-tone lithography, 10–13
optical maskless technology, 13–15
standard photolithography, 9–10
origin, 3–4
photochemical quantum efficiency, 45–46
photoresist deposition methods
curtain coating, 18, 20
definition, 15–16
dip coating, 20
electrophoretic photoresist deposition (ED), 19–20
electrostatic spraying, 19
extrusion coating, 18, 20–21
meniscus coating, 18–19
plasma-deposited resist, 19
roller coating, 18, 20
silkscreen printing, 19
soft baking/prebaking, 21–22
spin coating, 16–18
spray coating, 17–19
planarization
4-MB DRAM, 68
reflective notch, 67–68
strategies, 68–70
postexposure treatment, 24
projection lithography
k_1 evolution *vs.* time, 82
minimum line width and exposure wavelength *vs.* year, 80–81
numerical aperture evolution, 81–82
projection printing, resolution
DOF, 58–59
error types, 63–64
mask aligner, 60
mask alignment, 63–65
MTF, 61–63
numerical aperture (NA), lens, 57
optical, 56
Pendry's near-field amplifying superlens (NFSL), 57
resist profiles, 65–67
scanning projection system, 59
step and scan systems, 60–61
stepper system, 60
resist profiles
image reversal, 50–52
liftoff profile, 49–50
types of, 46–49
RET
exposure technology, 78–80
mask technology, 76–78
miniaturization science, 70
resist performance, 71–76
shadow printing, 9–10
contact printing, 54–55
light distribution profiles, 54
proximity printing, 55
self-aligned masks, 55–56
silicon-based device, 8–9
spray developer, 24
wafer cleaning
clean room classification system, 5–6
clean room contaminant sources, 5–6
supercritical cleaning, 7–8
wet *vs.* dry cleaning attributes, 7
wet development, 24
Photoresist deposition methods
curtain coating, 18, 20
definition, 15–16
dip coating, 20
electrophoretic photoresist deposition (ED), 19–20
electrostatic spraying, 19
extrusion coating, 18, 20–21
meniscus coating, 18–19
plasma-deposited resist, 19
roller coating, 18, 20
silkscreen printing, 19
soft baking/prebaking, 21–22
spin coating
resist spinner, 16
spinning process, 16–17
troubleshooting chart, 17
spray coating, 17–19
Photoresist stripping, 42
Photosensitive glasses/ceramics
composition, 222
development of, 224
fluidic components, 222–223
FotoForm glass application, 222
foturan, 223–224
heat treatment, 222
microfluidic structures, 223
Photosensitive plastics, 224–225
Photosensitive polyimide process, 33–34
Physical etching
and chemical etching
carbon-containing additives, 195–196
chemically assisted ion-beam etching (CAIBE), 185–186
dopant-driven anisotropy, 189–190
energy-driven anisotropy, 185–188
etch rate experiment, 186–187
etch ratios, 192, 194
F/C ratio, 193–194
features of, 190–191
gas compositions, 190–192
gas mixtures evolution, 192–193
hexode reactor, 185
III–V compounds, 195
inhibitor-driven anisotropy, 188–189
mask materials, 192, 195
metal etching, 195
organic films, 195
plasma etchants, 192, 194
reactive plasma gases, 192
selective *vs.* unselective, 194–195
substrate bias, 195
trench profile manipulation, 188
energy requirements, 171
etching profiles
angle-dependent sputter rate, 175–176
angular distribution, incident ions, 177
backscattering, 177
erodible or sacrificial masks, 176
limitations, 176
redeposition, 177
sputtering yield, 175
ion-beam milling (IBM)
cyclotron resonance frequency, 174
DC triode setup, 172
electron path, 173
magnetically enhanced ion etching (MIE), 172–173
sputter yield, 174
sputter deposition, 171
sputtering, 171–172
temperature control, 177–178
Physical vapor deposition (PVD)
aerosol deposition (AD)
characteristics of, 412
piezoelectrics in, 412–414
schematic illustration of, 412
zirconate ceramics, 413
applications, 393
ion cluster beam (ICB) deposition
cluster beam technology, 406
film formation, with atom clusters, 406
nanoparticles, 405–406
setup for, 406
ion plating, 405
molecular beam epitaxy (MBE)
advantages, 410
heteroepitaxy, 410
schematic of, 410
pulsed laser, laser ablation deposition, 407–410

sputtering
amount of material, 400
cathode shield, 401
continuous deposition, 403–404
DC plasma, 402
vs. evaporation, 398–400
magnetron, 402–403
polysilicon, 404–405
reactive sputtering, 403
setup for, 399
sputter deposition, characteristics of, 403
sputter yield, 400–401
type of material in, 403
steps, 392
thermal evaporation
arrival rate, 397
conformal and nonconformal step coverage, 398
data for, 394
deposited film thickness, 397
electron beam (e-beam), 394–395
flux, 393, 395
fraction of particles scattered, 395
geometric considerations in, 396
heat sources, comparison of, 395
Knudsen number, 396
radiofrequency (RF) induction heating, 395
reactive evaporation, 398
resistive heating, 394
SAMPLE computer program, thin film profile, 398–399
shadowing effect, 397
thin film deposition, 394
types of, 393
Planarization
ECRCVD, 434
4-MB DRAM, 68
reflective notch, 67–68
strategies
CMP process, 70
CVD, 69
ECR-CVD, 69–70
etch-back process, 69
spin-on glasses, 69
Planar parallel-plate reactor, 167
Plasma-deposited resist, 19
Plasma-enhanced CVD (PECVD)
polysilicon, 432–433
reactor configurations, 431
silicon nitride, 432
Plasmas
arc cutting/machining, 348–349
arc spray deposition, 584–585
breakdown voltage, 166
corona discharge, 166
DC
cathode dark space characteristics, 161
discharge, in lightning, 163
glow discharge, 164–165
sputter etching, 163–164
voltage distribution, 160
definition, 159
electron temperature, 162
etching (*see* Radical etching)
ion temperature, 162
obstructed glow, 162
Paschen's law, 165–167
polymerization
advantages, 520
polymers deposition, 519
schematic of, 519
Yasuda parameter, 519
RF plasma reactors
DC bias, 167–168
device damage, 170
electron loss, 168–169
etch performance, 170
planar parallel-plate reactor, 167
positive argon ions, 169
reactive ion etching, 169
sheath voltages, 169–170
trenches, 170–171
spraying
multilayer planar oxygen sensor, 582–584
plasma arc deposition, 584–585
spray nozzle, 581–582
thermal barrier coating (TBC), 582
Plasmon lithography, 127–128
Plastic molding
compression molding
advantages, 599
cyclic and continuous process, 599–600
hot embossing, 600–601
principle of, 598
pros and cons, 601
injection molding
conventional IM, 595–596
equipment, 597–598
setup, 595
stamper substrates, material selection for, 595–596
thermostatting mold parts, 594
thin-wall IM, 596–597
typical conditions for, 596
liquid resin molding techniques
pros and cons, 595
reaction injection molding (RIM), 594
transfer molding (TM), 594
Polydimethylsiloxane (PDMS), 138–139
Polymer
casting, 517
chain scission resist, 26–27
dip coating, 516
doctor's blade/knife coating, 517–518
dry film lamination, 516
masks, electrodeposition, 556–557
spraying
electrostatic spray, 516–517
thermosetting and thermoplastic coatings, 516
Polymethylmethacrylate (PMMA)
commercial resist sheets, 615–616
cracking of, 614
positive photoresists, 27
stress-induced cracks, 617–618
Polysilicon
analog devices accelerometer
deposition and IC metallization, 498
IC chip preparation for, 498
doped, 455–457
hinged, 487–489
LPCVD, 429–430
PECVD, 432–433
PVD, 404–405
SOI, 486–487
sputtering, 404–405
undoped, 453–455
Polyvinylidene fluoride (PVDF), 634
Porous polysilicon, surface micromachining, 493–495
Postbaking, *see* Descumming
Powder blasting, AJM, 375–377
Prandtl number, 419
Preferential etching, 246
PREVAIL, *see* Projection reduction exposure with variable axis immersion lenses (PREVAIL)
Printed circuit boards (PCB), copper electroplating, 553–554
Projection lithography
k_1 evolution *vs.* time, 82
minimum line width and exposure wavelength *vs.* year, 80–81
numerical aperture evolution, 81–82
Projection microstereolithography, 543–545
Projection printing, 10
DOF, 58–59
error types, 63–64
mask aligner, 60
mask alignment, 63–65
MTF, 61–63
numerical aperture of lens, 57
optical, 56
Pendry's near-field amplifying superlens (NFSL), 57
resist profiles, 65–67
scanning projection system, 59
step and scan systems, 60–61
stepper system, 60
Projection reduction exposure with variable axis immersion lenses (PREVAIL), 108
Protein patterning, 560–561
antigens, UV radiation of, 562
immunosensor, biological grating, 562
with lithography, 527–528
Proximity printing, 55
Pulsed laser deposition (PLD)
advantages and disadvantages, 408
direct writing in, 410
schematic of, 407
steps, 407

Q

Quantum lithography, 124–125

R

Radical etching
  atmospheric plasma etching
    arc-free plasma etching, 184–185
    downstream plasma, 183–184
    voltage current plot, 182–183
  macroscopic loading effects, 180–181
  microloading effects, 181
  plasma etch process, 178–179
  reaction mechanism, 178–180
  reactor configurations, 178
  silicon grass/black silicon, 181–182
Radiofrequency (RF) plasma reactors
  DC bias, 167–168
  device damage, 170
  electron loss, 168–169
  etch performance, 170
  planar parallel-plate reactor, 167
  positive argon ions, 169
  reactive ion etching, 169
  sheath voltages, 169–170
  trenches, 170–171
Rapid prototyping (RP)
  3D-printer, Z Corporation, 538
  FDM method, 538–539
  LOM process, 538
  techniques, 536–537
Reaction injection molding (RIM), 594
  fabrication, 629
  gate plate, 629–630
  Ni honeycomb structure, 630–631
Reactive ion etching, 169
Resist performance
  antireflection coatings, 73–74
  chemically amplified resists, 71–73
  thin film imaging (TFI)
    DESIRE process, 75–76
    multilayer resist, 76
    single-layer resist, 74–76
Resist profiles
  cross-linking, 51–52
  dual-tone resist chemistry, 51
  image reversal, 50–52
  liftoff profile, 49–50
  presoak process, 50
  projection printing, 65–67
  types of
    edge-scattered radiation profile, 48
    lightly scattered zones, 49
    overview, photoresist profiles, 47
    re-entrant, 46–47
    reverse, 46–47
    undercut, 46–47
Resists
  descumming
    dry, 38–40
    negative, 28–31
    permanent, 31–38
    positive, 26–28
    tone, 26
  stripping
    dry, 43–44
    wet, 42–43
Resolution enhancement technology (RET)
  exposure technology
    immersion lithography, 79–81
    Köhler illumination, 79
    OAI, 78–79
  mask technology
    OPC, 78
    phase-shifting masks, 76–78
  miniaturization science, 70
  resist performance
    antireflection coatings, 73–74
    chemically amplified resists, 71–73
    thin film imaging, 74–76
Reynolds number, 418–419
Rohm and Haas dual-tone resist, 11–12
Roller coating, 18, 20

S

Salt-assisted spray pyrolysis, 572
Scanning electrochemical microscope (SECM), 230
Scanning electron microscope (SEM), 203, 288, 307–308, 370, 491, 612
Scanning projection system, 59
Scanning proximal probe lithography
  ANSOM, 135–136
  atom lithography/mechanosynthesis, 131–133
  dip pen lithography, 133–135
  electrochemical effect, 129–130
  using electrons, bias effect, 130–132
  NSOM, 135–136
  scratch lithography, mechanical effect, 129
  STM/AFM, 128–129
Scanning tunneling microscope (STM), 128–129
Scattering with angular limitation projection electron-beam lithography (SCALPEL), 107–108
Schottky emission (SE), 105–106
Screen printing/silk screening
  thick films, 576
  types of inks/pastes and application, 578–579
  working of, 577–578
Seidel et al.'s model
  backbond surface states, 270–271
  fluctuating energy level model, 268–269
  key points, 273
  Si crystal property correlation, 273–274
  silicon oxidation reaction, 269
  two proton detachment, 271
  water concentration effect, 271–272
Selective epitaxy surface micromachining, 485–486
Selective laser sintering (SLS), 587–588
Self-assembled monolayers (SAMs), 116–117, 523–524
Semiconductor bias
  anodic dissolution, 280–281
  cyclic voltammogram, 280
  electroless etching, 280
  electropolishing, 281–282
  silicon, 282–296
Semiconductor Industry Association (SIA) road map, 91
Shadow printing
  contact printing, 54–55
  light distribution profiles, 54
  proximity printing, 55
  self-aligned masks, 55–56
Sheath voltages, 169–170
Short dry etching, 157
Silicon
  carbide, 365
  grass, 181–182
  macroporous
    2D photonic crystal, 287–288
    electrochemical etching, 287
    properties, 287
  microporous
    electrical resistivity, 286–287
    formation, 285–286
    nanoporous structure, 287
    surface morphology, 285
  nitride, PECVD, 432
  porous
    aging, 289
    applications, 284, 292–296
    definition, 284
    discovery, 282–283
    drying, 288–289
    geometry of, 284
    models of, 289–292
    photoluminescence (PL) effect, 283
    pore size distribution, 284–285
    properties, 286
    visible light emission, 283–284
  wafer laser machining, 339
Silicon fusion bonded (SFB) micromachining, 484
Silicon on insulator (SOI), surface micromachining
  advantages, 487
  layer of, 481
  in MEMS, 483–484
  vs. polysilicon surface micromachining, 486–487
  transistors, 481
  wafer fabrication techniques, 482–483
Silk screening/screen printing, 19
  electronic ink, 521–522
  OLED, 521–522
  sensors, 520–521
  thick film technology, 520
  of T-shirts, 520

Single-crystal Si
anisotropic etching, doping dependence
etch stop region, 273
mechanisms, 272
moderate dopant concentration, 272–273
atomistic model, Gosalvez-Kelly
chemical oxidation, 267
electrochemical mechanisms, 266–267
nucleophilic attack representation, 265–266
origin, 265
Si etching process, 267
surface microstructure, 268
surface morphology, 267–268
Elwenspoek et al.'s model
activation energy, 274
free energy change, 275–276
isotropic etching condition, 276–277
kinetic roughening, 276–277
roughening transition temperature T_R, 276
empirical data
isotropic/anisotropic etches, 234–236
laser spoiling, 239
[100]-/[110]-oriented wafers, 238–239
Wagon-Wheel masking patterns, 236–238
Madou model
band model, 277–278
focus on, 279–280
n-type Si band diagram, 278
reactive redox species, 278–279
surface bonds density, 279
Seidel et al.'s model
backbond surface states, 270–271
fluctuating energy level model, 268–269
key points, 273
Si crystal property correlation, 273–274
silicon oxidation reaction, 269
two proton detachment, 271
water concentration effect, 271–272
Single-point mechanical machining
CCF, 364
chip formation, during diamond milling, 360
diamond ultrahigh-precision tools, 361
Fanuc's ROBOnano Ui, 351, 363
flexible manufacturing system, 364
Micro-Car, 362–363
micromachining processes, 362
Moore Nanotechnology Systems LLC, 364
tools for, 360
types of, 360
Soft baking/prebaking, 21–22
Soft lithography
advantages, 141
curved glassy carbon structure, 140
disadvantages, 141
microcontact printing (μ-CP), 139–140
micromolding in capillaries (MIMIC), 141
microreplica molding, 141
microtransfer molding (μ-TM), 140
polydimethylsiloxane (PDMS), 138–139
Sol-gel technique
chemical forming
aerogel, 513–514
definitions, 512
gelation, 513
nanoparticle products and manufacture, 512
processes and products in, 513
silica sol formation, 514
xerogel, 514
LIGA, 635
Solid-gas resist stripping, 43
Spark discharge deposition, 585
Spin coating process
resist spinner, 16
spinning process, 16–17
troubleshooting
air bubbles, wafer surface, 82
chart, 17
fluid velocity, 83
high exhaust volume, 83
spin bowl exhaust rate, 82
spin chuck, 83
thick film, 82
thin film, 82
variable exhaust/ambient conditions, 83
Spin-on glasses (SOGs), 69
Spray coating, 17–19
Spray pyrolysis
atomization, 568
chemical vapor deposition (CVD), 568–569
decomposition of precursors, 570
droplet evolution, 571
setup for
film deposition, 569–570
powder synthesis, 570–571
surface precipitation, 571
in Taguchi gas sensor, 575–576
with and without chemical reaction, 571–572
Sputter etching, 163–164
Sputtering, PVD
amount of material, 400
cathode shield, 401
continuous deposition, 403–404
DC plasma, 402
vs. evaporation, 398–400
magnetron, 402–403
polysilicon, 404–405
reactive sputtering, 403
setup for, 399
sputter deposition, characteristics of, 403
sputter yield, 400–401
type of material in, 403
Step-and-flash imprint lithography (SFIL), 113
Stepper system, projection printing, 60
Stereolithography/microphotoforming process
advantages, 544
microstereolithography
projection, 543–544
vs. rapid prototyping, 540
and two-photon polymerization, scanning and parallel, 541–543
principles
fixed surface method, 540
photomask set, exposure with, 541
scanning and projection stereolithography, 540–541
Stiction, in surface micromachining
in-use stiction, 473–474
phenomenon in, 472
stand-off bump/antistiction bump, 472–473
STM, *see* Scanning tunneling microscope (STM)
SU-8 resist, 34–36, 38
Surface micromachining, 391
analog devices accelerometer, 496–499
vs. bulk micromachining, 495–496
dimensional uncertainties, 476–477
extensions of, 459–460
film stress, control of, 474–476
history of
ADXL-50 chip, from Analog Devices, 461
DMD, 461
micromotors of, 460–461
IC compatibility
and CMOS, 477–480
embedded micromechanics approach, 481
LPCVD deposition, critical process temperatures for, 478
SUMMiT-V™, 481
tungsten metallization, 479
modifications
buried cavities, silicon fusion bonding, 484
hinged polysilicon, 487–489
milliscale molded polysilicon structures, 489–491
polysilicon surface *vs.* SOI, 486–487
porous polysilicon, 493–495
pressure sensors, in separated by implanted oxygen (SIMOX) surface, 484–485

selective epitaxy, 485–486
silicon on insulator (SOI), 481–484
structural elements and molds in, 492
thick polysilicon, 491–492
UV depth lithography, 492–493
polysilicon anchor, 458
processes
buffer/isolation layer, pattern transfer, 466–467
etchant-spacer-microstructure combinations, 471–472
lithography levels, 462
MUMPs, 461–462
physical layers, 462
polyimide corrugated structure, 472
PolyMUMPs, 463–465
polysilicon micromotor, 462, 466
selective etching, of spacer layer, 469–471
spacer/sacrificial layer deposition, 468–469
stiction, 471–474
structural material, deposition of, 467–468
wet etch pattern transfer, 467
resonant frequency, 475–477
sealing processes in, 477
Si single crystal *vs*. crystalline polysilicon, 496
TI micromirrors, 499–501
Synchrotron orbital radiation (SOR)
access to, 605–606
applications, 603
electron energy, 604–605
fabrication of x-ray masks, 603
pros and cons, 602
schematic of x-ray exposure station, 604
total radiated power, 605

T

Tape casting, 579–580
Tetraethylorthosilicate (TEOS), with oxide, 427–428
Texas Instruments' (TI's) micromirrors, 499–501
TFI, *see* Thin film imaging (TFI)
Thermal-dip pen lithography (tDPN), 135
Thermal evaporation, PVD, 393–398
Thermal machining, 148–149
Thermal removal processes
electrical discharge machining (EDM)
characteristics, 321
μ-EDM, 323–327
sinker/cavity-type, 320–322
wire cutting, 322–323
electron beam machining, 348
history of, 320
laser beam machining (LBM)
applications, 327–328, 340–348
CO_2 laser, 327
definition, 327
excimer, 327
laser ablation, 329–340
manufacturing, 329
Nd:YAG, 327
plasma beam machining, 348–349
Thermoforming
electron beam forming process, 588–589
electrospray pyrolysis
electrospinning, 574
electrospray ionization (ESI), 574
monodisperse droplets, 573
flame spray pyrolysis, 572–573
four-jet lathe burner, 567
laser machining forming process
laser-assisted chemical vapor deposition (LCVD), 586–587
selective laser sintering (SLS), 587–588
LUMEX 25C, 567
plasmas and nanoparticles, 585–586
plasma spraying
multilayer planar oxygen sensor, 582–584
plasma arc deposition, 584–585
spray nozzle, 581–582
thermal barrier coating (TBC), 582
salt-assisted spray pyrolysis, 572
screen printing/silk screening
thick films, 576
types of inks/pastes and application, 578–579
working of, 577–578
spray pyrolysis
atomization, 568
chemical vapor deposition (CVD), 568–569
decomposition of precursors, 570
droplet evolution, 571
setup for film deposition, 569–570
setup for powder synthesis, 570–571
surface precipitation, 571
in Taguchi gas sensor, 575–576
with and without chemical reaction, 571–572
tape casting, 579–580
Thin film imaging (TFI)
DESIRE process, 75–76
multilayer resist, 76
single-layer resist, 74–76
Thin films
adhesion, 445–447
characterization, 458
deposition characteristics, 441
processing methods, 442
read-write heads, copper electroplating, 551–553
requirements, 441
strength of
Al films, 445–446
boundary strengthening, 443
controls, 443
Hall–Petch behavior, 444
Hall–Petch equation, 443
polycrystalline material, dependence, 443
reverse Hall–Petch behavior, 444
strain hardening/cold work, 443
yield strength of, 445
stress in
coarse-grain, 455–456
deposition techniques, comparison of, 452
doped polysilicon, 455–457
fine-grain, 455–456
gradients, bending moment, 448
growth temperature, 449–450
intrinsic and extrinsic residual stresses, 447
intrinsic stress, 451–452
PECVD and sputtered polysilicon, 457–458
polysilicon deposition and film microstructure, 452–453
qualitative description, 447–450
quantitative analysis, 450
RF excitation, frequency of, 449
RF power effects, 449
tensile and compressive stress, 448
thermal stress, 450–451
total reactor pressure, 449
undoped polysilicon, 453–455
Thin resist layers
Langmuir–Blodgett film, 116
self-assembled monolayers, 116–117
ultrathin film (UTF), 117
Thin-wall injection molding, 596–597
Transfer molding (TM), 594
Transverse electromagnetic (TEM) modes, laser ablation, 334
Trenches
depth, silicon, 170–171
profile manipulation, 188
T-topping, 36–37
Tungsten structures, in focused ion beam, 378–379
Two-photon 3D lithography, 125–127

U

Ultrahigh-aspect-ratio structures, 37
Ultraprecision mechanical machining
abrasive machining
categories of, 365
conventional synthetic abrasives, 365
electrolytic in-process dressing, nanofinishing of, 367–368

grinding, 366–367
jet machining-powder blasting, 375–377
Knoop Hardness for, 366
ultrasonic machining, 368–372
water jet machining, 372–375
dimensional stability, 358
drilling, 358–359
grinding, 359
lathe, 359
mills, 358
single-point, 360–365
surface finish
arithmetic average roughness, 377
arithmetic mean value, 377
cost of, 378
fast atom beam, 379–380
focused ion beam, 378–379
roughness and center line, 377
turning, 358
Ultrasonic grinding, *see* Ultrasonic machining (USM)
Ultrasonic machining (USM)
abrasive, 370–371
advantages and disadvantages of, 372
coin with grooving, 371
components of, 370
hardness (H)/modulus of elasticity (E), 371
polycrystalline silicon slab, 371
SEM photomicrographs, 370
tool material, 369
ultrasonic mill, 372
Ultrathin film (UTF) resist layers, 117
Ultraviolet (UV) depth lithography
polyimide surface structures, 492–493
resists, 492–493

V

Vapor-phase etching, 199–201
Very large-scale immobilized polymer synthesis (VLSIPS), 525–527

W

Wafer cleaning
clean room classification system, 5–6
clean room contaminant sources, 5–6
supercritical cleaning, 7–8
Wafer priming, 41–42
Wagon-Wheel masking patterns, 236–238
Water jet etching, 229
Water jet machining
advantages, 372
laser machining, 338–339, 373
nonmetallic parts, 373
schematic illustration of, 373
Wet bulk micromachining
description, 232–233
etch-stop techniques (*see* Etch-stop techniques)
extensive real estate consumption
Bosch advanced porous silicon membrane process, 297
front etching, 296–297
silicon fusion-bonded (SFB), 297–298
high-aspect-ratio MEMS structures, 233
historical note, 216–218
single-crystal Si, empirical data
isotropic/anisotropic etches, 234–236
laser spoiling, 239
[100]-/[110]-oriented wafers, 238–239
Wagon-Wheel masking patterns, 236–238
undercutting
convex corner attacks, 299
corner compensation structures, 299–300
disadvantages, 302–304
underetching
disadvantages, 302–304
KOH etching, 300–301
V-groove widening, 298–299
Wet chemical etching
anisotropic etching
ammonium hydroxide-water/tetramethyl ammonium hydroxide-water, 254
Arrhenius plots, 255
description, 250
disadvantages, 251
etchants variety, 251
ethylenediamine pyrocatechol, 253–254
hydrazine (H_2N_4), 254–255
masking for, 256–257
potassium hydroxide, 252–253
principal characteristics, 251–252
procedure, 251
silicon wafer protection, 257
Si surface roughness, 255–256
electrochemical removal techniques
definition, 225
electrochemical jet etching, 229
grinding, 225–226
machining, 226–228
mask EMM, 230–231
metal removal rate determination, 231–232
microelectrochemical machining, 228–229
scanning electrochemical microscope (SECM), 230
etch-stop techniques (*see* Etch-stop techniques)
historical note, 216–218
isotropic etching
Arrhenius plot, 243
dopant dependence, 245–246
electrochemical, 246–250
iso-etch curves, 241–243
loading effects, 243–244
masking materials, 244–245
preferential, 246–247
risk of, 246
simplified reaction scheme, 241
uses, 240
milling techniques
chemical, 218–219
photochemical, 219–221
photofabrication
definition, 221
industrial application, 221–222
photosensitive glasses, 222–224
photosensitive plastics, 224–225
single-crystal Si
anisotropic etching, doping dependence, 271–273
atomistic model, Gosalvez–Kelly, 265–268
Elwenspoek et al.'s model, 274–277
Madou model, 277–280
Seidel et al.'s model, 268–271, 273–274
Wet vs.dry etching
dissolved wafer process, 305–306
electrochemical sensor array, 306–308
electrochemical valves, 308–309
environmental issues, 304–305
magnetic micro-optical matrix switch, 309–311
Wire electrode discharge grinding (WEDG), *see* Electrical discharge machining, wire cutting (EDM-WC)

X

Xerogel, 514
X-ray lithography
blurring (δ) expression, 102
electron-beam x-ray source, 102
features of, 98
LIGA, 98–100
light sources, 97
masks
vs. optical, 100–101
silicon membrane, fabrication of, 101
structure, 101
resists, 99–100
synchrotron, 97–98, 102–103
X-ray masks, LIGA process
absorber fabrication
CNC machined absorbers, 611
single-layer absorbers, 610–611
stepped absorber, 611

absorber materials, 609–610
alignment, 611–612
in IC industry, 607
membrane materials, 608–609
plastic molding and second electroforming, 606
substrate assembly in x-ray scanner, 607
transfer mask, 612–613
x-ray deep-etch lithography, 606
X-ray transfer mask
advantages, 613
cylindrical resonator, 612–613
process of, 612

Z

Zone plate array lithography (ZPAL)
coherent light expression, 122–123
Fresnel zone plates
focal length, 122
radius of nth zone, 122
resolution for coherent light, 123
maskless optical projection lithography (MOPL), 124
resolution, 124

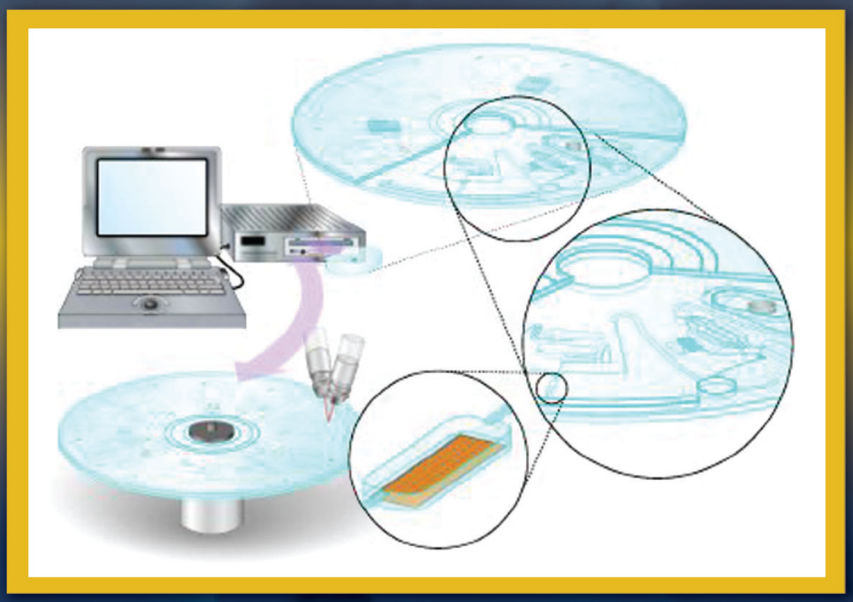
Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME III

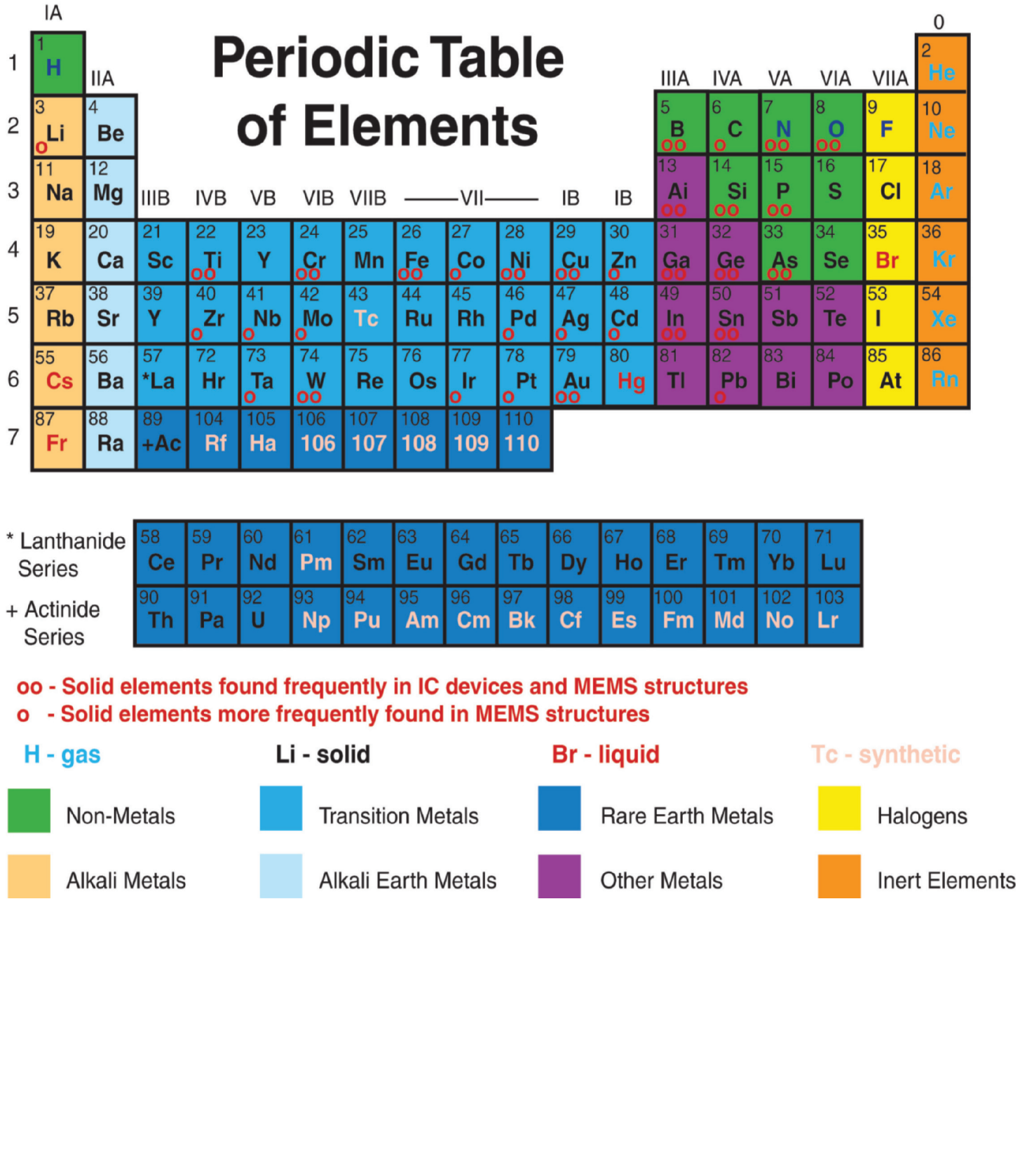
From MEMS to Bio-MEMS and Bio-NEMS

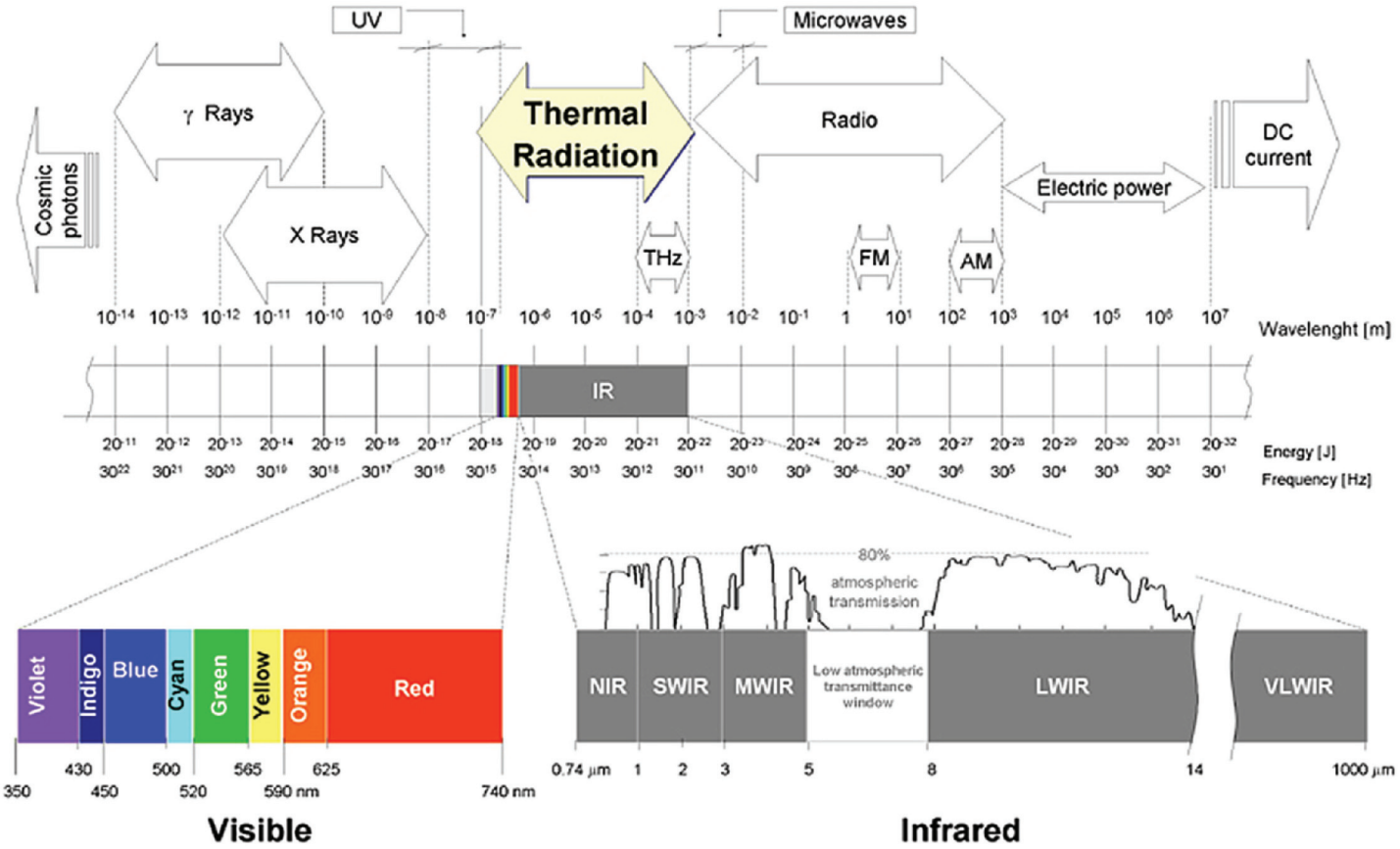
Manufacturing Techniques and Applications



Marc J. Madou

CRC Press
Taylor & Francis Group


Periodic Table of Elements
IA
IIA
IIIB
IVB
VB
VIB
VIIB
VII
IB
IB
IIIA
IVA
VA
VIA
VIIA
0
1 H
2 He
3 Li
4 Be
5 B
6 C
7 N
8 O
9 F
10 Ne
11 Na
12 Mg
13 Ai
14 Si
15 P
16 S
17 Cl
18 Ar
19 K
20 Ca
21 Sc
22 Ti
23 Y
24 Cr
25 Mn
26 Fe
27 Co
28 Ni
29 Cu
30 Zn
31 Ga
32 Ge
33 As
34 Se
35 Br
36 Kr
37 Rb
38 Sr
39 Y
40 Zr
41 Nb
42 Mo
43 Tc
44 Ru
45 Rh
46 Pd
47 Ag
48 Cd
49 In
50 Sn
51 Sb
52 Te
53 I
54 Xe
55 Cs
56 Ba
57 *La
72 Hr
73 Ta
74 W
75 Re
76 Os
77 Ir
78 Pt
79 Au
80 Hg
81 Tl
82 Pb
83 Bi
84 Po
85 At
86 Rn
87 Fr
88 Ra
89 +Ac
104 Rf
105 Ha
106 106
107 107
108 108
109 109
110 110
* Lanthanide Series
58 Ce
59 Pr
60 Nd
61 Pm
62 Sm
63 Eu
64 Gd
65 Tb
66 Dy
67 Ho
68 Er
69 Tm
70 Yb
71 Lu
+ Actinide Series
90 Th
91 Pa
92 U
93 Np
94 Pu
95 Am
96 Cm
97 Bk
98 Cf
99 Es
100 Fm
101 Md
102 No
103 Lr
oo - Solid elements found frequently in IC devices and MEMS structures
o - Solid elements more frequently found in MEMS structures
H - gas
Li - solid
Br - liquid
Tc - synthetic
Non-Metals
Transition Metals
Rare Earth Metals
Halogens
Alkali Metals
Alkali Earth Metals
Other Metals
Inert Elements


UV
Microwaves
γ Rays
Thermal Radiation
Radio
DC current
Cosmic photons
X Rays
Electric power
THz
FM
AM
Wavelenght [m]
IR
Energy [J]
Frequency [Hz]
80%
atmospheric transmission
Violet
Indigo
Blue
Cyan
Green
Yellow
Orange
Red
NIR
SWIR
MWIR
Low atmospheric transmittance window
LWIR
VLWIR
350
430
450
500
520
565
590 nm
625
740 nm
0.74 µm
1
2
3
5
8
14
1000 µm
Visible
Infrared

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME III

From MEMS to Bio-MEMS and Bio-NEMS

Manufacturing Techniques and Applications

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME III

From MEMS to Bio-MEMS and Bio-NEMS

Manufacturing Techniques and Applications

Marc J. Madou

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2012 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper

International Standard Book Number: 978-1-4200-5516-0 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

I dedicate this third edition of Fundamentals of Microfabrication to my family in the US and in Belgium and to all MEMS and NEMS colleagues in labs in the US, Canada, India, Korea, Mexico, Malaysia, Switzerland, Sweden and Denmark that I have the pleasure to work with. The opportunity to carry out international research in MEMS and NEMS and writing a textbook about it has been rewarding in terms of research productivity but perhaps even more in cultural enrichment. Scientists have always been at the frontier of globalization because science is the biggest gift one country can give to another and perhaps the best road to a more peaceful world.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Roadmap ix
Author xi
Acknowledgments xiii

PART I
From Traditional Manufacturing to Nanotechnology

Introduction to Part I 2

1 Nonlithography-Based (Traditional) and Lithography-Based (Nontraditional) Manufacturing Compared 3

2 Nature as an Engineering Guide: Biomimetics 49

3 Nanotechnology: Top-Down and Bottom-Up Manufacturing Approaches Compared 117

4 Packaging, Assembly, and Self-Assembly 185

5 Selected Materials and Processes for MEMS and NEMS 239

6 Metrology and MEMS/NEMS Modeling 285

PART II
Scaling Laws, Actuators, and Power and Brains in Miniature Devices

Introduction to Part II 358

7 Scaling Laws 359

8 Actuators 387

9 Power and Brains in Miniature Devices 475

PART III
Miniaturization Application

Introduction to Part III 526

10 MEMS and NEMS Applications 527

Index 627

Roadmap

From MEMS to Bio-MEMS: Manufacturing Techniques and Applications consists of ten chapters in three parts. In Part I, we compare available manufacturing options for microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS), introduce inspection options for the produced parts (metrology), and summarize available modeling software for MEMS and NEMS. Nonlithography-based (traditional) manufacturing techniques are contrasted with lithography-based (nontraditional) methods in Chapter 1. In Chapter 2, we investigate nature as an engineering guide, and we contrast top-down and bottom-up approaches in Chapter 3. In Chapter 4 we learn about packaging, assembly, and self-assembly from integrated circuits to DNA and biological cells. In these chapters we aim to help the reader decide upon an optimized manufacturing option to tackle specific manufacturing problems. We then introduce some selected new MEMS and NEMS processes and materials in Chapter 5. Finally, we cover metrology techniques for MEMS and NEMS and summarize MEMS and NEMS modeling in Chapter 6.

Of the three chapters in Part II, the first deals with scaling laws, the second with actuators, and the third with issues surrounding power generation and the implementation of brains in miniaturized devices. In Chapter 7, on scaling laws, we look at scaling from both intuitive and mathematical points of view. We are especially interested in deviations from linear scaling, where downscaling reveals new unexpected physical and chemical phenomena. The scaling chapter also constitutes an introduction to the subsequent treatises on actuators in Chapter 8 and on power and brains in Chapter 9. An actuator—like a sensor—is a device that converts energy from one form to another. In the case of an actuator, one is interested in the ensuing action; in the case of a sensor, one is interested in the information garnered. Scaling enables one to compare various actuator mechanisms, such as those proposed to propel a rotor blade, bend a thin silicon membrane, or move fluids in fluidic channels. Power generation poses quite a challenge for miniaturization science; the smaller the power source, the less its total capacity. We will consider the powering of miniaturized equipment, as well as the miniaturization of power sources themselves, in Chapter 9. In the same chapter, we also discuss the different strategies for making micromachines smarter.

In Part III, Chapter 10, we review the MEMS and NEMS markets. Today, the MEMS field has made the transition out of the laboratory and into the marketplace with a plethora of new products, including a strong entry into the consumer electronics field (mobile phones, cameras, laptop computers, games, etc.). If MEMS is entering adolescence, the word *nano* still seems to derive from a verb that means "to seek and get venture capital funding" instead of from the Greek noun for dwarf. Today, the market dynamics for MEMS and NEMS commercial products are very different, and we treat MEMS and NEMS applications separately in this chapter.

As highlighted in Chapter 1, which compares nonlithography-based (traditional) and lithography-based (nontraditional) manufacturing, we hope that a better understanding of how to match different manufacturing options with a given application will guide the identification of additional killer applications for MEMS and NEMS and encourage more companies and research organizations to innovate faster based on their in-house manufacturing tools and know-how.

Note to the Reader: *From MEMS to Bio-NEMS: Manufacturing Techniques and Applications* was originally composed as part of a larger book that has since been broken up into three separate volumes. *From MEMS to*

Bio-NEMS: Manufacturing Techniques and Applications represents the third and final volume in this set. The other two volumes include *Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology* and *Manufacturing Techniques for Microfabrication and Nanotechnology*. Cross-references to these books appear throughout the text and will be referred to as Volume I and Volume II, respectively. The interested reader is encouraged to consult these volumes as necessary.

Author

Dr. Madou is the Chancellor's Professor in Mechanical and Aerospace Engineering (MEA) at the University of California, Irvine. He is also associated with UC Irvine's Department of Biomedical Engineering and the Department of Chemical Engineering and Materials Science. He is a Distinguished Honorary Professor at the Indian Institute of Technology Kanpur, India, and a World Class University Scholar (WCU) at UNIST in South Korea.

Dr. Madou was Vice President of Advanced Technology at Nanogen in San Diego, California. He specializes in the application of miniaturization technology to chemical and biological problems (bio-MEMS). He is the author of several books in this burgeoning field he helped pioneer both in academia and in industry. He founded several micromachining companies.

Many of his students became well known in their own right in academia and through successful MEMS start-ups. Dr. Madou was the founder of the SRI International's Microsensor Department, founder and president of Teknekron Sensor Development Corporation (TSDC), Visiting Miller Professor at UC Berkeley, and Endowed Chair at the Ohio State University (Professor in Chemistry and Materials Science and Engineering).

Some of Dr. Madou's recent research work involves artificial muscle for responsive drug delivery, carbon-MEMS (C-MEMS), a CD-based fluidic platform, solid-state pH electrodes, and integrating fluidics with DNA arrays, as well as label-free assays for the molecular diagnostics platform of the future.

To find out more about those recent research projects, visit http://www.biomems.net.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Acknowledgments

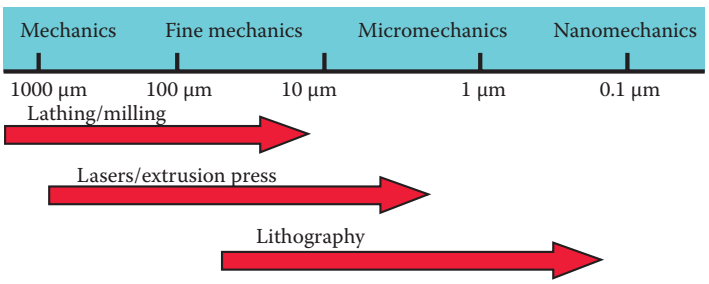
I thank all of the readers of the first and second editions of Fundamentals of Microfabrication as they made it worthwhile for me to finish this completely revised and very much expanded third edition. As in previous editions I had plenty of eager reviewers in my students and colleagues from all around the world. Students were especially helpful with the question and answer books that come with the three volumes that make up this third edition. I have acknowledged reviewers at the end of each chapter and students that worked on questions and answers are listed in the questions sections. The idea of treating MEMS and NEMS processes as some of a myriad of advanced manufacturing approaches came about while working on a WTEC report on International Assessment Of Research And Development In Micromanufacturing (http://www.wtec.org/micromfg/report/Micro-report.pdf). For that report we travelled around the US and abroad to visit the leading manufacturers of advanced technology products and quickly learned that innovation and advanced manufacturing are very much interlinked because new product demands stimulate the invention of new materials and processes. The loss of manufacturing in a country goes well beyond the loss of only one class of products. If a technical community is dissociated from manufacturing experience, such as making larger flat-panel displays or the latest mobile phones, such communities cannot invent and eventually can no longer teach engineering effectively. An equally sobering realization is that a country might still invent new technologies paid for by government grants, say in nanofabrication, but not be able to manufacture the products that incorporate them. It is naïve to believe that one can still design new products when disconnected from advanced manufacturing: for a good design one needs to know the latest manufacturing processes and newest materials. It is my sincerest hope that this third edition motivates some of the brightest students to start designing and making things again rather than joining financial institutions that produce nothing for society at large but rather break things.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Part I

From Traditional Manufacturing to Nanotechnology

In ten years the only manufacturing left in the United States will be 1) those facilities vital to the defense industry, 2) those industries that are uniquely high-tech, 3) those that cannot absorb long-distance freight charges, and 4) those industries that service "on the spot" instantaneous demand (although even that is questionable).



From macromachining to nanomachining.

Introduction to Part I

Chapter 1 Nonlithography-Based (Traditional) and Lithography-Based (Nontraditional) Manufacturing Compared
Chapter 2 Nature as an Engineering Guide: Biomimetics
Chapter 3 Nanotechnology: Top-Down and Bottom-Up Manufacturing Approaches Compared
Chapter 4 Packaging, Assembly, and Self-Assembly
Chapter 5 Selected Materials and Processes for MEMS and NEMS
Chapter 6 Metrology and MEMS, NEMS Modeling

Introduction to Part I

In Part I, we compare available manufacturing options for microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS), introduce inspection options for the produced parts (metrology), and summarize available modeling software for MEMS and NEMS. Nonlithography-based (traditional) manufacturing techniques are contrasted with lithography-based (nontraditional) methods in Chapter 1. In Chapter 2, we investigate nature as an engineering guide, and we contrast top-down and bottom-up approaches in Chapter 3. In Chapter 4 we learn about packaging, assembly, and self-assembly from integrated circuits (ICs) to DNA. In these chapters, we aim to help the reader decide upon an optimized manufacturing option to tackle specific manufacturing problems. We then introduce some selected MEMS and NEMS processes and materials in Chapter 5. Finally, we review metrology techniques for MEMS and NEMS and summarize MEMS and NEMS modeling in Chapter 6.

1

Nonlithography-Based (Traditional) and Lithography-Based (Nontraditional) Manufacturing Compared

Outline

Introduction

Why Use Miniaturization Technology?

From Perception to Realization

Beyond Si-Based MEMS

Serial, Batch, and Continuous Manufacturing Options

Comparison Table

Design from the Package Inward

Decision Tree for the Optimized Micromanufacturing Option

Scaling Characteristics of Micromechanisms

Partitioning

Fluid Propulsion Methods and MEMS Integration

Heating and Cooling and MEMS Integration

Sample Introduction

Manufacturing Technique of the Disposable Cassette

Substrate Choice

Checklist

Questions

References

1995–02, America lost 2 million industrial jobs, mostly to China. China lost 15 million of such jobs, mostly to machines.

Fortune

Despite the shrinking of America's industrial work force, our country's overall industrial output increased by 50% since 1992.

Economist

Charlie Chaplin's *Modern Times*.

Introduction

Micromachining, or microelectromechanical system/nanoelectromechanical system (MEMS/NEMS), is emerging as a set of new manufacturing tools to solve specific industrial problems rather than as a monolithic new industry with generic solutions for every manufacturing problem. It is important to evaluate the merit of using one certain MEMS/NEMS technique above all the other available micromanufacturing options [say, lithography-based LIGA* (i.e., nontraditional manufacturing) vs. nonlithography computer numerical control machining (CNC) (i.e., traditional manufacturing)] so as to find the technique that is optimal for the application at hand—in other words, one needs to zero-base the technological approach to the problem.[1] For example, micromachinists are often not aware of the capabilities of traditional machining (nonlithography) and use Si micromachining for parts that could have been made better with a more conventional manufacturing technology. By applying the right tool to the machining job at hand, we hope that micromachining will lead to many more successful commercial applications than there are today.

We start this chapter by listing reasons why one might want to miniaturize a given device at all. We follow this with a short treatise on the maturation of MEMS into a bona fide manufacturing technique that is complimentary today to traditional manufacturing methods. We go on contrasting serial, batch, and continuous manufacturing processes and distinguish between lithography-based (nontraditional) and nonlithography-based (traditional) manufacturing methods, capturing all of this in a large comparison table. In this same context, we explore the difference between truly three-dimensional (3D) manufacturing with equal versatility of machining along all three axes and the more constrained manufacturing methods, such as lithography-based techniques, where the capabilities in one dimension (say the height or z-axis) are very different from those in the other two dimensions (say the x, y plane). We then justify our recommendation to design MEMS from the package inward rather than starting from the MEMS itself. Finally, we introduce a decision tree to help determine the best manufacturing choice for a miniature device when given its detailed specifications. The utility of this decision tree is demonstrated using, as an example, the micromanufacture of a disposable glucose sensor with glucose meter.

* LIGA is a German acronym for "**Li**thographie, **G**alvanoformung, **A**bformung," which in English is (x-ray) lithography, electroplating, and molding. It is a process in MEMS/NEMS that was developed in the early 1980s by a team under the leadership of E. Becker and W. Ehrfeld at the Institute for Nuclear Process Engineering [*Institut für Kernverfahrenstechnik* (IKVT)] at the Karlsruhe Nuclear Research Center. For details, Volume II, Chapter 10.

Why Use Miniaturization Technology?

Over the last twenty years, many MEMS applications have proven their mettle in the marketplace, and consumer electronic applications that started emerging over the last five years (MEMS are now found in iPods, computers, cameras, GPSs, etc.) have catapulted MEMS once more in the public eye. A long list of reasons confirms why miniaturization presents so many opportunities for product innovation in so many different areas. Some of the most obvious reasons for miniaturization are summarized in Table 1.1. Usually, not all those reasons apply at once. For example, the small dimensions of micromachines might be crucial in medical and space applications but often lack importance in the automotive industry, where cost is the more important driver.

Table 1.1 Why Use Miniaturization Technologies?

• Minimizing of energy and materials consumption during manufacturing
• Redundancy and arrays
• Integration with electronics, simplifying systems (e.g., single-point vs. multipoint measurement)
• Reduction of power budget
• Taking advantage of scaling when scaling is working for us in the microdomain, e.g., faster devices, improved thermal management
• Increased selectivity and sensitivity
• Minimal invasiveness
• Wider dynamic range
• Exploitation of new effects through the breakdown of continuum theory in the microdomain
• Cost/performance advantages
• Improved reproducibility
• Improved accuracy and reliability
• Self-assembly and biomimetics with nanochemistry
• More intelligent materials with structures at the nanoscale

In Volume I, Chapter 1, when analyzing Moore's law describing the exponential growth in integrated circuit (IC) transistor density, we pointed out that similar exponential improvement in technology sophistication pertains to engineering skills in optics, genetics, and magnetic storage density. Humankind seems "hard-wired" to continue to innovate at an ever-increasing rate (see also Ray Kurzweil's Accelerating Returns, in Volume II, Chapter 2). Miniaturization is just one of the most important means by which humankind is shaping its own destiny rather than being at the mercy of evolution. Adaptation through evolution operates at a pace too slow to produce humans who thrive in the ever faster environmental changes they have created for themselves. In other words, we might have no choice but to adapt through ever-increasing technology sophistication rather than our DNA.

From Perception to Realization
MEMS Finally Succeeds in the Market

In the early 1980s, Si micromachining seemed often applied to show the world that University X also had a clean room or that research group Y could make a yet longer-lasting surface micromachined micromotor. The earlier predicament of Si-based micromachining as the "GaAs of the 1980s and early 1990s" [that is, a very good technology with little acceptance from an entrenched and highly standardized industry (think also Mac vs. PC)] had something to do with the intellectual/philosophical climate of silicon MEMS research and development in the early 1980s. Francis Bacon (1561–1626) explained how, in the advancement of knowledge, one is easily misled by what he calls "idolatry." Two of the "idols" he identifies point in the direction of a universal tendency to oversimplify, often manifested by the assumption of more order in a given body of phenomena than actually exists, and a tendency to be struck by novelty.[2]

Both idols apply to how Si micromachining became misrepresented. Little commonality exists between the many different microdevices made possible by micromachining. Actually, MEMS products represent discontinuous innovations—they are not simply incremental improvements on a previously existing technology. Each new product requires some new thinking on the part of the developers. The striking visual aspects of Si micromachined devices easily give a sense of novelty to any observer. This second idol drove most of the interest in micromachining by the popular press and the efforts of many academics. This climate contributed to overly optimistic expectations for very fast results, with a market size dwarfing the IC industry, and an overemphasis on Si as an answer to all miniaturization problems. In Chapter 10, on MEMS and NEMS applications, we suggest that a realistic number for the market of Si MEMS products today is less than 3% of the IC market (even the $6.3 billion MEMS sales in 2006, i.e., 3% of the $211 billion IC market for 2006, does include non-Si devices). Subsequently, government and industry funding sources experienced a hangover.

Bacon's idols aside, the very large academic involvement in the micromachining field had some other explanations. The IC industry technologically and financially outdistanced universities and forcibly pushed the latter to explore topics requiring lower startup expenses and areas where innovation still seemed likely. It was a perfect fit, as micromachining is an excellent topic for numerous PhD topics. In the 1980s, Si micromachining became a favored filler for clean room overcapacity, especially in Europe, where the IC industry suffered major setbacks against US and Japanese competition. By the end of the century, it was apparent that the Si micromachining industry was still very small compared with the IC industry (less than 3%) but had large numbers of people involved in its research and development. Middlehoek and Dauderstadt estimated, as early as 1994, that about 10,000 researchers worldwide were involved in Si sensor research and that $7.5 billion had been spent over the preceding twenty-five years.[3] Many MEMS researchers have now switched to nanotechnology (NEMS) research instead.

Another major shortcoming in Si MEMS, as practiced in academia, pertains to the lack of interdisciplinary teams. More than any other field, microengineering requires interdisciplinary teamwork rather than specialists. Multidisciplinary work from the design phase on is crucial in the

development of a successful sensor product. Many of the participating university groups characterize themselves as multidisciplinary, but the opposite is often true; this lack of multidisciplinary focus continues to obstruct the production of more practical results. It reminds us of a statement by C.P. Snow (1905–1980), who remarked that the separate departments at a university had ceased to communicate with one another. The prefix *uni* in university, he said, lost its meaning as the institution, striving for more and more power as government funds replenished research dollars, had turned into a loose confederation of disconnected mini-states instead of an organization devoted to the joint search for knowledge and truth.

For micromachining to lose the stigma of a technology looking for an application, used in research backrooms and excess industrial clean rooms only, it became important for MEMS to realize some product successes. These industrial successes started appearing principally in the early to mid-1990s and include pressure sensors (on a small scale as early as 1972 at National Semiconductor), accelerometers (1992, from Analog Devices), projection displays (digital mirror devices from TI in 1995), ink-jet nozzles (thermal ink-jet from HP, 1984), electrokinetic platforms (Caliper and Agilent, 1999), fiber optic switches (NTT, 1995), and the first digital mirror device-based products (Texas Instruments, 1996). The first mechanical MEMS applications in industry were modest success stories only, with dollar evaluation in initial public offerings for mechanical MEMS companies typically only in the $5 million to $45 million range. Ever since biotechnology and information/communication industries started taking an interest in MEMS, huge stock market success stories began to appear, and large corporations bought small MEMS companies for billions of dollars. But today, many investors still cringe when they hear the terms *MEMS* or *NEMS*, because of the ill-fated optical MEMS bubble of the late 1990s and a BIOMEMS field that never seems to live up to its promises.

Fortunately, as we will learn in Chapter 10, on MEMS and NEMS applications, mechanical and optical MEMS in consumer electronics, started taking off over the last five years, and even MEMS for large optical switches in telecommunications is now regaining a foothold. The real MEMS breakthroughs continue to originate overwhelmingly from start-up companies (e.g., Illumina and Cepheid) and large corporations (e.g., ADI and Bosch) rather than academia. Micromachining in industry is more correctly seen as a diverse set of tools to solve practical problems in the crafting of subminiaturized 3D structures rather than as a goal by itself. The multidisciplinary approach and matrix organization of MEMS teams in industry further explains their more impressive MEMS accomplishments.

In academia, with its very vertical departmental structure and isolation from practical manufacturing and application understanding, miniaturization science is still often hyped beyond recognition in order to secure grants, tenure, or money from government agencies, large corporations, and venture capitalists. Describing academic MEMS efforts as unbiased fundamental work or fine-tuning of micromachining skills, rather than proclaiming each new result as the latest breakthrough in sensor technology or analytical chemistry, would often serve society better.

The new wave of nanomachining and nanochemistry efforts, mostly grounded in the fundamental research phase at this stage, holds the promise that university scholars might regain the miniaturization playing field in the near future. We can only hope that the hype surrounding nanotechnology settles down and that funding agencies and companies give researchers and scientists enough time to realize the stupendous opportunities that nanotechnology holds. To realize the economic potential of MEMS and NEMS, society will need a new generation of engineers who are comfortable with chemical and biological issues as well as many different types of micro- and nanomanufacturing considerations. Academia will have to become more matrix organized, if only to prepare the proper workforce for the coming age of nanomachining. From the fundamental point of view, micromachining and nanomachining are also of extreme interest, as these fields provide an excellent opportunity for operating experiments in a regime where continuum theory breaks down, holding the potential for discovery of important new chemical and physical phenomena. Unfortunately,

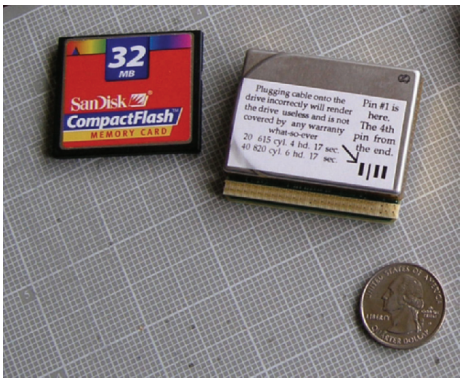
mounting funding pressures in the United States are forcing scientists to sell research as development and development as manufacturing. By putting their work in the context of a practical solution too early, academics might have lost some of industry's trust in the capabilities of micromachining and nanomachining. A breakthrough in deep, cryogenic etching or yet thicker vertical photoresist walls should stand by itself and does not necessarily need to be illustrated with the fabrication of wheels for micromotors! A better practice in the future might be to compare these new academic results (say, on new deep UV photoresists) with other, more traditional approaches so that the work becomes of generic value to a potential user for his or her own specific application.

Beyond Si-Based MEMS

New Si-based MEMS products are only a small part of the total projected MEMS markets, and there is ample room for non-Si-based companies in the MEMS markets of the future (e.g., microplastic molders, microelectroplaters, microceramic parts producers, etc.). Market predictions from the 1980s and early 1990s about the Si MEMS revolution were very mistaken; it is only by broadening the definition that MEMS is living up to the numbers that were quoted at that time. In Chapter 10 of this volume on MEMS and NEMS markets, we will see that when including non-Si MEMS, which many market studies do include now, and especially when also counting nonlithography-based MEMS (e.g., precision engineered microproducts; see below) (which very few market studies do yet), MEMS might very well constitute 10% of all current IC sales.

The market does not care about the type of miniaturization method employed for a new product, and for some time the Si micromachining community, especially in the United States, did not appreciate the tremendous accomplishments of alternative miniaturization methods. The HP Kittyhawk Personal Storage Module, a 1.3-inch disk drive, a compact disc (CD) optical pickup, and IBM's thin-film magnetic read-write head shown in Figure 1.1 are triumphs of precision engineering, and none involve Si-MEMS. The Kittyhawk (Figure 1.1a), introduced by HP in 1992, never caught on and seems to have been introduced before its time (Sony has continued its efforts on a similar MiniDisc). Kittyhawk was discontinued in September 1994. Approximately 160,000 units were sold (compared with a projected two-year sales of 700,000 units). In 1996, largely because of Kittyhawk's failure, Hewlett Packard closed its Disk Memory Division and exited the disk drive business. (The story of HP Kittyhawk was described in a Jan 26, 2006 Harvard Business School business case, "Hewlett-Packard: The Flight of the Kittyhawk."[64]) CD optical pickups and CD technology in general are other marvels of precision engineering, funded by billions of dollars in research and development (R&D) at Philips and Sony (see Figure 1.1b and 1.1c). This technology has not been exploited nearly as well as it should outside its traditional application area. We envision many nontraditional applications for this platform, such as its use as a sample preparation station and a molecular diagnostic tool as explored in this chapter, in Volume I, Chapter 6, and in Chapter 5, this volume. In diagnostic applications, the same CD technology may be used to read both analytical data (e.g., as a color change of a chromophore) and information (in the form of pits on the CD), to serve as a very inexpensive microscope (see Volume I, Figures 6.93 and 6.94) or as a high-precision x-y stage, etc. In this respect, one could legitimately ask the question, why are optical pickup heads not included in MEMS market studies when hard disc drive read-write heads often are? (See Chapter 10, this volume.) The answer is that these choices were made quite arbitrarily, underscoring the need to identify the criteria/definitions of MEMS and the manufacturing means used to make the microcomponents covered in any MEMS marketing report. The thin-film magnetic heads, shown in Figure 1.1d, are an excellent example of the commercial success one can achieve when identifying the correct machining option and materials combinations for a miniaturization task at hand. At $10 billion in sales in 2007, the thin-film magnetic head is still the most successful MEMS product today. The manufacture of the thin-film head is principally based on photolithography and electroplating of the coils and does not involve any Si. In 2009, read-write (RW) heads are still projected to have a market the size of all other Si MEMS

(a)



(b)

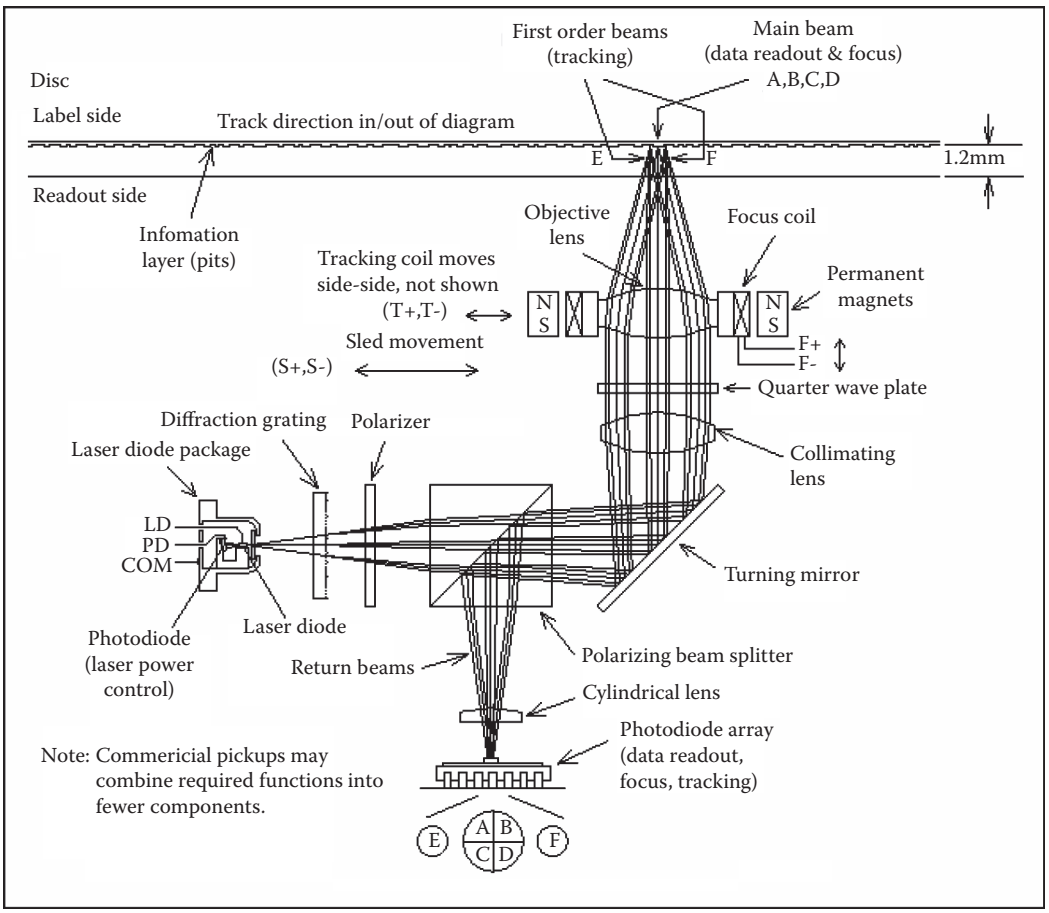
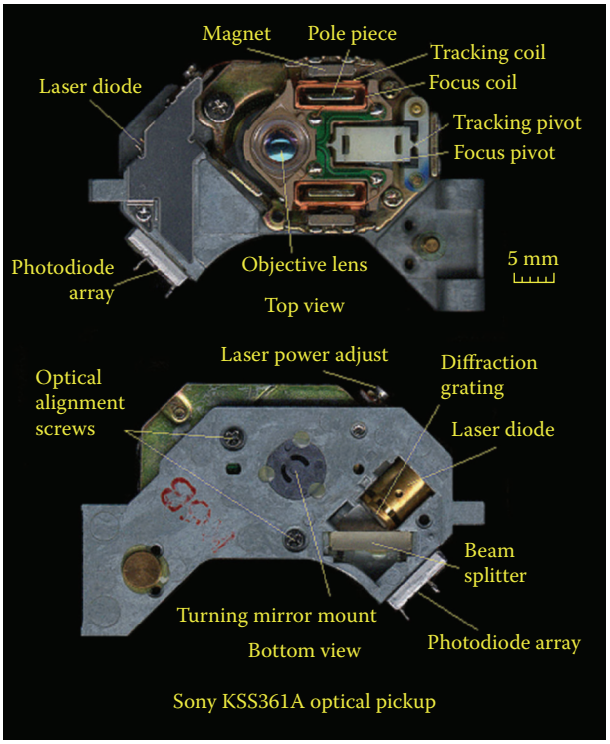


FIGURE 1.1 Triumphs of mechatronics. (a) HP's Kittyhawk. (From http://www.homebrewcpu.com/Pictures/kitty-hawk.jpg.) (b) CD optical pickup. General three-beam optical pickup organization. (From http://www.repairfaq.org/REPAIR/ F_cdfaq1.html.) (c) Physical realization embodied by the Sony KSS361A Optical Pickup (see http://www.repairfaq.org/REPAIR/kss110c.gif). (d) The evolution of magnetic read-write heads. From 1980 on, the magnetic coils were made by semiconductor-type technology (lithography and Cu plating) (http://dspace.cusat.ac.in/dspace/bitstream/123456789/2121/1/ GIANT%20MAGNETORESISTANCE%20EFFECT.pdf.)

applications put together (see RW in Figure 10.15, this volume).

Because there are few established engineering rules and practices available to decide upon an ideal manufacturing technique from the wealth of options available for a given microapplication, good matches of machining methodology and application are often missed or rediscovered. Figure 1.2 illustrates the evolution of the types of manufacturing methods attempted for the construction of microfluidic elements. In the 1960s, microfluidic elements were principally made from photoglass; after attempting

(c)



(d)

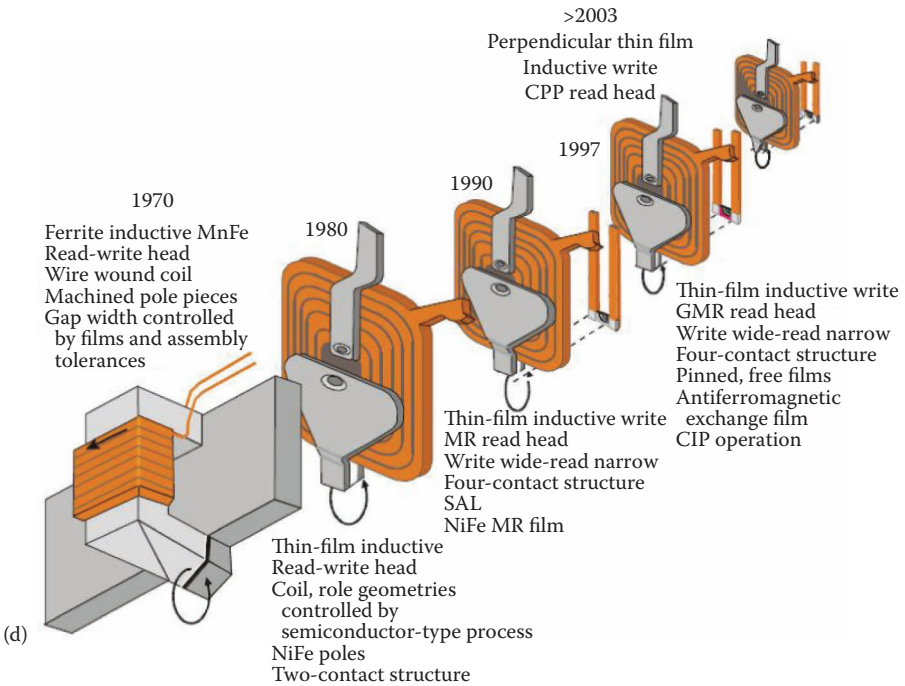


FIGURE 1.1 (*Continued*)

to craft the same devices by dry etching in Si in the 1980s at Stanford, LIGA was employed in the early 1990s. But by the mid-1990s, photoglass was rediscovered as the best machining option for some microfluidics applications. One might hope that by following a decision process flow researchers can avoid these costly loss and rediscovery cycles in the future.

As pointed out above, recent market studies cover Si and non-Si technology miniaturization applications, thereby achieving earlier projected large revenues that were expected to derive from Si micromachining alone. Some early market data on Si-based miniaturization were misleading in another way: they included big commercial successes

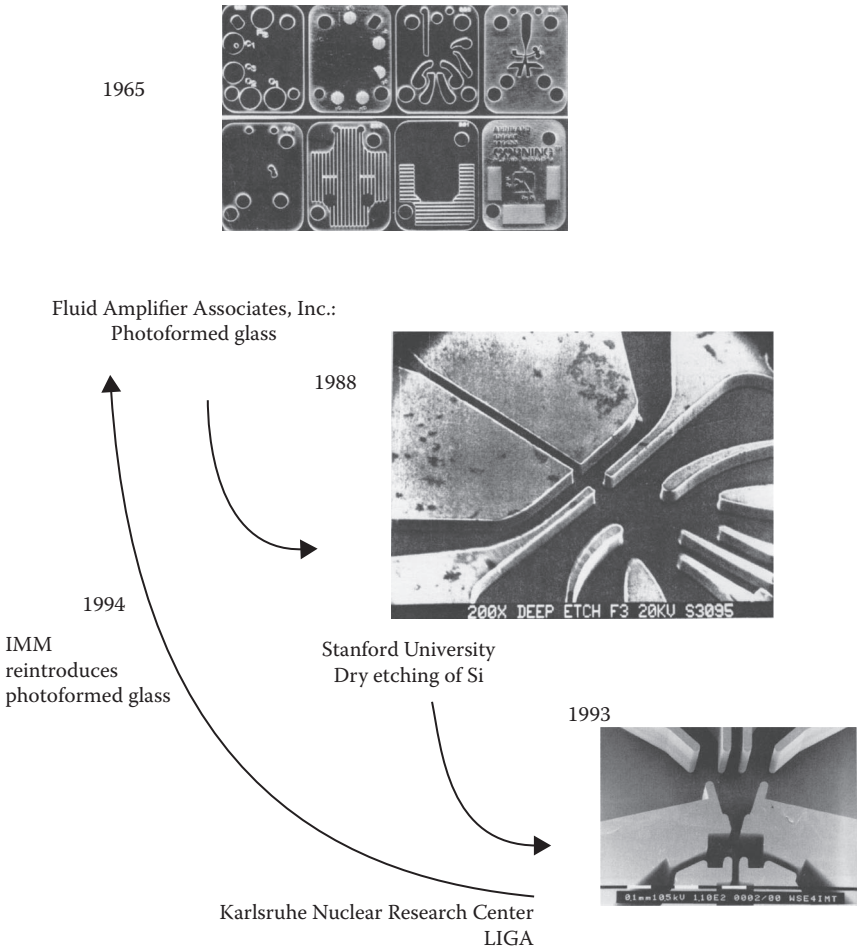


FIGURE 1.2 The rediscovery of photoformed glass for the manufacture of microfluidics.

in Si sensors such as simple photodiodes, charge-coupled detectors (CCDs), and Hall elements. The latter were, for the most part, developed before 1970 and are almost exclusively based on standard Si manufacturing techniques.

Serial, Batch, and Continuous Manufacturing Options

In serial manufacturing, a manufacturing station produces one product at the time. As an example for serial machining consider focused ion-beam milling, a mechanical removing technique detailed in Volume II, Chapter 6. Vasile et al., for example, describe a lathe mounted 20-keV Ga ion beam as the cutting tool for in-vacuum micromachining.[4,5] The ion beam they use is 0.3 μm in diameter, and they crafted tungsten needles, hooks, tuning forks, and specialized scanning probe tips. Two of their tungsten microstructures are shown in Volume II, Figure 6.33. Start-to-finish fabrication time for one of the objects shown here is about 2 hours! It should be noted that this expensive serial manufacturing technique might not make much sense for mass production of inexpensive consumer products but could be justified, for example, for the fabrication of an atomic force microscope probe tool set that could not be produced otherwise. On the other hand, to make a microshovel, as also shown in Volume II, Figure 6.33, is merely a fancy demonstration of the capabilities of this technique (call it a micro-"BS shovel"*).

In batch fabrication, a "group" or "batch" of identical products is produced in one single machining step. Batch fabrication techniques lend themselves to economies of scale that are unavailable with serial techniques. The use of batch processes is widespread throughout the manufacturing industry, producing a diverse range of products, including ICs,

* BS used to stand for *bullshit*; now it stands for *Bush Shit*.

pharmaceuticals, polymers, biochemicals, food products, and specialty chemicals. In the more advanced manufacturing techniques used for micromachining and nanomachining, batch fabrication typically involves some sort of a mask containing a desired repeated pattern that is positioned between the tool and work-piece. Then, that mask design is transferred to the work-piece using, for example, light (as in photolithography; Volume II, Chapter 1) or an abrasive slurry (as in abrasive jet machining; Volume II, Chapter 6). Alternatively, the tool itself contains an array of patterns that is transferred to the work-piece, for example, mechanically, through a slurry (as in ultrasonic machining; Volume II, Chapter 6), or thermomechanically (as in hot embossing; Volume II, Chapter 10). Most nontraditional micromachining carried out today, such as IC fabrication, relies on batch fabrication; that is, repetitive features are simultaneously defined on a work-piece. In the case of IC fabrication, an array pattern is photolithographically defined on a silicon wafer, and many wafers are then processed further to fabricate desired structures. We present as an example of batch fabrication, an array of finished CCD image sensors (to use, for example, in a digital camera), produced all in parallel on the same Si wafer, as shown in Figure 1.3. After production, these sensors need to be cut out of the Si wafer.

It is important to recognize that the use of a mask severely restricts the machining possibilities in the direction perpendicular to the mask. The resulting products, when using a mask, are projected shapes of the image on the mask only, with limited opportunities to shape the sidewalls of the work-piece perpendicular to the image on the mask. To illustrate this, think about precision mechanical machining, which is truly 3D since a diamond tool can cut in any direction, and contrast this with using a photomask in photolithography or a patterned tool in ultrasonic machining, where material removal forces work only in one direction and the resulting shapes are principally 2D. In the latter case, the dimension perpendicular to the mask may have differing heights (and indeed, we could say that we are dealing with a 2½D structure), but the shape of the sidewall can be modified only in a separate machining step.

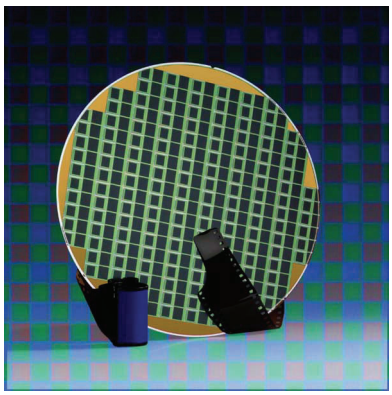


FIGURE 1.3 Charge-coupled detector (CCD) image sensors are formed on Si wafers and then cut apart. CCDs have replaced film as the image capture medium of preference. (Courtesy of IBM.)

Beyond batch manufacturing, the ultimate in automation of manufacturing is continuous or web-based manufacturing, with a web moving from supply roll to take-up roll [also called roll-to-roll (R2R) manufacturing*], with visions of low-cost photovoltaics (PVs), thin-film transistors (TFTs), large-area organic light-emitting diode (OLED) lighting, flexible displays, low-cost sensors, radio frequency identification (RFID) tags, etc. (see Figure 1.4). In this case, the product is made continuously, small or large, on flexible substrates, as shown in Figure 1.5. The flexible substrates replacing glass or Si must possess the desirable optical properties, thermal stability, low coefficient of thermal expansion, surface smoothness, and moisture and solvent resistance. Today, polymer films and stainless steel are the two flexible substrates of choice. R2R manufacturing dramatically increases throughput, with a significant reduction of capital and device costs.

Gordon Moore has argued that the only industry "remotely comparable" in its rate of growth to the semiconductor industry is the printing industry. It is worth noting that newspaper printing constitutes a continuous process; here, we promote the idea of continuous (beyond batch) miniaturization processes that, like a printer connected to a personal computer, will print out miniature objects in a continuous mode. Many of the manufacturing techniques we encountered in Volume II are already or can be modified to a roll-to-roll operation. This includes older

* Not to be confused with "run-to-run" or "run-by-run" control in semiconductor manufacturing.

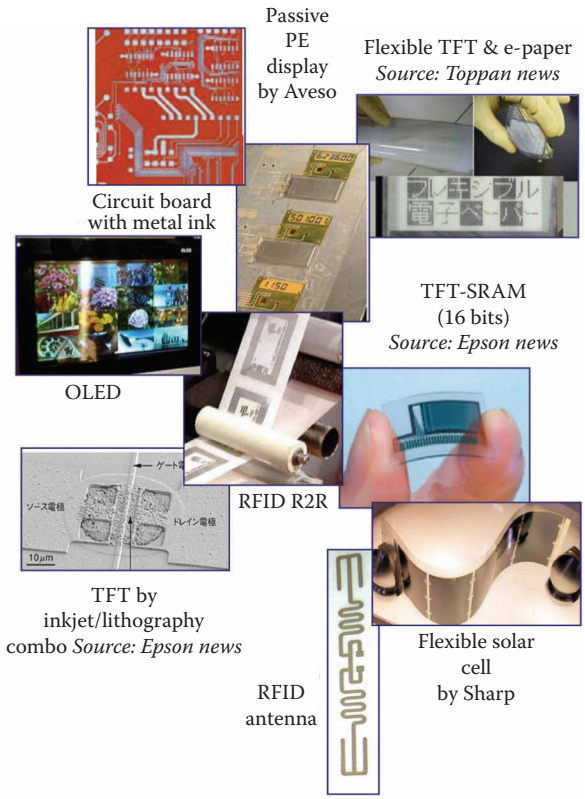


FIGURE 1.4 Examples of printed products on flexible substrates. For acronyms such as RFID, OLED, and TFT, please refer to the text.

technologies, such as doctor blade, ink-jet printing, laser machining, lamination, punching, drying and curing (in roll format ovens), picking and placing, hot-embossing, and sputtering (with cylindrical targets; see Volume II, Figure 7.12), and newer methods, such as imprint lithography [nanoimprint lithography (NIL) and step-and-flash imprint lithography (SFIL)], dip-pen lithography (DPL), etc. Soft lithography (patterning with flexible stamps; see Volume II, Figure 2.55) constitutes yet another possible means of transferring lithography patterns to a web-based process. In this case, the soft lithography pattern on the rubber stamp might be rolled up around a cylinder to ink a substrate continuously.

Processes, such as lithography, that need critical alignment and registration are the more challenging to incorporate in an R2R manufacturing process. But roll-to-roll or in-line equipment is also available for lithography. For illustration, in Figure 1.6, we show the lithography equipment procured by The Center for Advanced Microelectronics Manufacturing at Binghamton University in New York (http://camm.binghamton.edu). Established in 2005, this center is meant to spearhead development of next generation R2R electronics manufacturing capabilities. In Figure 1.6a we show a wet stripper/developer from Hollmuller-Siegmund that can handle a web of a width up to 15 in. The in-line lithography setup (Figure 1.6b) is from Azores Corporation and can handle webs that are 8 in. wide (expandable to 24 in.). This g-line machine comes with an accuracy of about 4 µm and can process photosensitive material at a speed of 230–760 mm/min. Hole-punch patterns are used for alignment. Many governments are taking notice of this important trend in electronics manufacturing. In Korea, for example, the Department of Printed Electronics Engineering in the World Class University (WCU) program of Sunchon National

FIGURE 1.5 Roll-to-roll (R2R) manufacturing. Production in kilometers and use and sale in centimeters.

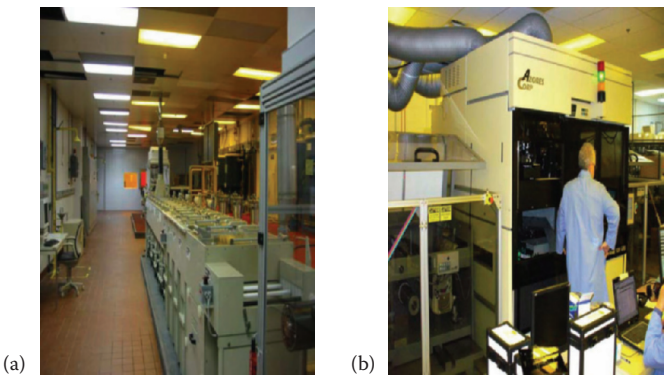


FIGURE 1.6 In-line lithography equipment. (a) wet stripper/developer from Hollmuller-Siegmund (can handle a web up to 15 in. wide). (b) in-line lithography setup from Azores Corporation can handle webs that are 8 in. wide (expandable to 24 in.).

University has been launched to lead in the fields of printed RFID and flexible lighting through Korean Government support that started in 2009.

Advances are being made continuously in "ink" technologies and printing processes and control; organic and printable electronics are expected to be a $35 billion industry by 2015 and reach over $300 billion in 2025, that is, almost twice the size of the Si industry today.

In Example 1.1, at the end of this chapter, we introduce a scenario for the continuous manufacture of polymer/metal-based biosensors—beyond batch.

Comparison Table

In Table 1.2, the most popular manufacturing techniques in crafting micromachines are compared. These tools are used to fabricate microstructures either in serial (Se), parallel (Ba for batch), or continuous (C) mode; the techniques can be organized in many different other ways, for example on the basis of the type of energy involved in the machining or as nonlithography-based (i.e., traditional—group 1) or lithography-based (i.e., nontraditional—group 2), i.e., truly 3D vs. truly 2½D. Table 1.2 is a listing of lithographic and nonlithographic miniaturization techniques with their respective resolutions, defined here as the smallest lateral features that have been generated. As a very rough rule of thumb, one might say that with dimensions below 100 μm one needs to switch from mechanical tooling (group 1) to optical tooling (group 2) to cut into a material.

Wet bulk micromachining, a very popular MEMS technique, had its genesis in the Si-based IC industry, but further development required the adoption and adaptation of several different processes and materials. Si-based wet bulk micromachining techniques involve anisotropic etching in the bulk of the material, often resulting in stunningly beautiful scanning electron microscope micrographs, are one of the most visual and "sexy" micromachining tools used—the MTV of miniaturization science. But in the trend toward miniaturization, almost a law of nature ever since the first oil crisis and the IC revolution, many other precision machining techniques on many different materials have also found commercial applications. Wet bulk Si micromachining, the most traditional form of Si micromachining (dominated by electrical engineers), has found acceptance in traditional precision machining circles (dominated by mechanical engineers) and vice versa: silicon micromachinists are starting to look at precision engineering techniques as a way to extend their tool kit. The most important characteristic for both disciplines is the ever-increasing quest for improved accuracies, higher aspect ratios, and reduced cost, and practitioners have recognized the complementary nature of their respective fields.

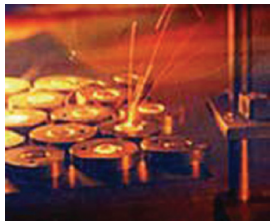
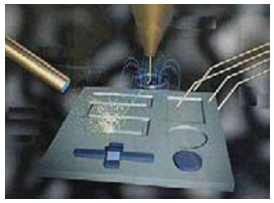
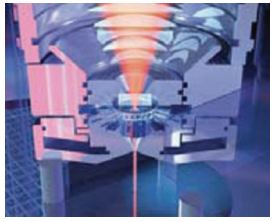
Whereas mechanical sensor manufacture is moving toward more integration of sensing functions with electronics, embodied in increased reliance on surface micromachining mostly involving the deposition and etching of thin films on top of a Si substrate, which is most compatible with IC fabrication, chemical and biosensors are moving away from integration and toward hybrid thick technology. This trend in chemical and biological sensors is caused by the need for modularity and the tremendous compatibility problems that arise when attempting to integrate chemical sensor materials with ICs.

From Table 1.2, Si micromachining then emerges as only one of the many options for precision machining. To emphasize the need for micromachinists to look beyond Si as the ultimate substrate and/or building material, we have presented many examples of non-Si micromachinery throughout this book. The need to incorporate new materials and processes is especially urgent for progress in chemical sensors and microinstrumentation, which rely on non-IC materials and often are relatively large.

The renewed interest in some of the non-Si-based machining methods stems principally from two major deficiencies of IC-based machining techniques: the difficulty of creating truly 3D microstructures and the problem of interfacing microstructures with the macroworld. A machining method with a range that covers macroscopic- to microscopic-scale devices is needed, as often micromachines cannot easily be handled. Laser beam machining (LBM), electrodischarge machining (EDM), ultraprecision machining, dry reactive ion etching (DRIE), LIGA, and pseudo-LIGA techniques are steps in that direction, forming so-called "handshake technologies," bridging the micro- and macroworlds. These

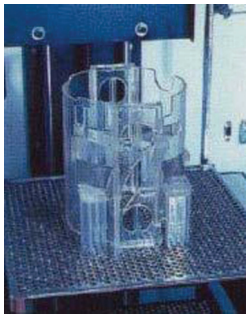
Table 1.2 MEMS Miniaturization Methods: Nonlithography-Based (Traditional) and Lithography-Based (Nontraditional)

Machining Method	Material/ Application	Typical Min/Max Size Feature	IC Compatible?	Tolerance	Important Reference on Technique	Aspect Ratio (Depth/ Width)	Initial Investment Cost/Access
Group 1: Traditional Techniques (Not Involving Photolithography-Defined Masks)							
Chemical milling (S), (Ba)	Almost all metals	From submillimeter to a few meters (x, y); max thickness (z) ±1 cm	Yes	Lateral tolerance 0.25–0.5 mm	Harris[6]	~1	Low/good
Electrochemical machining (S/A), (Ba)	Hard and soft metals, turbine blades, pistons, fuel-injection nozzles	Minimum-size devices larger than in chemical milling because of contacting need	Fair	Lateral tolerance <10 μm	(A) Romankiw,[7,8] (S) Phillips[9]	(S) 100	~$400,000/good
Electrodischarge machining (S), (Se)	Hard, brittle, conductive materials used for tools and dies	Minimum holes of 0.3 mm in 20-mm-thick plate	No	Lateral tolerance 5–20 μm	Kalpajian[10]	100	High/good: equipment with numerical control is common
Electrodischarge wire cutting (S), (Se)	Hard, brittle materials; many punch-and-die applications	Minimum rods 20 μm in diameter and 3 mm long	No	Lateral tolerance 1 μm	Saito[11]	>100	High/good: does not require special electrodes

Electron beam machining (S/A), (Se) 	Hard-to-machine materials	(S) Most suited for large numbers of simple holes (<0.1 mm)	Fair	(S) ~10% of feature size (5 μm on a 50-μm hole)	Taniguchi[12]	(S) 10 is typical but 100 is possible	Very high/fair ~$100,000 when using a modified scanning electron microscope
Continuous deposition (A), (C); e.g., doctor's blade technology, tape casting	With all materials available in inks, e.g., glucose sensors	Most suited for inexpensive disposables, from 100 μm to a few millimeters	No	15 μm	Harper,[13] Fiori and De Portu[14]		Low/good
Focused ion-beam on a lathe (S/A), (Se) 	Very pure IC materials	From submicrons to millimeters	Yes	(S) 50–100 nm	Vasile et al.[5]		High/poor
Hybrid thick film (A), (Ba)	Wide variety of materials available in inks	Minimum feature size 90 μm	Fair	12 μm	Harper[13]		~$30,000/good
Laser beam machining (S/A), (Se) 	Complex profiles in hard materials	(S) Holes from 10 μm to 1.5 mm at all angles	Fair	1 μm	Helvajian[15]	(S) 50	~$50,000 but up to ~$400,000 for a five-axis system/ good
Plasma-beam machining (S/A), (Se/Ba)	Very-high-temperature materials	(A) Only used for thick films >25 μm; (S) for very thick films >2.5 mm	No	(A) 20 μm for a 25-μm-thick film; (S) typical ± 3 mm but 0.8 mm is possible	Pfender[16]		~$600,000/fair

(*Continued*)

Table 1.2 MEMS Miniaturization Methods: Nonlithography-Based (Traditional) and Lithography-Based (Nontraditional) (*Continued*)

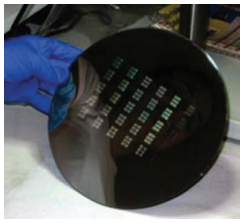
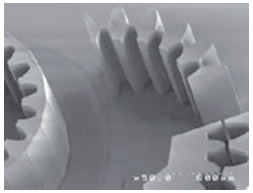
Machining Method	Material/ Application	Typical Min/Max Size Feature	IC Compatible?	Tolerance	Important Reference on Technique	Aspect Ratio (Depth/ Width)	Initial Investment Cost/Access
Stereolithography (A), (Se) 	Polymeric photosensitive materials	Maximum 10 × 10 × 10 mm (*x, y, z*)	Yes	Minimum solidification 5, 5, 3 μm (x, *y, z*)	Ikuta,[64] stereo lithography[17]		$100,000–500,000/ good
Ultraprecision mechanical machining (S), (Se)	Form-stable materials	From submillimeters (e.g., 0.2-mm hole) to meters	No	1 nm	Boothroyd and Knight[18]		$400,000/good
Ultrasonic machining (S), (Se)	Hard and brittle materials	Holes from 50 μm to 75 mm	No	Lateral tolerance 10 μm	Bellows and Kohls[19]	2.5 μm for a 250-μm hole	$20,000/good
Microdispensing (A), (Se/Ba)	Wide variety of materials available in inks, especially suited to biomaterials	Minimum membrane volume: nanoliters to microliters	Yes	~3% volume	Li et al.[20]		Medium/poor
Injection molding, hot embossing (A), (Ba)	Plastics, ceramics, and metals	Minimum 250 μm; maximum several centimeters	No	±0.13% (e.g., ±2.6 μm for a 1.998-mm diameter)	Sakai,[21] Rak[22]	>5	Medium/good

Powder blasting (S), (Se)	Ceramics and hard materials >40 HRC	Minimum 50 μm up to several millimeters	No	1–2 μm	Belloy et al.[23]		Medium/poor
Abrasive water jet (S), (Se)	Metals, plastics, and ceramics	Min 0.8 mm typical, 80 μm possible	No	±50 μm	Waterjet[24]	>100	$80,000–200,000/ poor

Machining Method	Material/ Application	Typical Min/Max Size Feature	IC Compatible?	Tolerance	Important Reference on Technique	Aspect Ratio (Depth/Width)	Shape and Height/Depth	Initial Investment Cost/Access
Group 2: Nontraditional Techniques (Involving Photolithography-Defined Masks)								
Photofabrication (S)	Plastic, glass (ceramic), e.g., fluidic elements	Max *x, y* = 40 × 40 cm and max *z* = 0.6 mm	Yes	Lateral tolerance 20 μm	Trotter[25]	~3 for photoplastics; ~20 for photoglass	*x, y* is free; *z* up to 6 mm	Medium/ poor to medium
Photochemical milling (S)	Printed circuit boards, lead frames, shadow masks	Max 60 × 60 cm, max thickness <0.5 mm	Yes	13 μm (printed circuits)	Allen[26]	~1	*x, y* is free; *z* up to 0.5 mm	Medium/ good
Wet etching of anisotropic materials (S)	Crystal Si, GaAs, quartz, SiC, InP	Max wafer size, min feature a few microns	Fair	1 μm	Kern[27]	100	*x, y, z* shape locked in by crystallography, *z* height of the wafer	Low/good
Dry etching (S)	Most solids	Max wafer size, min feature submicron	Good	0.1 μm	Manos and Flamm[28]	10	*x, y* shape free; *z* up to 200 μm	High/good
Poly-Si surface micromachining (S/A)	Poly-Si, Al, Ti, etc.	Max wafer size, min feature submicron	Good	0.5 μm	Johnstone and Parameswaran[29]		*x, y* free; *z* 0.1–10 μm, but preferably 1–2 μm	High/fair

(*Continued*)

Table 1.2 MEMS Miniaturization Methods: Nonlithography-Based (Traditional) and Lithography-Based (Nontraditional) (*Continued*)

Machining Method	Material/ Application	Typical Min/Max Size Feature	IC Compatible?	Tolerance	Important Reference on Technique	Aspect Ratio (Depth/Width)	Shape and Height/Depth	Initial Investment Cost/Access
Si on insulator (S) 	Crystalline Si	Max wafer size, min feature submicron	Good	0.1 μm	Diem et al.[30]		*x*, *y* free; *z* height depending on the epilayer, e.g., 100 μm	High/poor
LIGA (S/A)	Ni, PMMA, Au, ceramic, etc.	Max 10 × 10 cm or more, min 0.2 μm	Fair	0.3 μm	Ehrfeld[31]	>100	*x*, *y* free, *z* up to 100 μm	High (>$35M)/ poor
UV-transparent resists (S/A) 	Polyimide, SU-8, AZ-4000	Max size equals the wafer size	Good	0.5 μm	Ahn et al.,[32] SOTEC[33]	25	*x*, *y* free, *z* up to 1 mm	High/poor
Molded polysilicon HEXSIL (Keller) (S/A)	Poly-Si, Ni, etc.	Max several millimeters, min few microns	Good	0.5 μm	Keller and Ferrari[34]	10	*x*, *y* free; *z* up to 100 μm	High/poor
Erect polysilicon (Pister) (S/A)	Poly-Si	Max several millimeters, min submicron	Good	0.5 μm	Pister[35]	10	*x*, *y* free, *z* up to 1 mm	High/poor
Micromolding (A), (Ba)	Ceramics, plastics, metals	Max wafer size, min feature several microns	No		Bride et al.,[36] Zhao et al.,[37] Sander et al.[38]	3–20	*x*, *y* free, *z* up to 200 μm	Medium/ poor to fair

A, additive; Ba, batch; C, continuous; IC, integrated circuit; max, maximum; MEMS, microelectromechanical system; min, minimum; PMMA, poly(methylmethacrylate); S, subtractive; Se, serial. LIGA is a German acronym for "**Li**thographie, **G**alvanoformung, **A**bformung," which in English is (x-ray) lithography, electroplating, and molding. For a more in-depth study on manufacturing options, we refer to the literature: Harris,[6] Shaw,[39] DeVries,[40] Slocum,[41] and Evans.[42] Also visit the Louisiana Tech University's Institute for Micromanufacturing (http://www.latech.edu/ifm).

handshake technologies should be of importance beyond micromachines—for example, in ICs, where one is faced with the same dilemma of packaging, mounting, and wiring large-scale integrated chips (LSIs) and very large-scale integrated chips (VLSIs). Several of these new technologies are maskless and enable very fast prototyping. The merging of bulk micromachining with other new fabrication tools, such as surface micromachining and electroplating, and the adaptation of materials such as Ni and polyimides have fostered powerful new nontraditional precision engineering methods.

Table 1.2 comprises both additive (A) and subtractive (S) techniques. For additive techniques, a further comparison of thick-film versus thin-film deposition methods is presented in Table 1.3. The decision to go with a thin versus thick film has very important ramifications and has to be considered very deliberately. Resolution and minimum feature size for thin films are obviously superior. Also, the porosity, roughness, and purity of deposited metals are less reproducible with thick films. Finally, the geometric accuracy is poorer with thick films. On the other hand, the thick-film method displays versatility, which is often key in chemical sensor manufacture. Silk screening forms an excellent alternative when size does not matter but cost, in relatively small production volumes, does. For biomedical applications the size limitations, clear from Table 1.3, make thick-film sensors more appropriate for in vitro applications. For in vivo sensors, where size is more crucial, IC-based technologies might be more appropriate. Thin-film technology does not necessarily involve IC integration of the electronic functions. Table 1.4 provides a comparison of the economic and technical aspects in implementing thin-film technology, IC fabrication, thick film, and classic construction for sensors.[43] It should be mentioned that in sensor fabrication, 60–80% of the cost consists of packaging, an aspect not addressed in Table 1.4. In comparison with complementary metal-oxide-silicon (CMOS)-compatible thin-film sensors, the packaging of thick-film sensors usually is more straightforward. Since packaging expenses overshadow all other costs, this is a decisive criterion. In the case of chemical sensors, where the sides of the conductive Si substrate might shunt the sensing function through contact with the electrolyte, encapsulation is especially difficult compared with the packaging of an insulating plastic or ceramic substrate (see also Chapter 4, this volume, on packaging).

Table 1.3 Comparison of Thin- versus Thick-Film Technology

Property	Silicon/Thin Film	Hybrid/Thick Film
In-plane resolution	0.25 μm and better	12 μm
Minimum feature size	0.75 μm and better	90 μm
Temperature range	<125°C	>>125°C
Sensor size	Smaller	Small
Geometric accuracy	Very high	Poor
Deposition methods	Evaporation, sputtering, CVD*	Screen printing, stencil printing
Patterning methods	Etch through photomask; lift-off stencil	Screen photomask, etched stencil, machined stencil
Reliability nonencapsulated device	Low	High
Electronic compatibility	Good	Moderate
Versatility	Low	Very good
Roughness, purity, and porosity of deposited materials	Superior	Moderate
Energy consumption	Low	Moderate
Handling	Difficult	Easy
Approximate capital costs per unit	Very large, but very low in large volumes	Low; very low in moderate volumes

*CVD, chemical vapor deposition

Table 1.4 Comparison between Different Sensor Technologies: Economic and Technical Aspects

	Classic Construction	Thick-Film Technology	Thin-Film Technology	IC Technology*
Technology substrate	Wires and tubes	Screen printing Al_2O_3, plastic	Evaporation-sputtering Al_2O_3, glass, quartz	IC techniques silicon, GaAs
Initial investment	Very low	Moderate	High	High
Production line cost	>$10,000	>$100,000	>$400,000	>$800,000
Production	Manual production	Mass production	Mass production	Mass production
Units per year	1–1,000	1,000–1,000,000	10,000–10,000,000	100,000 and up
Prototype	Cheap	Cheap	Moderate	Expensive
Sensor price	Expensive sensor	Low cost per sensor	Low cost per sensor	Low cost per sensor
Use	Multiple use, in vitro, in vivo	Disposable, in vitro	Disposable, in vivo	Disposable, in vivo
Markets	Research, aerospace	Automotive, industrial	Industrial, medical	Medical, consumer
Dimensions	Large	Moderate	Small	Extreme miniaturization
Solidity	Fragile	Robust	Robust	Robust
Reproducibility	Low	Moderate	High	High
Maximum temperature	800°C	800°C	1,000°C	150°C (silicon)
Interfacing	External discrete devices	Smart sensors, surface mount	Smart sensors, surface mount	Smart sensors, CMOS,† bipolar

*IC, integrated circuit.
†CMOS, complementary metal-oxide-silicon.
Source: From Lambrechts, M., and W. Sansen. 1992. *Biosensors: Microelectrical devices*. Philadelphia: Institute of Physics Publishing. With permission.[43]

Selection criteria for the additive processes reviewed in Volume II depend on a variety of considerations, such as the following.

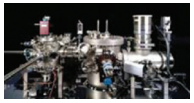
1. Limitations imposed by the substrate or the mask material: maximum temperature (T_{max}), surface morphology, and substrate structure and geometry
2. Apparatus requirement and availability
3. Limitations imposed by the material to be deposited: chemistry, purity, thickness, T_{max}, morphology, crystal structure, etc.
4. Rate of deposition to obtain the desired film quality
5. Adhesion of deposit to the substrate; necessity of adhesion layer or buffer layer
6. Total running time, including setup time and postcoating processes
7. Cost
8. Ease of automation
9. Safety and ecological considerations

Table 1.5 compares some of the additive processes important in microsensors and micromachining. This table and the above criteria complement the questions in the check-off list presented in Table 1.11. The latter is meant as a guide toward a more intelligent choice of an optimum machining substrate for the micromachining task at hand.

A subset of subtractive machining techniques from Table 1.2 is compared in Table 1.6. Here, metal removal rates, tolerance, surface finish, damage depth, and required power are listed.

The metal removal rate by electrochemical machining (ECM) and plasma arc machining (PAM) is much higher than that of the other machining processes. Compared with conventional mechanical CNC machining, the metal removal rates by ECM and PAM are, respectively, 0.3 and 1.25 times that of CNC, whereas others are only a small fraction of it. Power requirements for ECM and PAM are also comparatively high. The tolerance obtained

TABLE 1.5 Comparison of Additive Processes Important in Microsensors and Micromachining

	Species Deposit Ion Mechanism	Deposit Ion Rate	Depositing Species	Coverage of Complex-Shaped Objects	Coverage into Small Blind Holes	Metal Deposit Ion	Alloy Deposit Ion	Refractory Compound Deposit Ion	Energy of Depositing Species	Bombard-ment of Substrate/ Deposit by Inert Ions	Substrate Heating by External Means	Cost
Evaporation	Thermal energy	Very high (up to 750,000 Å/min)	Atoms and ions	Poor line-of-sight coverage	Poor	Yes	Yes	Yes	Low (0.1–0.5 eV)	No, normally	Yes, normally	Low
Sputtering	Momentum transfer	Low except for pure metals (e.g., Cu, 10,000 Å/min)	Atoms and ions	Good, but non-uniform thickness distribution	Poor	Yes	Yes	Yes	Can be high (1–100 eV)	Depends on geometry	Not generally	High
CVD 	Chemical reaction	Moderate (200–2500 Å/min)	Atoms	Good	Limited	Yes	Yes	Yes	High in plasma-enhanced CVD (PECVD)	Possible	Yes	High
Electrodeposition	Precipita-tion of redox species	Low to high	Ions	Good	Limited	Yes, limited	Limited	Limited	Can be high	No	No	Low
Thermal spraying	Solidifica-tion of molten droplets	Very high	Droplets	No	Very limited	Yes	Yes	Yes	Can be high	Yes	No, normally	Very high

CVD, chemical vapor deposition.

Table 1.6 Machining Characteristics of Different Processes

	MRR (mm^3/min)	Tolerance Maintained (μm)	Surface Finish Required (μm)	Surface Damage Depth (μm)	Power Required for Machining (W)
USM	300	7.5	0.2–0.5	25	2,400
AJM	0.8	50	0.5–1.22	2.5	250
ECM	15,000	50	0.1–2.5	5	100,000
CHM	15	50	1.5–2.5	5	N/A
EDM	800	15	0.2–1.2	125	2,700
EBM	1.6	25	0.5–2.5	250	150–2,000
LBM	0.1	25	0.5–1.2	125	2–200
PAM	75,000	125	Rough	500	50,000
CNC	50,000	50	0.5–5	25	3,000

AJM, abrasive jet machining; CHM, chemical machining; CNC, computer numerical control machining; EBM, electron beam machining; ECM, electrochemical machining; EDM, electrodischarge machining; LBM, laser beam machining; MRR, metal removal rate; PAM, plasma arc machining; USM, ultrasonic machining.
Source: Based on http://www.cemr.wvu.edu.

by various processes except PAM is within the range of CNC machining, which means satisfactory dimensional accuracy can be maintained. Satisfactory surface finishes are obtained by all processes except PAM. Depth of surface damage is very small for abrasive jet machining (AJM), electrochemical machining (ECM), and chemical machining (CHM) processes, whereas it is very high in the case of electrodischarge machining (EDM), electron beam machining (EBM), PAM, and the like. [As we saw in Volume II, these techniques give rise to a heat-affected zone (HAZ).] For this reason, ECM can be employed for making dies and punches. AJM may be suitably employed for machining super alloys and refractory material; ECM has good applicability for machining steels and super alloys; CHM is very useful for aluminum and steel; EDM may well be adopted for machining steels, super alloys, titanium, and refractory material; and PAM can be suitably employed for machining aluminum, steels, and super alloys. Capital cost for ECM is very high compared with conventional CNC machining, whereas capital costs for AJM are PAM are comparatively low. EDM has a higher tooling cost than other machining processes. Power consumption cost is very low for PAM and laser beam machining (LBM) processes, whereas it is greater in case of ECM. Metal removal efficiency is higher for EBM and LBM than for other processes. Tool wear is very low for ECM, CHM, EBM, LBM, and PAM, whereas it is high for electrical discharge machining.

Given the tremendous variety of micromachining tools available, as evident from Tables 1.2–1.6, a truly multidisciplinary engineering education will be required to design miniature systems with the most appropriate building philosophy.

Design from the Package Inward

When approaching the development of a microdevice, we prefer using the terminology *miniaturization* rather than *MEMS*, as the latter term is still too loaded with the notion that Si technology represents the only solution. The choice of processes for manufacturing a 3D microdevice can best be made after studying the detailed requirements of the application. Since the package serves as the interface between the microstructure and the macroworld, and since it is the major contributor to cost and size, one should start the design with the package and work toward the best machining process for the micromachine inside. The package is often made with a traditional machining method such as turning and grinding or possibly with ultrasonic or wire electrodischarge machining.

For the micromachine inside the package, there may be many machining options, as both batch IC and traditional serial production techniques, perhaps in combination with micromolding, must be considered. Several new, batch Si micromachining techniques derived from the IC industry are

gaining acceptance, but serial, precision machining techniques with specialized tools (e.g., ultrasonic machining or laser beam machining) are often the only commercially available methods to make intricate, small, 3D microcomponents.

Decision Tree for the Optimized Micromanufacturing Option

Introduction

In this section we develop a checklist on how to arrive at an optimum miniaturization strategy for a new MEMS/NEMS product, choosing between MEMS manufacturing options: from nonlithography-based to lithography-based, as captured in Table 1.2. To become a proficient MEMS/NEMS engineer one needs to first understand the application and market for the product very well. These aspects should be apparent from a very detailed specification list and discussions with the user/client. Once an appropriate sensor/actuator principle for the MEMS/NEMS has been chosen, on the basis of the specifications, through brainstorming sessions with the user and/or client and by making preliminary designs, one needs to develop a clear understanding on how this sensor/actuator principle scales into the microdomain. Then one must choose the optimum manufacturing approach and the optimized materials, including substrate material, for the most cost-effective practical implementation. These choices are tested through a new design phase and evaluation of critical process steps. A major challenge in this regard often concerns partitioning, i.e., how far to push the integration of electronics with the sensing function (hybrid vs. monolithic)? If the MEMS-based product compromises both a disposable and a fixed instrument the question becomes, what should be included with the MEMS disposable and what should be included in the fixed instrument? With partitioning decisions made, a computer-aided design (CAD) is followed by research prototyping, engineering prototyping, and finally α and β products for the user/client to test. With a satisfied customer, manufacturing of a new MEMS product may begin.

Throughout the rest of this chapter, we illustrate the use of our proposed decision tree with the search for a sensor to measure glucose in a drop of blood. Glucose sensing represents the single largest biosensor market, and a significant amount of all R&D in biosensors worldwide is geared toward improving this technology. Glucose self-testing by diabetics is one of the fastest-growing segments of the world diagnostics market. The World Health Organization estimates a worldwide diabetic population of more than 200 million (1–2% of the world population). Diabetes is the fifth leading cause of death in the United States, and 30% of our children will be diagnosed with diabetes in the future. The US healthcare costs related to diabetes are $132 billion per year. The specifications for a glucose self-test meter with disposable sensor include very low cost disposable (<15 cents), in vitro use, millions of devices manufactured per month, size of active area on the sensor of a few millimeters square (dictated by the size of the drop of blood), linear response with glucose concentration, and a simple handheld measuring device. Examples of glucose sensors/meters for diabetic care are shown in Figure 1.7. Our analysis will clarify why the current glucose sensors are constructed the way they are.

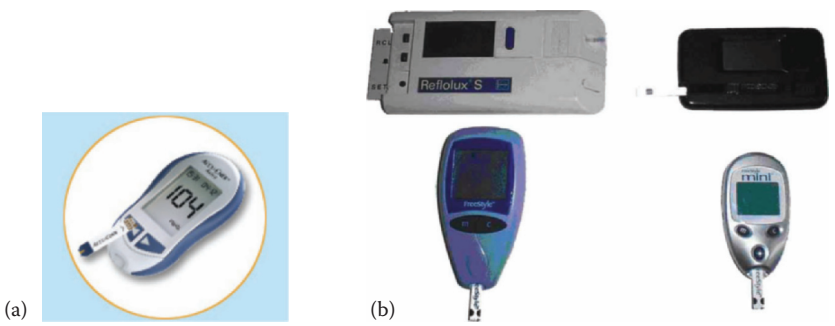


FIGURE 1.7 Blood glucose monitor. (a) Glucometer from Roche (Accu-Chek). (b) Other examples of blood glucose meters, 1993–2005.

Specifications

Applied miniaturization starts with a thorough understanding of the sensor or microsystem specification list, interviews with client/users, understanding of the application environment, and, equally important, a firm market appreciation. In choosing a sensor/actuator principle for a miniaturization application, the various possible sensing/actuating principles (see Table 8.2) need to be challenged with this detailed specification list. During discussions of the specifications required for measuring a certain measurand, jargon confusion is common; making face-to-face meetings with clients/user is mandatory. A specification list should include static and dynamic performance characteristics. Static specifications for a biomedical product with both a reader and a disposable might include the following.

- Package/interface with the world.
- Esthetics/design of package.
- Sensitivity. Change in signal (response) per change in concentration (or mass).
- Lower limit of detection (LOD). The minimum concentration (mass) that can be detected at a given confidence level. The detection limit is affected by the level of noise in the data. One can determine the level of noise by repetitively measuring the signal of a blank.
- Resolution. The smallest increment of change in the measured value that can be determined from the instrument's readout scale; e.g., for a pH electrode the accuracy is ±0.2 pH and the resolution is 0.1 pH.
- Selectivity/specificity. The ability to accurately quantify one species or component in the presence of variable amounts of other components. It has a great effect on the accuracy of the method.
- Stability (long-term and short-term).
- Response time.
- Ambient conditions (e.g., temperature, under or above water, pressure).
- Sample size.
- Operating life.
- Drift. Gradual change in signal that is not due to change in analyte content.
- Shelf life.
- Cost, size, weight of reader.
- Cost, size, weight of disposable.
- Number of devices needed.
- Accuracy.
- Precision.
- Output format/display.
- Operating voltage, current, power.
- Quality.
- Reliability, i.e., mean time between failures.
- Hysteresis.
- Threshold.
- Nonlinearity.

Typical dynamic specifications for such a biomedical product are as follows:

- Dynamic error response.
- Dynamic range. Typical sensor response reaches saturation because of either saturation of binding sites or instrument factors.
- Hysteresis.
- Instability and drift.
- Noise.

Given the very fragmented nature of the microsystems business, it is impossible to present a generic specification list, as a good list must also detail all application-related intricacies. As an example, consider a pH sensor for sensing pH in the secondary cooling flow loop of a nuclear reactor. The presence of radiation and water at high temperature and pressure will significantly reduce the potential sensor principles one may consider. The example also clarifies why a specification list can hardly ever be detailed enough.

The specifications for our example disposable glucose sensor and glucose meter are as follows.

- Package/interface. Handheld device with a slot that accepts disposable glucose sensor strips.
- Esthetics/design. Esthetics are more and more important: good design, easy to use. Letters must be large. Many diabetics are elderly people and have poor eyesight.
- Sensitivity. Sensitivity is about 100 nA/M.
- Lower LOD. Lowest measurable level should be 0.06 mM glucose or less.
- Dynamic range. Glucose assay range is from 1.1 to 33.3 mmol/L.

- Selectivity/specificity. Glucose test results are automatically adjusted for common interfering medications, vitamins, and endogenous substances, including up to 100 µg of acetaminophen, 20 mg/dL uric acid, 3 mg/dL ascorbic acid, 500 mg/dL cholesterol, and 3000 mg/dL triglycerides.
- Response time. The time it takes to read a glucose sensor may range from 5 to 60 seconds (modern meters are typically below 15 seconds).
- Ambient conditions. Sensor operating temperature range = 18–30°C.
- Sample size. Sample sizes vary from 30 to 0.3 µL.
- Operating life. Nonreplaceable battery life = 4000 tests.
- Shelf life. Six months.
- Cost, size, weight of reader. The glucose reader should be approximately the size of the palm of the hand, and it should be battery powered.
- Cost, size, weight of disposable. Very low cost (less than 15 cents). Strip is a few centimeters square in size, and the active spot is a few millimeters square. A consumable sensor element containing chemicals, which react with glucose in the drop of blood, is used for each measurement. Since sensors may vary from batch to batch, a provision might be required for the user to enter in a code that may be found on the batch of sensors. By entering the code into the glucose meter, the meter will be calibrated to that batch of glucose sensors.
- Number of devices needed. Each strip can be used only once and is then discarded. Millions of glucose sensor strips will be manufactured per month.
- Accuracy. The minimum acceptable accuracy for results produced by a glucose monitoring system shall be as follows: 95% of the individual glucose results shall fall within ±15 mg/dL (±0.83 mmol/L) of the results of the manufacturer's measurement procedure at glucose concentrations less than 75 mg/dL (4.2 mmol/L) and within ±20% at glucose concentrations greater than or equal to 75 mg/dL (4.2 mmol/L).
- Precision. Determined as a coefficient of variation on replicate tests on fresh blood in the 3.5–6% range.
- Output format/display. The glucose value in mg/dL or mmol/L displayed in a window. The preferred measurement unit varies by country. In the United States, mg/dL is preferred, and mmol/L is preferred in Canada and Europe. To convert mmol/L of glucose to mg/dL, multiply by 18.
- Reliability. Does not apply for glucose sensor: one-time use only.
- Hysteresis. Does not apply for glucose sensor: one-time use only.
- Nonlinearity. Linear glucose sensor response.

The specifications for a glucose meter listed above suggest a very low cost disposable (<15 cents) with an active area of a few millimeters square (dictated by the size of the drop of blood). These two data points alone preclude the use of Si technology, as this would represent a too-expensive solution. Since we desire a small, simple, and inexpensive measurement device, and a linear response to the glucose in blood, an amperometric approach stands out as the most attractive sensing option (see Volume I, Chapter 7). An amperometric sensing approach is very simple and inexpensive—one pair of thin-film metal electrodes on an insulating substrate—and the method produces a signal linear in glucose concentration. See, for example, the MediSense/Abbott sensor, the Precision QID, in Volume I, Figure 7.80. The linear response makes it more sensitive than a potentiometric device that features a log dependency of the voltage signal on concentration. As we learned in Volume I, Chapter 7, an optical approach tends to be more expensive than an electrochemical approach (see Volume I, Table 7.21).

Having established a sensing mechanism, we now need to know how the chosen effect (amperometry) scales into the microdomain.

Scaling Characteristics of Micromechanisms

To make the right choice from the different available sensing/actuating principles, it is also important to know how the phenomena scale in the microdomain.

Scaling effects are explored in detail in Chapter 7, this volume, and some intuitive guidance (rules of thumb) is provided in Table 1.7.

Scaling understanding is crucial early on when deciding on a sensing/transduction principle. For example, both absorption and amperometric detection approaches scale poorly compared with a luminescent or a potentiometric sensing approach.

We apply these scaling rules now to the glucose sensor example. From the above, we appreciate that an amperometric device scales more poorly than a potentiometric device (the smaller an amperometric device, the smaller the current, while the signal for a potentiometric device is scaling invariant), but an in vitro glucose sensor may be relatively large, and sensitivity is more important in this case. A potentiometric device would be as inexpensive as our chosen amperometric approach, but its sensitivity is not as good. Moreover, we saw in Volume I, Chapter 7, that by making an array of electrodes, the signal to noise ratio of a microamperometric glucose sensor may be enhanced because of increased mass-transport at the electrode edges. So we stick with the amperometric sensor as our choice for the proposed glucose meter.

Partitioning

Introduction

A major challenge in the manufacture of MEMS and NEMS revolves around correct partitioning, that is, how far to push integration of electronics with the MEMS or NEMS sensing function (hybrid vs. monolithic). What do we include with the MEMS disposable, and what can we put into the fixed reader instrument? Onboard or offboard fluidics? Battery or mains power? We first compare hybrid versus monolithic approaches and then illustrate the challenges of partitioning a MEMS instrument by looking at an example from the burgeoning field of microfluidics. Partitioning of electronics and instrument components obviously affects packaging needs, and we will come back to this point in Chapter 4, this volume, on packaging.

Monolithic versus Hybrid MEMS

Hybrid integration in the IC industry means combining thin-film ICs with thick-film technology.

Table 1.7 Some Important Effects of Scaling on Design of Miniaturized Devices

1. Scaling of forces: weak forces are preferred in the microdomain. Electrostatic forces [with a quadratic dependency on dimension (1^2)] under certain conditions outperform magnetic torque [with a cubic dependency on dimension (1^3)] since they become relatively stronger in small devices.
2. Increasing strength of materials in the microdomain. Single-crystal whiskers may be more than 1000 times stronger than the bulk material.
3. The dominance of surface effects in the microworld: for example, because of an increase in S/V (surface-to-volume ratio), better heat dissipation results (thermal isolation is difficult and cooling is excellent) and frictional forces increase.
4. Decrease of manufacturing accuracy of micromachines (see Volume II, Figure 6.3).
5. The need for error-insensitive design (fewer parts, flexible materials) and the use of nature as a guide in design philosophy for very small devices.
6. Amperometric and absorption measurements scale poorly into the microdomain compared with potentiometric and chemiluminescence (see Volume I, Chapter 7, and Chapter 7, this volume, for details). Our proposed glucose sensor is amperometric, so we need to look in some more detail into whether this averse scaling should influence our choice.
7. Mechanical, thermal, and electromagnetic response time constants and many other time constants shrink with advanced miniaturization, generally resulting in shorter response times. Microsystems, with their small inertial mass, are faster, leading to higher speed or frequency, but higher frequency of a cyclic process generally leads to larger losses and consequently to lower efficiency.
8. Power consumption is often dramatically reduced, especially if good thermal isolation is possible for elements with small heat capacity.
9. Energy sources scale very poorly (1^3), making onboard energy sources (e.g., batteries and fuel cells) less attractive for powering micromachines than radiating in energy (e.g., solar or laser light with micromachined photovoltaic converters or microwaves if extremely small receivers and converters were available).
10. Most actuators, which rely on power for inducing a movement, also scale poorly with reducing size, and it is often better to consider a "macroactuator," which can produce a micromotion, than a micromachined actuator with too small of a stroke.

Viewed from the Si MEMS sensor angle, a hybrid sensor keeps the electronics separate from the sensor. The hybrid might then consist of two pieces of Si on the same substrate connected by a short wire bridge, or it might be a Si sensor mounted in a header

plugged into an electronics board. In a monolithic MEMS approach, on the other hand, electronics and MEMS elements are cofabricated within one single (lengthy!) sequential Si process. Each step in such a sequence has a nonzero failure probability and, since probabilities multiply fast, the yield may be dramatically reduced for long process sequences. In this light, there are some positive aspects to a hybrid MEMS approach. In the latter, MEMS and electronics are fabricated in separate high-yield processes and joined within a common package/substrate, improving the overall yield. Microassembly in hybrid manufacture enables the mixture of components from separate optimal processes, including nonstandard processes and materials that are incompatible. For example, many optoelectronics require GaAs substrates that are incompatible with standard Si electronics (see also under microassembly in Chapter 4, this volume). In addition, in hybrid MEMS, *all* components needed for a complete MEMS function can be accommodated. Somewhat unexpectedly, parasitics are sometimes lower in a hybrid than in an integrated approach. The device components in a hybrid remain serviceable and, because the sensor package determines the final size, hybrids often are not much larger than their Si integrated counterparts. Hybrids usually are simpler to design and can be produced more economically in smaller-production volume runs. For chemical sensors and biosensors, they often represent the most viable option to implement micromachining technology. Prices for Si mechanical sensors given in market reports mostly refer to hybrid devices with the Si sensor mounted in a header and the signal conditioning kept on separate chips. As an example, Figure 1.8 represents an accelerometer built using a 3-inch Si wafer and involves a fusion bonding process to produce a proof mass, spring, and fixed capacitor plate. The interface circuits are designed to function with different sensors. The specific sensor configuration and calibration data are written into an EEPROM. It is a ±2-g device with 2-mg resolution at 60 Hz. These accelerometers are used for aerospace navigation, in navigation in small commercial airplanes, and in helicopters.

From the precision engineering angle, a hybrid device is one in which parts produced by different

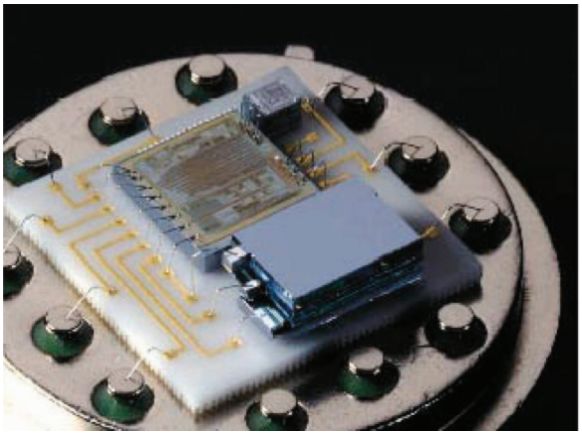


FIGURE 1.8 Commercially available Si accelerometer based on hybrid technology. CSEM 6100 accelerometer with interface circuitry in a T08 package. (From CSEM.)

technologies are joined and assembled. In the electronics industry thick film is often synonymous with hybrid manufacturing and thin film with front-end IC manufacturing. For more details on choosing between thin- and thick-film processes, consult Tables 1.3 and 1.4 above.

For the proposed glucose meter, we will have all the electronics in the reader rather than in the disposable. It makes no sense to throw the electronics away with each used glucose sensor. Moreover, the glucose sensor will be constructed on an inexpensive nonsilicon substrate.

Many chemical sensors come with a pump and some fluidics, say for sample collection, calibration, rinsing, etc. In what follows, we decide what pumping mechanism to employ and how to partition the fluidics between sensor disposable and permanent instrument.

Partitioning in a Microfluidic Instrument

As we saw in Volume I, Chapter 6, microfluidic systems comprising nozzles, pumps, channels, reservoirs, columns, mixers, oscillators, diodes, amplifiers, and valves can be used for a variety of applications, including separation chemistry, drug dispensation, ink-jet printing, and general transport of liquids, gases, and liquid/gas mixtures. The advantages of these devices include lower fabrication costs, enhancement of analytical performance, lower power budget, and lower consumption of chemicals than traditional pumping and spraying systems. Ultimately, these developments will

revolutionize applications where precise control of fluid flow is a necessity.

Generic technical challenges in making a microfluidic instrument—for a biotechnology application for example—include the partitioning of the various instrument functions such as power supply, polymerase chain reaction (PCR) for DNA sample amplification, mechanism for getting the sample into and out of the instrument, providing fresh viable reagents, and detection. Depending on the specific biotechnology application (say, clinical diagnostics or high-throughput drug screening), the electronics, the power for the propulsion mechanism, heaters (e.g., for PCR), etc., may be integrated with a disposable cassette holding the disposables or within the fixed instrument. For diagnostics, both the disposable and the reader instrument should be inexpensive; the disposable should thus involve as little complexity as possible—perhaps only reagents, plastic substrate, and fluidic conduits. The reader instrumentation should be inexpensive, small (perhaps hand-held), and very user friendly. For high-throughput screening, much more sophisticated fluid handling is typically incorporated into a tabletop, expensive research instrument. The disposable in the latter does not need to contain sealed fluids or dry stored chemicals, often making it less sophisticated than disposables in diagnostic application. In one extreme, one might envision a totally integrated option such as the micro-total analysis system[44] illustrated schematically in Figure 1.9a.[45] More realistically, one might look at a hybrid construct, as sketched in Figure 1.9b. For most

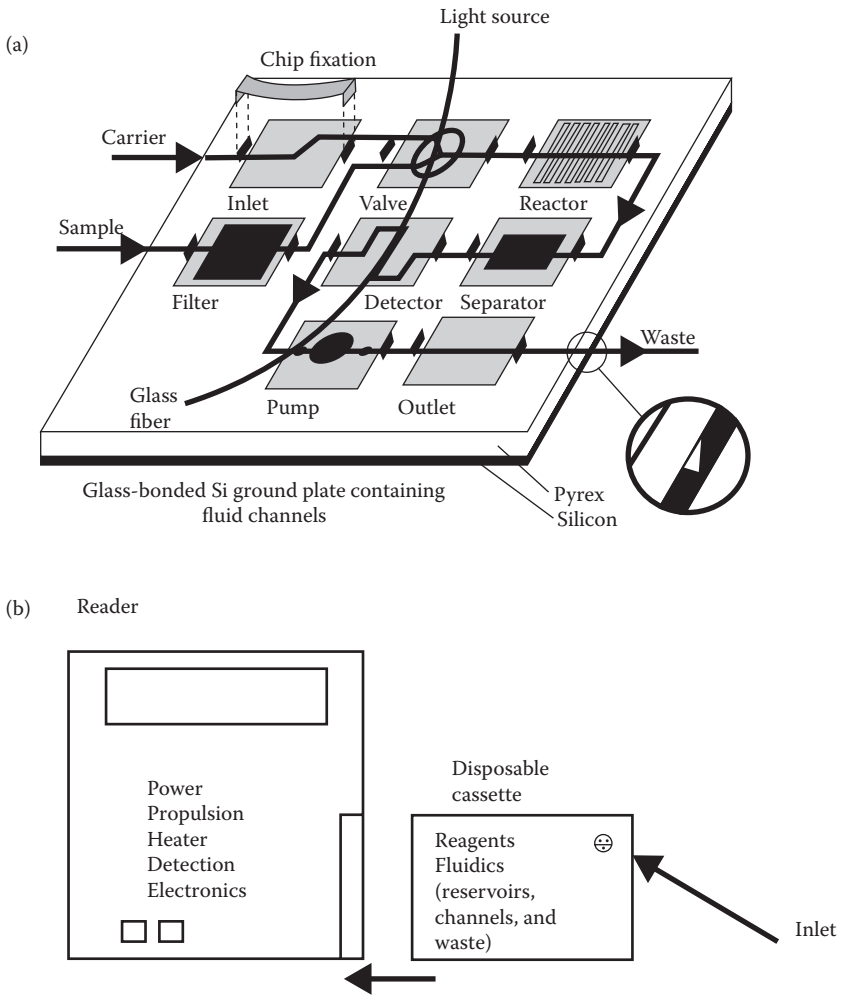


FIGURE 1.9 Partitioning of functions between instrument and a disposable. (a) Configuration for a micro-total analysis system (μTAS). (Based on Bergveld, P. 1995. In *Micro total analysis systems*, eds. A. van den Berg and P. Bergveld, 1–4. Twente, the Netherlands: Kluwer Academic Publishers. With permission.[45]) (b) Generic hybrid microinstrument, with inlet, valves, fluid propulsion, detection, waste, and thermal control. Various options exist for partitioning between cassette and reader.

diagnostic microinstruments, it indeed makes more sense to envision a disposable cassette incorporating the specific reagents needed for a certain test and a separate reader instrument. In a diagnostic application, a major challenge is how to store wet or dry chemicals in sealed reservoirs within the disposable package, making a microfluidic diagnostic instrument actually more challenging to develop than a high-throughput drug screening station.

Our disposable glucose sensor will only have a substrate, electrodes, a chemical cocktail and perhaps a very simple fluidic mechanism to make blood collection on the sensor surface easier. In the context of glucose sensing we will now briefly refer to systems that employ pumps to deliver insulin depending on the glucose sensor reading.

Fluid Propulsion Methods and MEMS Integration

Introduction

As we saw in Volume I, Chapter 6, there are various technologies for moving small quantities of fluids or suspended particles from reservoirs to mixing and reaction sites, to detectors, and eventually to waste or to a subsequent instrument. Methods to accomplish this include syringe and peristaltic pumps, electrochemical bubble generation, acoustics, magnetics, direct current (DC) and alternating current (AC) electrokinetics, and centrifuge. In Table 1.8 we compare some promising fluid propulsion means.[46] Only the simplest and least expensive pumping schemes should be integrated within the disposable itself.

Mechanical Pumps

From Volume I, Chapter 6, we remember that the pressure that mechanical pumps have to generate to propel fluids through capillaries will be higher the narrower the conduit (see Volume I, Equation 6.54). Piezoelectric (acoustic), electro-osmotic (DC electrokinetic), electrowetting (DC electrokinetic), and electrohydrodynamic pumping (AC electrokinetic) all scale more favorably in the microdomain and lend themselves better to pumping in MEMS devices. Despite this promise, almost all biotechnology equipment today is based on traditional external syringe or peristaltic pumps. Although integrated micromachined pumps based on two "one-way" valves may achieve a precise flow control on the order of 1 µl/min with fast response, high sensitivity, and negligible dead volume, these pumps require complicated fabrication processes, generate only modest flow rates and low pressures, and consume a large amount of chip area and considerable power. As a consequence of the low pumping pressure, they often become useless when device dimensions are reduced and resistance to fluid flow and required driving pressure increase. An example of a Si micromachined pump from Animas-Debiotech is shown in Figures 1.10 and 1.11. This belt-worn drug delivery pump features a piezoelectric design pioneered at the

Table 1.8 Comparison of Microfluidics Propulsion Techniques

	Fluid Propulsion Mechanism				
Property	Centrifuge (see Figure 1.14 and Volume I, Figure 6.89)	Pressure (e.g., blister pouch)	Acoustic	Electro-osmosis	Electrowetting
Valving solved	Yes, for liquids	Yes, for liquids and vapor	No	Yes, for liquids	Yes, for liquids
Maturity	Research and development	Products	Research	Products	Research
Propulsion force influenced by	Density and viscosity	Generic	Generic	pH, ionic strength	pH, ionic strength
Power source	Rotary motor	Mechanical roller (see Figure 1.12b)	5–40 V (peak to peak)	10 kV	1–10 kV
Materials	Plastics	Plastics	Piezoelectrics	Glass, plastics	Glass
Flow rate	From less than 1 nL to greater than 100 µl/s	20–500 µl/s	20 µl/s	0.001–1 µl/s	0.3 µl/s

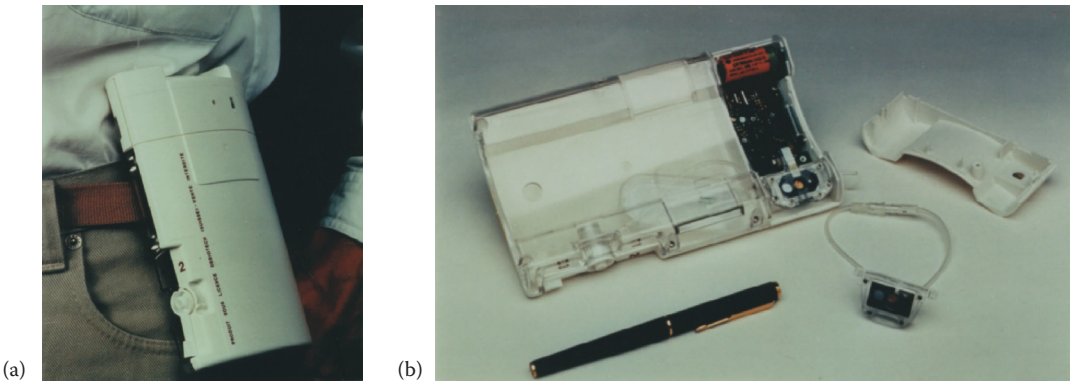


FIGURE 1.10 Micropump technology. (a) Animas-Debiotech's micropump is a Si micromachined piezopump meant to be a belt-worn drug delivery mechanism. The Oncojet is disposable and may deliver a drug at a rate of 10–50 mL/day. (b) The glass-Si sandwich pump is about 2 × 1 cm in size. (Courtesy of Ary Saaman, Animas-Debiotech.)

University of Twente, and a generic example of this MEMS pump is shown in Figure 1.11a.[47] A close-up of the actual Animas-Debiotech MEMS pump is shown in Figure 1.11b. The size of most Si micromachined pumps is quite large, and one wonders why Si was chosen as a substrate (http://www.diabetesnet.com/diabetes_technology/insulinpumps_debiotech.php). Reasonably, plastic molding technology is now being explored as a viable solution.[48,49]

The advantages of using larger peristaltic pumping today are that it relies on well-developed and commercially available components, that a wide range of flow rates is attainable, and that the pumping mechanism is not easily fouled.

Another example of a mechanical pump is embodied in the Johnson & Johnson blister pouch human immunodeficiency virus (HIV) test shown in Figure 1.12.[50] Automated sample preparation and amplification coupled with detection systems would revolutionize the medical diagnostic application of DNA amplification with PCR. An elegant example of progress in this direction comes from Johnson & Johnson (see also US patent 5422271).[50] This company's simple fluidic structure for the detection of the HIV is shown in Figure 1.12. The device was invented at Kodak and bought by Johnson & Johnson in 1994. The disposable pouch-based PCR instrument features a total amplification time of 43 min and a sensitivity and specificity adequate for the detection of HIV. An unknown sample and the PCR cocktail are injected into the amplification compartment through a fill port in the plastic pouch. Eight pouches are

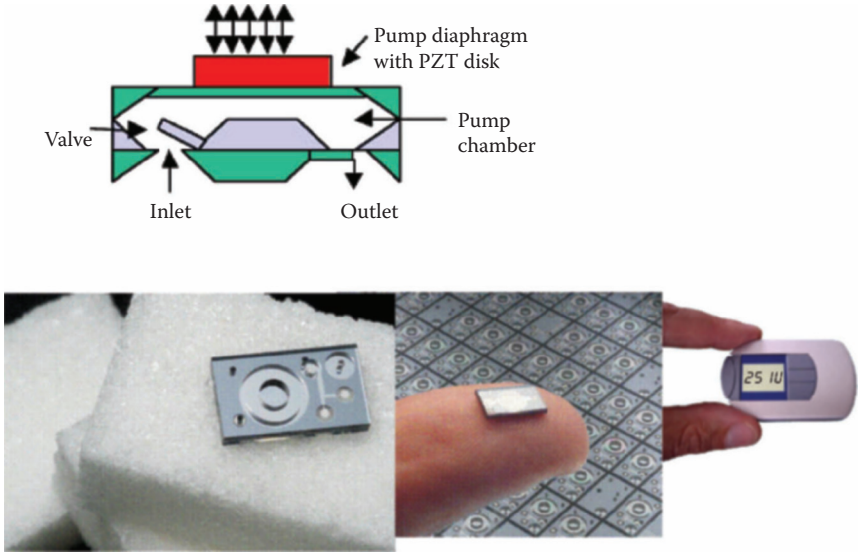


FIGURE 1.11 (a) Generic design of a piezopump.[47] (b) The Animas-Debiotech Si-based microelectromechanical system micropump (http://www.diabetesnet.com/diabetes_technology/insulinpumps_debiotech.php).

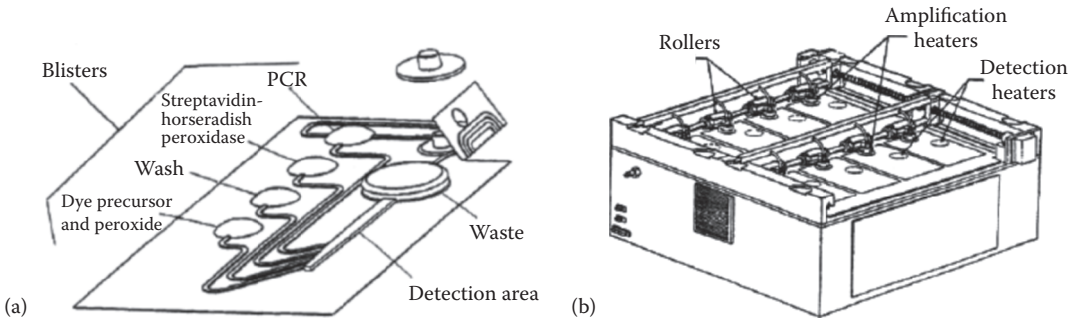


FIGURE 1.12 Johnson & Johnson blister pouch HIV test (a) and tabletop instrument with rollers, amplification, and detection heaters (b). See also reference 50 and US patent 5422271.

aligned on a fixed instrument with the amplification compartments on the upper amplification heaters, and the detector areas on the lower detection heaters. Upon completion of the amplification, the next analysis step is accomplished by an automatic roller that forces detection chemistry fluids and amplified samples from compartments through channels leading to a temperature-controlled detection area. The plastic channels in the pouch contain small nips that close off the channel and break open upon pressurization with a roller system, similar to the one used in credit card image transfer. These nips in the plastic form very inexpensive, single-use, normally closed valves. In the detection zone, biotinylated PCR products are captured via hybridization to probes immobilized on latex particles. Multiple probe sites can be accommodated in the detection areas. Streptavidin-horseradish peroxidase is mixed with those probes to allow for binding between the biotinylated probes and the streptavidin-horseradish peroxidase. Any material left unbound is then removed with the subsequent wash solution. In the presence of peroxidase, leuco dye is oxidized to a visible blue color, indicating a positive hybridization event. Replacing the relatively thick walls of traditional, low surface-to-volume ratio, polypropylene reaction tubes with the pouch "blisters," with their relatively high surface-to-volume ratio and very thin walls (100 μm), which are readily conformable to the shape of the instrument heaters, permits rapid heating and cooling.

The above PCR-based instrument is an example of optimum partitioning between disposable and fixed instrument for use in a medical diagnostic application. For this purpose, sample receptacle, reagents, and all fluidics, such as mixing, valving, and waste area, are integrated in an inexpensive disposable. Power supply, heaters, coolers, readout electronics, etc., are kept separate in an inexpensive disposable plastic pouch. This inexpensive and ingenious way to couple PCR and detection is difficult for a Si-based micromachining approach to compete with. Despite its promise, this HIV test was taken off the market, presumably because of manufacturing problems.

A disadvantage of the pouch approach is that it is rather difficult to further miniaturize because it is a pressure-driven system. The disposable cartridge in the i-STAT (now Abbott) blood analyzer device illustrated in Figure 10.42b of this volume also contains a blister pouch containing calibrant solution.

The proposed glucose meter of our example does not come with an insulin delivery pump, but, in addition to the Animas-Debiotech pump (Figures 1.10 and 1.11), many such insulin delivery pumps are commercially available. Some glucose meters communicate wirelessly to the pump, and some pumps include the meter.

Acoustic Streaming

Acoustic streaming is a constant (DC) fluid motion induced by an oscillating sound field at a solid-fluid boundary (see Volume I, Chapter 6, on fluidics). Transducers do not need to be integrated with the disposable; the cassette with capillary flow channels can simply be laid on top of the acoustic pump network in the reader instrument. Acoustic streaming is also well suited for mixing reagents.[51] The method is considerably more complex to implement than electro-osmosis (see below), but the insensitivity of acoustic streaming to the chemical nature of the fluids inside the fluidic channels

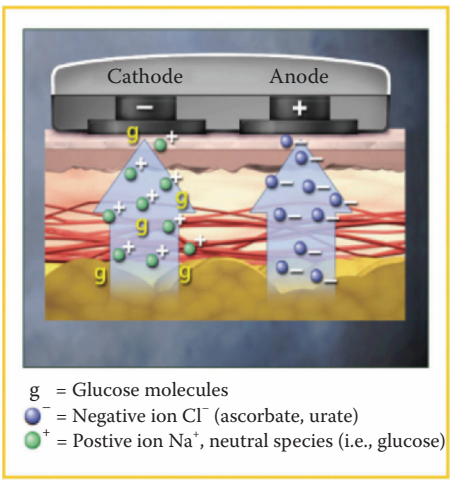
and its ability to mix fluids very effectively make it a potentially viable approach. A typical flow rate measured for water in a small metal pipe lying on a piezoelectric plate is 0.02 mL/s at 40 V, peak to peak.[52] Today acoustic streaming as a propulsion mechanism remains in the research stage.

Electro-osmosis and Iontophoresis

Electro-osmosis is capable of delivering an appreciable DC electrokinetic flow rate in a capillary, does not involve any moving parts, and is easily implemented (see Volume I, Chapter 6, on fluidics). All that is needed is a metal electrode in some type of a reservoir at each end of the flow channel. Typical electro-osmotic flow velocities are on the order of 1 mm/s, with a 1200 V/cm applied electric field. For example, Jorgenson reported electro-osmotic flows of 1.7 mm/s.[53] This is fast enough for most analytical purposes. Some disadvantages of electro-osmosis are the required high voltage (1–30 kV power supply) and direct electrical-to-fluid contact with resulting sensitivity of flow rate to the charge of the capillary wall and to the ionic strength and pH of the solution. It is consequently more difficult to make it into a generic propulsion method. For example, liquids with high ionic strength cause excessive Joule heating; it is therefore difficult or impossible to pump biological fluids such as blood and urine. The sensitivity of the electrokinetic potential to the nature and strength of the electrolyte brings with it the insidious problem of pressure buildup at intersections where liquid columns of different ionic strengths meet. Micromachinists are starting to address the connection between the outside world and the capillaries in electrokinetic equipment. Gluing glass reservoirs for holding the platinum bias electrodes to the glass plate substrate with room temperature vulcanizing (RTV) silicone is a research solution but not an attractive manufacturing option.[54]

In the context of our glucose meter example, we need a blood collection/sampling mechanism. One method for interstitial fluid collection uses iontophoresis. The GlucoWatch® Biographer (cost, $599–799), shown in Figure 1.13, collects glucose through the skin, and no finger pricks are needed (Cygnus Corp.: http://www.glucowatch.com).* In this method, one draws out interstitial fluid by iontophoresis, and the charge and size exclusion properties of the reverse iontophoretic extraction lead to a very clean sample. Iontophoresis utilizes the passage of a constant, low-level electric current (~0.3 mA/cm^2) passed through the skin between two electrodes applied onto the surface of the skin. Electrolyte ions in the body act as the charge carriers for this current to iontophorese substances from within the body outward through the skin. Only small compounds pass through the skin,

(a)



(b)

FIGURE 1.13 The GlucoWatch® G2™ Biographer (a) collects glucose through the skin via iontophoresis (b). (Cygnus Corp., http://www.glucowatch.com; http://www.animas.com.*)

* On March 1, 2006, Cygnus, who made the GlucoWatch Biographer, was closed down, and its assets were bought by Animas (who makes insulin pumps, see Figure). Animas is no longer making GlucoWatch Biographers. The U.S. Food and Drug Administration reports that at least 50% of users experienced mild or moderate skin irritation.

so there are no proteins (e.g., hemoglobin) in the extract. Because the skin has a net negative charge at physiological pH, positively charged sodium ions are the major current carriers. This migration of sodium ions toward the cathode induces an electro-osmotic flow toward this electrode. Uncharged molecules (e.g., glucose) are carried along by this electro-osmotic flow and thus preferentially extracted at the cathode. While glucose is collected at the cathode, interfering species (ascorbate and urate) collect at the anode. Micromolar concentrations of glucose are extracted by the process of iontophoresis compared with 5–10 mM concentrations normally seen in blood samples. To accurately measure this small amount of glucose, the GlucoWatch® Biographer utilizes an amperometric biosensor. Glucose oxidase enzyme located in the hydrogel discs of the AutoSensor reacts selectively with glucose to produce hydrogen peroxide (H_2O_2). The hydrogen peroxide is detected at a platinum-containing working electrode via an oxidation reaction producing an electric current that is measured by the circuitry of the Biographer. After a 2-hour warm-up period and a single-point calibration, the GlucoWatch® G2™ Biographer can provide glucose readings every 10 min for up to 13 hours. Obtaining a blood glucose value from a traditional finger stick glucose meter and then entering the value directly into the Biographer during the calibration phase constitute a single-point calibration. This calibration phase occurs at the end of the 2-hour warm-up cycle. The GlucoWatch provides information on trends and not necessarily in real time for treatment adjustments.

Concerns with the merit of trending only, calibration time, risk of infection and irritation,* and the irreproducibility of the watch eventually spelled the demise of this approach. So we will combine our glucose meter with a more conventional lancet or laser pricking (see below) to obtain a blood sample.

Centrifugal Pumping

Using a rotating disc, centrifugal pumping provides flow rates ranging from less than 10 nL/s to greater than 100 µL/s depending on disc geometry, rotational rate (revolutions per minute), and fluid properties (see Figure 1.14).[55] Pumping is relatively insensitive to physicochemical properties such as pH, ionic strength, or chemical composition (in contrast to AC and DC electrokinetic means of pumping). Aqueous solutions, solvents (e.g., dimethyl sulfoxide), surfactants, and biological fluids (blood, milk, urine) have all been pumped successfully. As explained in Volume I, Chapter 6, fluid gating may be accomplished using "capillary" valves in which capillary forces pin fluids at an enlargement in a channel until rotationally induced pressure is sufficient to overcome the capillary pressure (at the so-called "burst frequency"). Since the types and the amounts of fluids one can pump on a centrifugal platform span a greater dynamic range than for electrokinetic pumps, this approach seems more amenable to sample preparation tasks. The example fluidic platform based on a CD in Figure 1.14 is used by RotaPrep for cell lysis and DNA purification. This specific product uses magnetism, centrifugal force, and capillary action for system operation.

The disposable in this case is an inexpensive polycarbonate disc, containing only fluid manifolds, glass beads, and a metal disc. The permanent instrument has a motor, controls, and permanent magnets.

Vacuum Pressure Reservoir

Fluid transport can also be accomplished by employing chambers within the cassette or reader that are at a higher or lower pressure than the sample and reagent reservoirs. Valves are required to separate

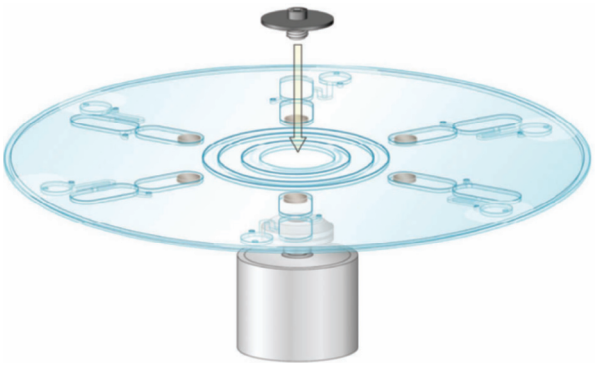


FIGURE 1.14 RotaPrep system for cell lysis (http://www.rotaprep.com). This specific product uses magnetism, centrifugal force, and capillary action for system operation. (Courtesy of Dr. Kido, CEO, RotaPrep.)

* On March 1, 2006, Cygnus, who made the GlucoWatch Biographer, was closed down, and its assets were bought by Animas (who makes insulin pumps). Animas is no longer making GlucoWatch Biographers. The U.S. Food and Drug Administration reports that at least 50% of users experienced mild or moderate skin irritation.

the evacuated or pressurized chambers from the rest of the fluid path until fluid motion is required. Opening a valve initiates the flow of fluid toward an evacuated chamber or away from a pressurized chamber. An advantage of this mechanism over those previously described is that it is a straightforward way to generate a large fluid-motive force. Its drawbacks are that it can only be actuated once, valves may be difficult to implement, and the reliable manufacture of pressurized or evacuated chambers may be challenging. Unless one uses osmotic pumping by implementing a salt reservoir, external actuation seems the safest implementation method for pressure-gradient actuation.

Heating and Cooling and MEMS Integration

Heating and cooling of fluid within a microinstrument, like fluid propulsion, can be accomplished with elements that are either contained within a disposable cassette or the reader. There are, superficially, some arguments to be made for integrating heating and cooling in the disposable. We learned in Volume II, Example 9.1, "Spray Pyrolysis Application: Taguchi Gas Sensor" that electrical current passing through resistors integrated on thin membranes can cause a very rapid temperature increase. This is because the resistor has very little thermal mass and is in intimate thermal contact with the membrane and the fluid to be heated. Integrated heaters are, in principle, simple to fabricate on a variety of materials. On the other hand, temperature control can also be accomplished with a heating and cooling system that is external to the microsystem. For example, the Johnson & Johnson pouch, surveyed earlier, uses two external heaters (Figure 1.12). The heaters are cooled with forced air to speed cycling times. Fast heating and cooling are possible with external heaters. Wittwer and Garling[56] and Wittwer et al.[57] have demonstrated 30-s PCR thermal cycling using a hot-air cycler and glass capillary PCR chambers (see the discussion on PCR in Chapter 10, this volume). External heaters offer the advantages of simpler (less expensive) microinstruments and no cassette-to-reader electrical connections. In general, however, external heaters are slower and require more electrical power than integrated heaters.

A third option is to heat fluid within the microinstrument using radiation generated in the reader. The radiation could be infrared, radio wave, or microwave; it could be adsorbed by the fluid or by absorptive structures (such as black ink) placed on the cassette. An advantage of radiative heating is that it avoids electrical contact from reader to disposable. Our proposed glucose sensor strip has no heater on board.

Sample Introduction

The exact mechanism for introducing samples into the microinstrument depends on the target sample. This is one of the most difficult areas—not only is pretreatment of the sample often required, but one also must transition from a relatively bulky external receptacle structure, which can be handled and accessed manually, to the microfluidic components inside without incurring too much dead space. Very few machining techniques cover the whole dynamic size range required. Because the sample introduction method may place important constraints on the manufacturing process, a few sample introduction options are considered here. Traditional sample injection systems, such as that for flow injection analysis, rely on some sort of commutating valve, either rotary or sliding, actuated manually or mechanically. Harrow and Janata[58] compared different commercially available sample injection systems, and Shoji et al.[59] were among the first to micromachine a sample injector consisting of two three-way microvalves and a channel. These valve systems today are too complex and expensive to consider their integration in a disposable. Perhaps the simplest way to introduce the sample into the microinstrument is by creating wells into which the sample is dropped. For example, most of the substrates with electro-osmotic pumps have plastic tubes glued to them to make sample reservoirs. Similarly, the Biotrack blood coagulation device requires that a drop of sample be placed in a depression on the instrument surface.[60] The sample is then wicked into an internal chamber by capillary action. Problems with sample wells include the large volume that they require and the need to accurately align the sample with the well if the sample well is small. Because the wells are open, one risks sample

contamination or the inclusion of air bubbles in the sample. The latter are a major problem in miniature fluidic systems because of the associated surface tension, and debubblers, which are typically bulky, are often a necessity. Samples may also be injected directly into the microinstrument or disposable cassette. An example of this is the Johnson & Johnson pouch HIV test, where an injection septum is provided in the disposable pouch (Figure 1.12). This reduces the chance of sample contamination but again requires exact alignment of the injection needle with the sample injection port. Furthermore, the manufacture of the cassette pouch is complicated by the need for an injection septum. In an integrated sample, the collection device could be the cassette. For example, the cassette could be manufactured in the form of a blood collection tube (Vacutainer), a syringe, a pipette, or a swab. However, this would complicate the cassette manufacture considerably and may limit the cassette design to a specific type.

For the sample collection of our example glucose sensor, we came to understand earlier that iontophoresis still poses problems. Therefore, we will use a traditional lancet (Figure 1.15a). The Transdermal Biophotonic System from SpectRx (Figure 1.15b) collects interstitial fluid through micropores punctured in the outer layer of the skin by a laser (not FDA approved at this time), and glucose is measured in a patch that contains the glucose sensor. Measuring in interstitial fluids rather than blood, this collection method is similar to the GlucoWatch (see above) in that it provides information only on trends and not necessarily real time for treatment adjustments.

Manufacturing Technique of the Disposable Cassette

The choice of how to partition fluid transport and heating techniques must go hand in hand with the choice of disposable/cassette manufacturing technique. The principal advantages of nontraditional methods for making microinstruments are that very intricate structures with very fine features can be

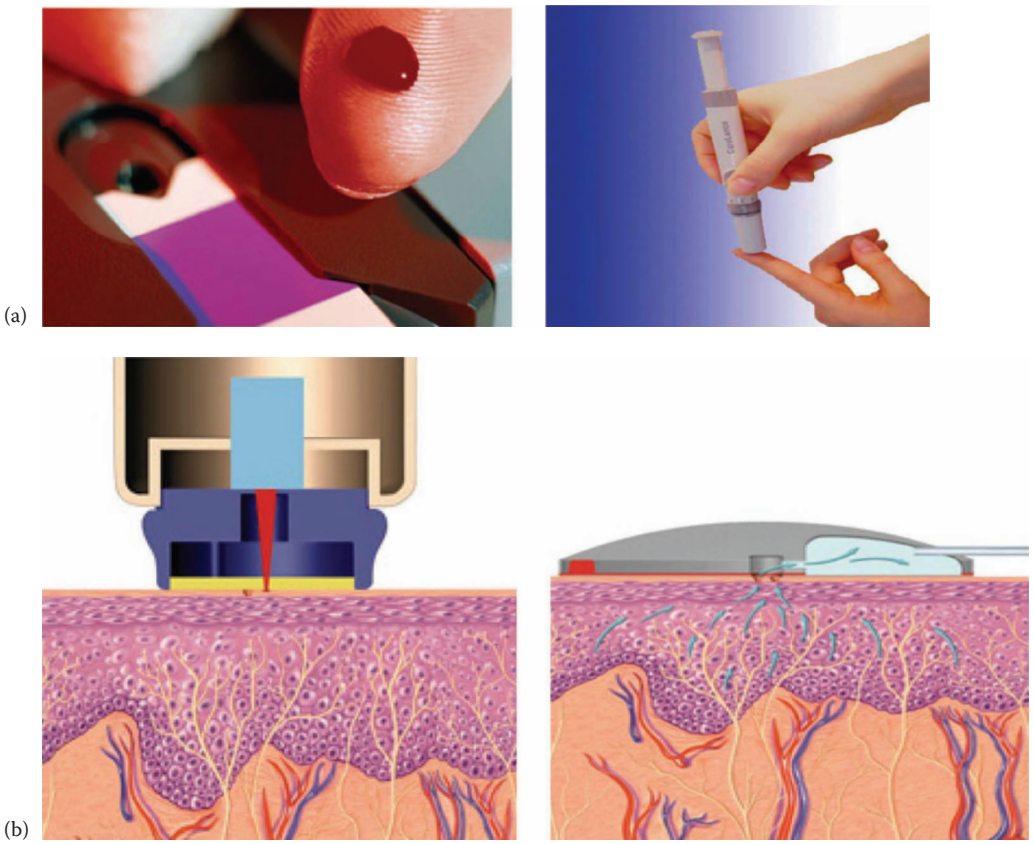


FIGURE 1.15 (a) Lancet for blood collection. (b) The Transdermal Biophotonic System from SpectRx collects interstitial fluid through micropores punctured in the outer layer of skin by a laser (http://www.spectrx.com).

produced, and heaters, electro-osmosis electrodes, and acoustic streaming pumps can be readily realized. On the other hand, nontraditional microfabrication is optimized for making devices that are considerably smaller than most biochemical reaction sample volumes. A 100-µl reagent volume is the same size as a Si chip; that is, 14 mm on a side—huge by microelectronics standards. It is certainly possible to make Si micromachined parts that are as large as the starting wafer (typically 4 in. in diameter); however, the costs of the wafer and wafer processing are likely to be much higher than for other manufacturing techniques. In addition, variation from a previously developed microfabrication sequence can require an expensive and time-consuming process development effort. In the meantime, traditional methods have provided answers to most of the miniaturization needs in diagnostic applications. Where MEMS contributes is in realizing some crucial components of the instrument, for example, the detector, the fluidic conduits, and the actuators. Most solutions rely on plastic rather than Si. We will settle on a choice for our glucose sensor's substrate and manufacturing method after we introduce some background on substrate choices and have launched our check-off list.

Substrate Choice

Here we compare substrate choices such as Si, ceramic, glass, plastics, and others in terms of traditional mechanical machinability, metallization ease, and cost.

For many mechanical sensor applications, single-crystal Si, based on its intrinsic mechanical stability and the feasibility of integrating sensing and electronics on the same substrate, presents an excellent substrate choice. For chemical sensors, on the other hand, Si, with few exceptions,* is merely the substrate and as such it is not necessarily the most attractive option.

In Table 1.9, we show a performance comparison of substrate materials in terms of cost, metallization ease, and machinability. Both ceramic and glass substrates are difficult to machine, and plastic substrates are not readily amenable to metallization. Silicon has the highest material cost per unit area, but this cost can often be offset by the small feature sizes possible in a Si implementation. Silicon, with or without passivating layers, due to its extreme flatness, relative low cost, and well-established coating procedures, is often the preferred substrate—especially for thin films. A lot of thin-film deposition equipment is built to accommodate Si wafers, and as other substrates are harder to accommodate, this lends Si a convenience advantage. There is also a greater flexibility in design and manufacturing with Si technology compared with other substrates. In addition, the initial capital equipment investment, although much more expensive, is not product-specific. Once a first product is on-line, a next generation of new products will require changes in masks and process steps—but not in the equipment itself.

The disadvantages of using Si are usually most pronounced with increasing device size and low production volumes and when electronics do not need to be, or cannot be, incorporated on the same Si substrate. The latter could be either for cost reasons (e.g., in the case of disposables, such as glucose sensors) or for technological reasons (e.g., the devices will be immersed in conductive liquids, or they must operate at temperatures above 150°C).

An overwhelming determining factor for substrate choice is the final package of the device. A chemical sensor on an insulating substrate is almost always easier to package than a piece of Si with conductive edges in need of insulation.

Table 1.9 Performance Comparison of Substrate Materials

Substrate	Cost	Metallization	Machinability
Ceramic	Medium	Fair	Poor
Plastic	Low	Poor	Fair
Silicon	High	Good	Very good
Glass	Low	Good	Poor

* The ion-sensitive field effect transistor (ISFET) is one example where chemistry and electronics are intimately integrated.[61] The difficulties encountered while integrating chemistry with electronics in an ISFET were highlighted in Volume I, Chapter 7. This device, thirty years after its invention, is finally commercially available. A second example is Nanogen's high-density DNA array (see Volume II, Chapter 8, on chemical-, photochemical-, and electrochemical-based forming). The problem of wiring a very high-density DNA array (e.g., 10,000 sites) makes on-site electronics attractive in this case, outweighing both packaging and cost issues.

As we pointed out earlier, packaging is so important in sensors that, as a rule, sensor design should start from the package rather than from the sensor. In this context, an easier-to-package substrate has a huge advantage. The latter is the most important reason why chemical sensor development in the industry retrenched from an all-out move toward integration on Si in the 1970s and early 1980s to a hybrid thick film on ceramic approach in the late 1980s and early 1990s. In US academic circles, chemical sensor integration with electronics continued until the late 1980s, stopped for a while, and then re-emerged with the successful introduction of a commercial ion-sensitive field effect transistor (ISFET)[61] and DNA arrays with integrated electronics;[62] in Europe and Japan, such efforts were never completely abandoned.

A good engineering guideline to determine whether Si is suitable for a given mechanical application is the following: there should be at least two benefits arising from the use of Si over other substrates (aside from the possibility of integrating the electronics on the same substrate). For example, in the case of a torsional micromirror, as shown in Figure 1.16, fabricated based on a silicon-on-insulator (SOI) approach, the two Si-derived benefits are 1) excellent mirror-like surface and 2) superior torsional behavior of the Si mirror hinges. On the basis of these two important properties, Si has become a formidable contender for making arrays of micromirrors for optical switching. Note that the mirror in Figure 1.16 is not optimized for a high-density mirror switch array. For such an application, the hinges should be very short or hidden behind the mirror, as the mirror surface exposed to the incoming light should be maximized. The design as shown could be used, for example, for a linear scanner.

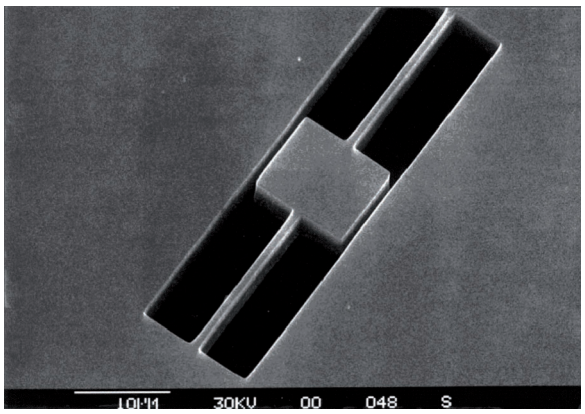


FIGURE 1.16 Silicon-on-insulator (SOI) micromirror.

In Table 1.10 we compare some of the material properties for single-crystal SiC, Si, GaAs, and diamond with relevance to their role as substrates in electronics and MEMS. Not enough good data are available yet to assemble a similar table for the polycrystalline version of the listed semiconductors. From a comparison of Si and GaAs for micromachining applications in this table, we conclude that GaAs is a better material for thermal isolation and for higher-temperature operation. Single-crystal GaAs lends itself as a material for micromechanics because of its attractive optoelectronic and thermal properties. The material is less attractive for mechanical devices, with a yield load smaller by a factor of 2 compared with Si. The many heterostructures possible with GaAs make a wide variety of optical components, such as lasers and optical waveguides, feasible. The piezoelectric effect enables piezoelectric transducers. On the basis of the high-electron mobility, the material is also ideal for the measurement of a magnetic field through the Hall effect. "Macro" pressure, temperature, and vibration sensors, making use of the influence of external pressure and temperature on the band gap, have been built.

Micromachining is proving to be a good method to incorporate micromirrors in resonant optical cavities for tunable lasers. Moreover, micromachining using gallium arsenide and group III–V compound semiconductors is a practical way to integrate radio frequency switches, antennas, and other custom high-frequency components with ultrahigh-speed electronic devices for wireless communications (see Chapter 10, this volume).

Checklist

Some twenty years ago, Stephen Senturia (Massachusetts Institute of Technology) posed a rhetorical question about building micromachines: "Can we design microrobotic devices without knowing the mechanical properties of the materials involved?" In subsequent years, characterizing mechanical properties of MEMS materials became popular. However, an equally important question—"Would

Table 1.10 Material Properties of Four Important MEMS Materials at 300 K

Property	3C-SiC	GaAs	Si	Diamond
Melting point (°C)	2,830 (pressure is 35 bar; decomposes)	1,238	1,415	4,000 (phase change occurs)
Maximum operating temperature (°C)	873	460	300	1,100
Thermal conductivity (W/cm °C)	4.9	0.5	1.57; comparable to metals such as carbon steel (0.97) and Al (2.36)	20
Linear thermal expansion coefficient (× 10–6 °C–1)	4.7	5.9	2.35; the low expansion coefficient of Si is closer to that of quartz (7.1) than that of a metal (e.g., 25 for Al), making it insensitive to thermal shock	0.08
Young's modulus (GPa)	448	75	190 (111); the elastic modulus is similar to that of steel (206–235)	1,035
Physical stability	Excellent	Fair; sublimation of As is a problem	Good	Excellent
Energy gap (eV)	2.39	1.42, direct	1.12, indirect	5.5
Chemical resistance	Very good	Poor	Good	Excellent
Electron mobility (cm2/V s)	1,000	8,500	1,500	2,200
Hole mobility (cm2/V s)	50	400	600	1,600
Density	3.2	5.3	2.32; lower density than Al (2.7); thus, it has a high stiffness-to-weight ratio	3.5
Yield strength (GPa)	21 (for 6H-SiC)	2.0	7 (steel is 2.1); IC-grade Si is stronger than steel	53
Breakdown voltage (× 106 V/cm)	2	0.4	0.3	10
Dielectric constant	9.7	13.1	11.9	5.5
Lattice constant (Å)	4.36	5.65	5.43	3.57
Knoop hardness (kg/mm2)	3,980	600	1,000 (stainless steel is 660)	10,000
Saturated electron drift velocity (× 107 cm/s)	2.2	2	1.1	2.7

you build a microstructure without knowing all the available construction methods?"—often remains unaddressed. Choices of materials and machining processes are intertwined, and both must be confronted early on in the preliminary design phase.

In Table 1.11, we introduce a checklist to help make the decision of optimized manufacturing option and substrate material easier. For the checklist, which serves as a control gate, we start from the assumption that the market and specifications list are well understood and that the detailed specifications list, scaling appreciation, and market understanding agree best with one particular sensing approach (in the example of the glucose sensor, we decided on an amperometric approach).

The checklist starts with a question about the design of the interface of the micromachine with the real world—in other words, with its package. Once the package design is established, one can look for the best manufacturing technique for the micromachined part inside. This strategy dictates

Table 1.11 Checklist to Determine Substrate and Machining Approach for a New Micromachining Application

1. What will the package of the sensor or system be, and how does it interface with the real world? The package and the interface with the environment determine size and cost of the total product and the nature of the microdevice inside, as well as the answers to most of the following questions.
2. Is there a need for integration of electronic functions in the micromachine? Such need (e.g., because the sensor has a very high impedance) necessitates a semiconductor substrate. Above, we proposed a rule of thumb to justify Si as a substrate; there must be at least two compelling reasons besides the possibility of integrated electronics [for the Si on insulator (SOI) micromirror in Figure 1.16, e.g., the reasons are the optical quality mirror surface of silicon and its excellent torsional behavior].
3. How many of the microdevices will be made (production volume or number of units) and what is the unit complexity (number of devices in a unit)? The number might suggest a serial (small number), hybrid (large number), batch (very large number), or even a continuous (moving web-based) approach.
4. Cost per device? The cost may suggest a serial (high cost, >$40) or Si batch (low cost, <$2) approach. For very low cost (glucose sensor strips, <$0.20), a very large batch or continuous fabrication process becomes mandatory.
5. Does the substrate merely function as a support? If so, glass, ceramic, or even plastic and cardboard all become options. If the substrate has a mechanical function, Si is an excellent candidate. If the substrate must have good optical properties, materials such as GaAs and poly(methylmethacrylate) are candidates.
6. Is there a need for modularity? Modularity is important with chemical sensor arrays, in which integration often is counterproductive because of the incompatibility issues faced when depositing different chemical sensor coatings on an integrated chip.
7. What are the expected relative tolerances on the lateral dimensions, and what is the required depth-to-width aspect ratio of features built into the substrate? Very small relative tolerances on lateral dimensions cannot be achieved yet; 1% lateral tolerance on a 100-μm line, in optical lithography, is considered good, but 10^{-3} on 1 cm in diamond milling is very poor (see Volume II, Figure 6.3). Large aspect ratios (say, >20) might necessitate wet anisotropic etching of a single-crystalline material, room temperature and cryogenic dry etching, or LIGA and pseudo-LIGA processes.
8. To what environment (air, water, or other) will the device be exposed? Sensors exposed to aqueous environments such as blood present more packaging problems and make integration of electronics difficult.
9. Which substrate makes packaging requirements less stringent? For a sensor in aqueous solutions, for example, a ceramic substrate requires no protection of the sides of the ceramic strip; a Si sensor, on the other hand, is difficult to insulate and package since a conductive medium might electrically short out the chemical sensor signal via the conductive Si sidewalls.
10. Is the desired microstructure a truly 3D part (e.g., a contact lens), or is it a projected form? A truly 3D part suggests traditional precision machining, for example, diamond turning or a molding process; projected forms allow for a lithography process.
11. What are the thermal requirements?
    - Maximum temperature of exposure (temperatures >150°C may make integrated Si functions impossible). If integration and high temperature are needed, specially processed Si, SOI, or GaAs become candidates.
    - Is thermal conductivity needed?
    - Is a thermal match with other materials important?
12. What are the flatness requirements (often in connection with the optical properties of the substrate)?
    - Average roughness, Ra?
    - One or both sides polished?
13. What are the optical requirements?
    - Transparency in certain wavelength regions?
    - Index of refraction?
    - Reflectivity?
14. What are the electrical and magnetic requirements?
    - Conductor versus insulator?
    - Dielectric constant?
    - Magnetic properties?
15. What is the process compatibility?
    - Is the substrate part of the process?
    - Chemical compatibility?
    - Ease of metallization?
    - Machinability?
16. What are the strain-dependent properties?
    - Piezoresistivity?
    - Piezoelectricity?
    - Fracture behavior?
    - Young's modulus?

itself, because the packaging cost contributes largely to the overall cost of any micromachined product and often overwhelms size specifications as well. The reason for expensive packaging can be ascribed to the many serial labor processes involved. Also, each device may require individual attention. The size of the package is dictated by the need to interface to a bigger structure, by limitations of traditional machining methods, or by the need for manual handling of the structure. Micromachining techniques such as fusion bonding, anodic bonding, and other integrated encapsulation techniques, commonly used in bulk and surface micromachining, might help transform more and more of the serial packaging steps into parallel batch steps even for ICs, which will prove advantageous in keeping costs of the total structure down. In the author's opinion, micromachined batch packaging solutions represent one of the most unmined micromachining opportunities.

Next in the checklist, one needs to establish whether to integrate active electronics on the substrate and confirm the expected temperature range of operation (if too high, the latter might void the integration option). A further consideration concerns number and cost of the devices and which substrate will make packaging easier. When faced with a new micromachining problem, one should consider all available precision manufacturing options—that is, zero-base the approach, be willing to forget all preconceived notions about how the part should be manufactured.

Different markets dictate different machining approaches. For example, chemical sensors, serving a fragmented market, are better constructed in a modular hybrid thick-film approach, but automotive accelerometers, serving a huge and uniform market, are better built in an integrated polysilicon design. The choice of materials for miniaturization of sensors, actuators, and power sources will depend also on the particular sensing principle employed.

We now return once more to the glucose meter example summarized in Figure 1.17. In this exercise, the first round of brainstorming (based on the specification list, and the scaling input) for a disposable glucose sensor resulted in the suggestion of an amperometric approach. We are ready now to face the first design iteration by looking at material and machining options for the proposed electrochemical glucose sensor. For implementation of the preferred sensor materials and manufacturing technology, we challenge the electrochemical conversion option with the list of manufacturing options presented in Tables 1.2 to 1.6 and substrate materials options (Tables 1.9 and 1.10). In the decision process, we use the checklist in Table 1.11 as the control gate. For the glucose sensor this results in thick-film sensor electrodes on disposable plastic (polyvinyl chloride) as the best suggested approach. Screen printing is a well-known technique for making patterns on various substrates (i.e., paper, plastic, wood, glass, ceramics, etc.). This coincides with the principal solution pursued by industry. See, for example, the MediSense/Abbott sensor, the Precision QID, which was shown in Volume I, Figure 7.80. This sensor consists basically of a two-electrode system with a carbon working electrode (carbon ink) and counter—and reference—electrode combined in a Ag/AgCl electrode (Ag/AgCl ink) (Figure 1.18). The working electrode is covered with another ink for the enzyme/mediator system.

Example 1.1: Proposed Scenario for Continuous Manufacture of Polymer/ Metal-Based Biosensors

While the number of new tools available for microfabrication has grown dramatically, few methods have been gainfully applied to disposable biosensor construction for clinical applications. In contrast to mechanical sensors, where the excellent mechanical properties of single-crystalline Si often tend to favor Si technology, the choice of the optimum manufacturing technology for chemical sensors is far less evident. For chemical sensors, Si in the sensor often serves no other role than that of a passive substrate. This past decade the following trend emerged: mechanical sensors (pressure, acceleration, temperature, etc.) moved toward more integration (for example, in accelerometers with CMOS-compatible surface micromachining), while chemical sensors and biosensors moved away from integration and from Si as a substrate. For the latter, hybrid technology on plastic or ceramic substrates with silk screening

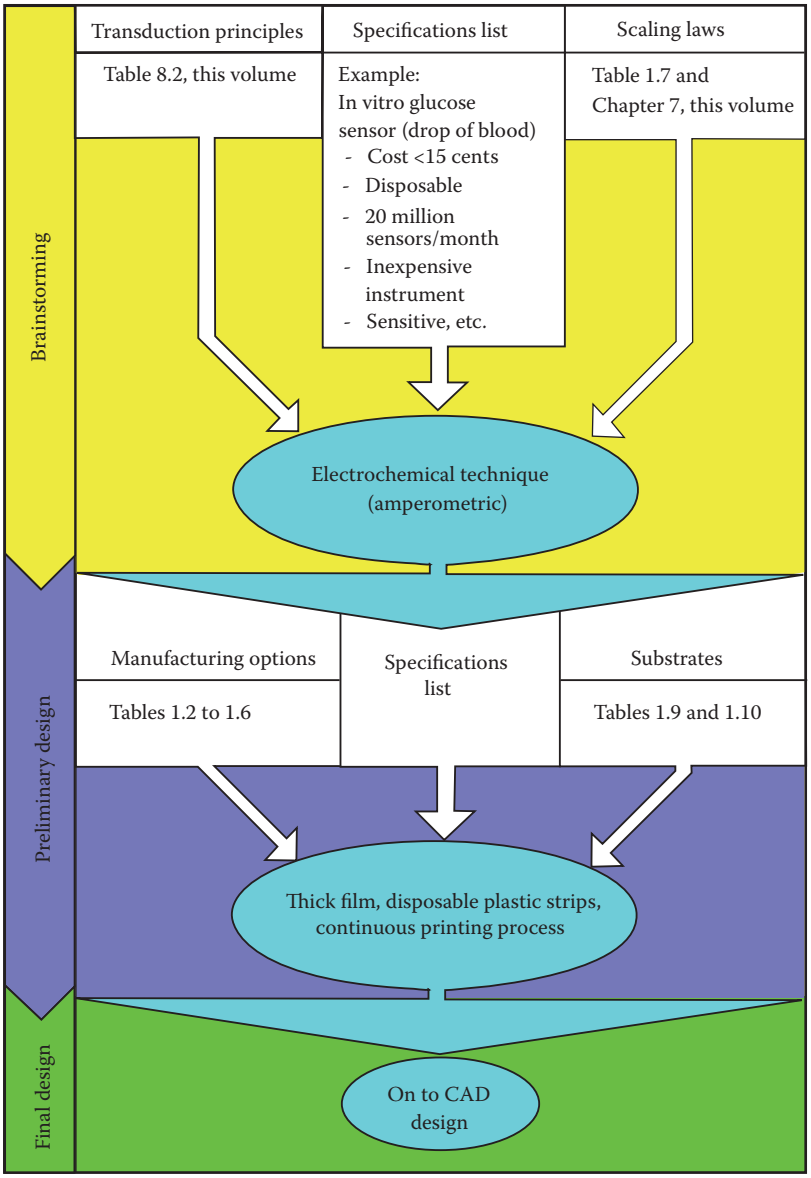


FIGURE 1.17 Confrontation of detailed specification list with sensing/transduction principles and scaling considerations in the brainstorming phase and subsequent searching of manufacturing options by going over a check-off list in the preliminary design phase in the process of developing a disposable glucose sensor (see text).

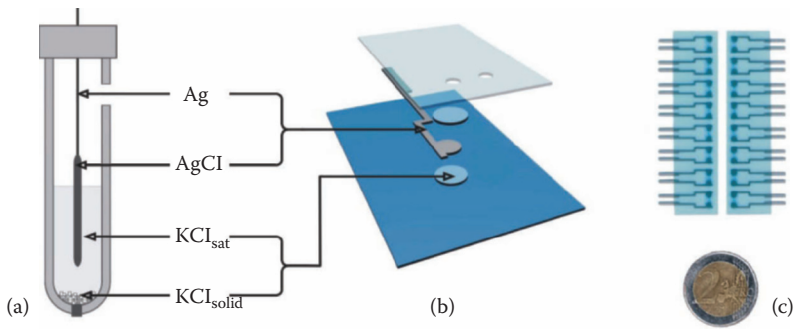
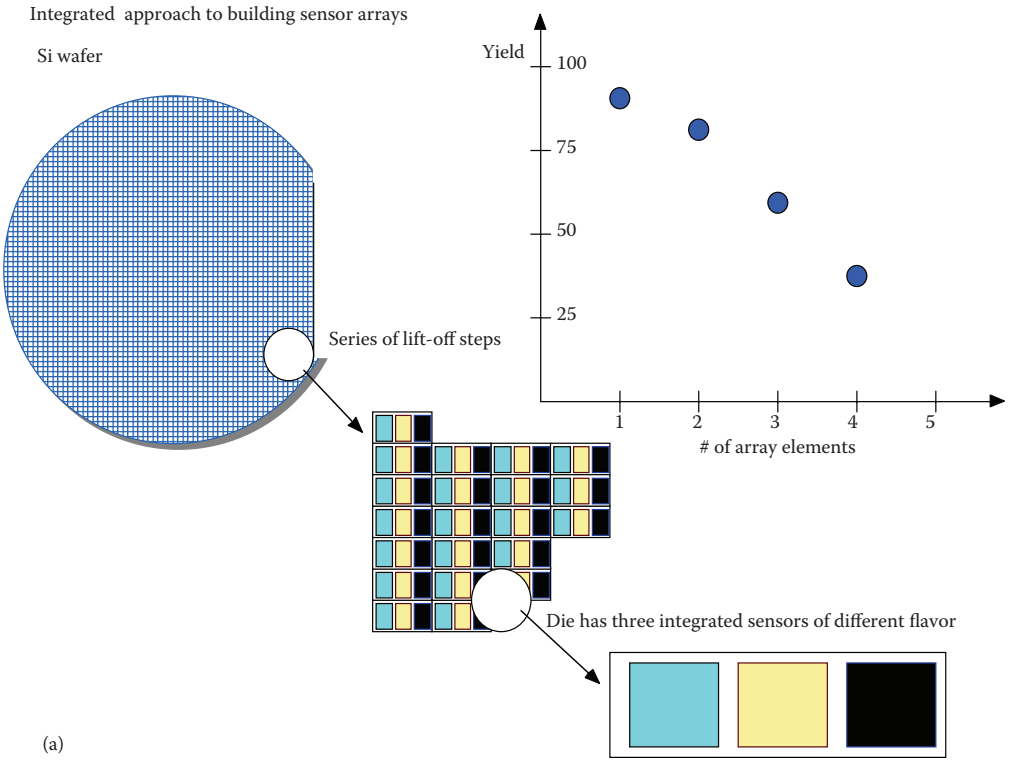


FIGURE 1.18 Structure of a conventional reference electrode of Ag/AgCl/KCl type (a) and design of a planar reference electrode fabricated by means of screen printing (b and c).



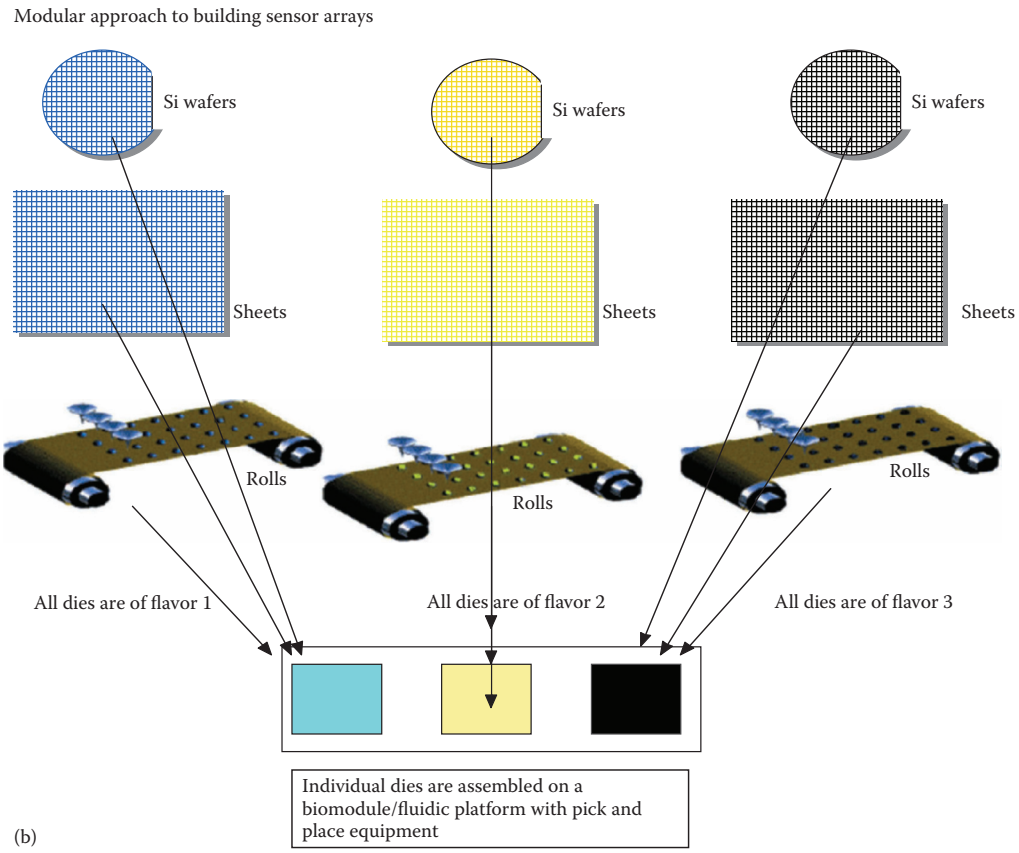


FIGURE 1.19 (a) Integrated sensor array fabrication method. (b) Modular approach to making a chemical sensor array. Wafers, sheets, or rolls of substrate with sensors of one type only are optimized separately to ensure optimum yield. They are then cut out and with pick-and-place equipment put into a biomodule/fluidic platform. Different sensor panels can be put together easily, as there are no compatibility issues to deal with.[63]

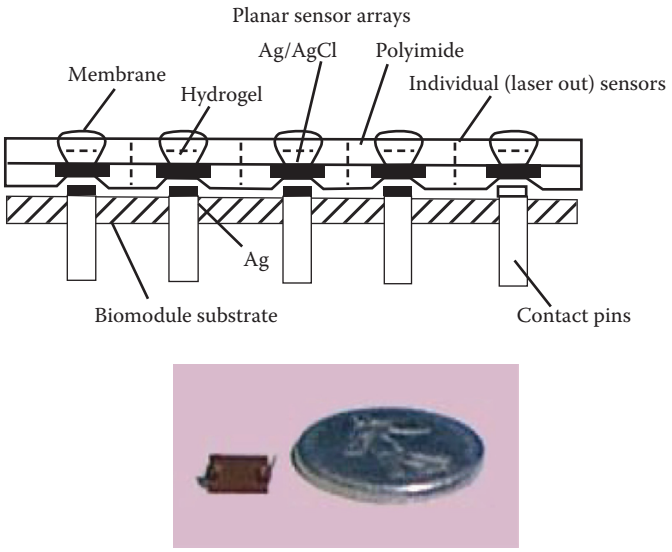


FIGURE 1.20 The individual biosensors [ion-selective electrodes (ISEs)] are combined in a so-called "biomodule" with pick-and-place equipment.[63]

or drop delivery systems for the application of the organic layers became more popular. Microfabrication of Si-based electrochemical sensors, using IC technology, has been especially challenging, mainly due to process incompatibility issues, packaging problems, failure to incorporate a true reference electrode, and the difficulties involved in patterning relatively thick organic layers such as ion-selective membranes and hydrogels [see also i-STAT (Abbott Point of Care, http://abbottpointofcare.com) in Chapter 10 of this volume]. The low cost requirements and the tremendous variety and fragmentation of biomedical sensor applications necessitates a nonsilicon, modular approach. Having addressed materials choice and modularity, various competing manufacturing processes must be compared. Today, in fabricating somewhat larger disposable biosensors, continuous manufacturing processes are used that yield much less expensive devices than silicon batch processes can afford. As an example, consider the mass production of amperometric glucose sensors where the cost per sensor target is 10 cents. Using a Si batch approach, it is very difficult to make a glucose sensor for less than $1 (or any disposable biosensor for that matter). The current process to mass produce glucose sensors involves such proven technology as doctor's blade on a continuous moving web, making a 10-cent cost per sensor quite possible. It is our belief that for disposable microsensors to become an economic reality such continuous or semicontinuous manufacturing processes will have to be further developed and adapted to microsensor construction. In Figure 1.19, we contrast an integrated approach to chemical sensor array fabrication with a more modular manufacturing approach.[63] The integrated sensor array with three types of sensing functions on a silicon substrate, depending on process compatibility, often has a rapidly decreasing yield. The more modular approach shows how one type of sensor might be made on a silicon wafer, a plastic sheet, or a continuous roll. The biosensors fabricated here are electrochemical sensors that can be combined in "biomodules" with pick-and-place equipment after they have been laser-cut from their individual processed sheets (see Figure 1.20). By making sensors of only one type on the same substrate, this approach allows for the fabrication of a sensor array composed of sensors that may otherwise have fabrication incompatibilities, thus increasing the yield of the final array. Since any panel could be put together in a factory upon demand without reconsidering process incompatibilities, more products could be introduced into the market all at once.

The ultimate goal is a manufacturing process for disposable sensors that merges traditional machining options with IC-based manufacturing options and is more akin to packaging processes (e.g., drop delivery, pick-and-place) than to the

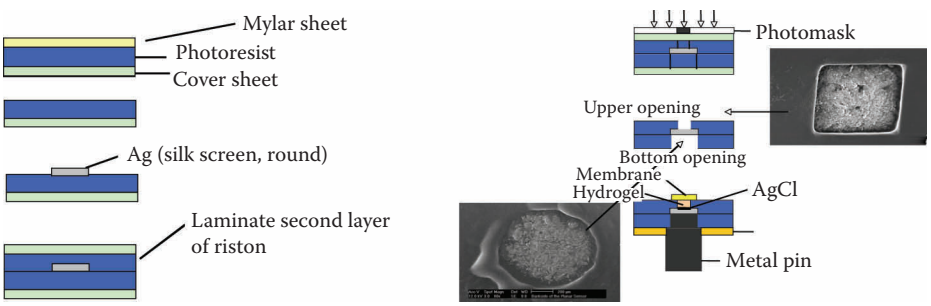


FIGURE 1.21 Schematic illustration of the procedure for fabrication of an individual array element.[63]

Table 1.12 Basic Properties of the Different Photoresist Materials Used to Build Fluidic Elements[63]

Photoresist	XP SU-8	Pyralin PI2721 (Photosensitive Polyimide)	Pyralux PC1025 Dry Resist Film	Riston 4600 Dry Resist Film
Resist tone	Negative	Negative	Negative	Negative
Developer	PGMEA*	DE6180	1% K_2CO_3	1% K_2CO_3
Etching profile	Vertical sidewall	Undercut	Vertical sidewall	Vertical sidewall
Smallest feature	2 μm	2 μm	20 μm	20 μm
Compatibility with plastic substrate in developing	No (surface coating and back cover are needed)	No (surface coating and back cover are needed)	Yes	Yes
Thickness of single layer	Any thickness up to 1 mm	~20 μm	64 μm	30 μm
Bonding to dry resist film	Excellent	Excellent	Excellent	Excellent
Adhesion to substrate	Excellent	Excellent	Excellent	Good
Uniformity	Good	Fair	Excellent	Excellent
Flexibility	Not applicable	Not applicable	Excellent	Brittle
Developing	Vibration needed	Vibration needed	Washing needed	Easy and quick
Multicoating availability	Yes	Yes	Multilayers by lamination	Multilayers by lamination

*PGMEA, propylene glycol methyl ether acetate.

front end part of the IC industry (e.g., lift-off, integration).

As a concrete example, the manufacturing procedure of an individual ion selective electrode array element, say, for sensing potassium, is illustrated in Figure 1.21. These electrochemical planar sensors may be made on a variety of polymer substrates; currently we use 5 × 5 in dry negative photoresist sheets (e.g., Pyralux). Pyralux comes in rolls, so it will eventually be possible to make these types of sensors in a continuous process (as depicted in Figure 1.19). An important simplification in the sensor manufacture depicted in Figure 1.21 is the self-aligned step. By exposing the dry resist from the top through the photomask, both cavities in the photoresist bilayer are made at the same time. The top cavity results from the pattern on the mask while the bottom cavity originates from the silk-screened Ag pattern, which acts as a mask for the lower part of the resist. No expensive double exposure system is required.

Fluid handling for calibration or washing steps may be implemented on the basis of the same dry photoresist technology. Fluidic structures have been made directly in all of the various resist systems listed in Table 1.12. The use of two or more photoresists in combination has enabled the fabrication of photopatternable multilayer microfluidic structures. For example, Madou et al.[63] have sealed photopatterned microchannels of SU-8 with sheets of the flexible photoresist, Pyralux, as shown in Figure 1.22a. Subsequent patterning of the Pyralux allows for the creation of more complex three-dimensional fluidic structures, as seen from Figure 1.22b. The

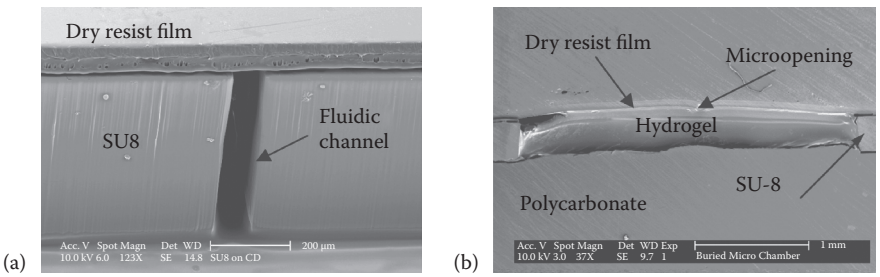


FIGURE 1.22 Buried SU-8 microchannel (a) and multilevel microchamber filled with hydrogel (b) (to be used as an onboard electrochemical reference electrode).[63]

latter structure can be used, for example, to seal liquids or hydrogels such as for an onboard electrochemical reference electrode.

The above "beyond batch" fabrication sequence is generic and applies equally well if optical sensor probes have been selected and will make it possible to fabricate biosensors in an affordable manner. This will lead to a wide variety of additional biomedical products currently impossible to fabricate using silicon technology.

Questions

1.1: List three different methods one can utilize to deposit a thin-metal film on a MEMS structure.

1.2: What do the following acronyms stand for? MEMS, VLSI, DUV, DOF, SEM, NGL, NA, MTF, OAI, OPC, RIE, and CMOS.

1.3: Compare laser-machining (LM) with e-beam machining (EBM).

1.4: The following bulleted list includes examples of micromaching tools:

(a) Classify them according to the applied energy appearance (W: wet chemical and electrochemical machining; M: mechanical machining; E: electrothermal machining) at the workpiece.

(b) Classify them according to machining methods (S: subtractive; A: additive; S/A: subtractive and additive).

- Photofabrication
- Laser beam machining
- Diamond milling
- Chemical and electrochemical milling
- Electron beam machining
- Photochemical milling
- Dry etching
- Ultrasonic machining
- Plasma beam machining
- Abrasive jet machining
- Electroplating and electroless plating
- Stereolithography
- Electrodischarge machining

1.5: Provide a comparison of the economic and technical aspects in the construction of sensors with thin-film technology, IC fabrication (semiconductor substrates), thick film, and classic construction.

1.6: Why use miniaturization technology?

1.7: Get calibrated: What is a mil or a thou? How many microns in a mil? How many thousandths of an inch (thou or mil) are there in a millimeter? Approximately how thick is a human hair? Approximately how big is a virus? A bacterium? A red blood cell? How thick is a standard sheet of paper?

1.8: How small is small? There are many manufacturing processes other than silicon micromachining for making small parts. Do some research to find specifications for various manufacturing processes, small devices, and tools. Search the Internet, call companies, use the Yellow Pages, use a Thomas's Register, ask a machinist, ask your instructor, etc. We're looking for the lower limits on traditional machining or tools that are currently commercially available at a local job shop or a tool distributor. The objective is to get a feel for the boundary between MEMS and traditional machining.

1.9: Reconstruct the decision tree to reach an optimized micromanufacturing option for a given miniaturization problem.

1.10: Explain the statement that MEMS engineers should design from the MEMS package inwards.

1.11: Make a comparison of additive thin- versus thick-film technology.

1.12: Explain the progress from serial to batch to beyond batch manufacturing (R2R).

1.13: Characterize objects made by laser machining, mechanical machining, E-beam machining, and plastic molding using the following criteria:

- Projected versus 3D
- Serial, batch, or continuous manufacturing
- Top-down or bottom-up machining

References

1. Block, B. Personal communication.
2. Van Doren, C. 1991. *A history of knowledge.* New York: Ballantine Books.

3. Middlehoek, S., and U. Dauderstadt. 1994. Haben Mikrosensoren aus Silizium eine Zukunft? *Technische Rundschau* July:102–05.
4. Vasile, M. J., C. Biddick, and S. A. Schwalm. 1993. Microfabrication by ion milling: The lathe technique. *Micromechanical systems*. A. P. Pisano, J. Jara-Almonte, and W. Trimmer, Eds. New York: ASME Press, p. 81.
5. Vasile, M. J., C. Biddick, and S. A. Schwalm. 1994. Microfabrication by ion milling: The lathe technique. *J Vac Sci Technol* B12:2388–93.
6. Harris, T. W. 1976. *Chemical milling.* Oxford, United Kingdom: Clarendon Press.
7. Romankiw, L. T. 1997. Electroforming of electronic devices. *J Plating & Surf Finish* 84:10–16.
8. Romankiw, L. T. 1997. Evolution of the plating through lithographic mask technology. *Electrochim Acta* 42:2985–3005.
9. Phillips, R. E. 1986. What is electrochemical grinding and how does it work? *Carbide Tool* 18:12–14.
10. Kalpajian, S. 1984. *Manufacturing processes for engineering materials.* Reading, MA: Addison-Wesley.
11. Saito, N. 1964. Mechanism of electric discharge machining. *Bull Jpn Soc Precis Eng* 1:39.
12. Taniguchi, N. 1984. Research on and development of energy beam processing of materials in Japan. *Bull Jpn Soc Precis Eng* 18:117–25.
13. Harper, C. A., ed. 1982. *Handbook of thick film hybrid microelectronics.* New York: McGraw-Hill.
14. Fiori, C., and G. De Portu. 1986. Tape casting: a technique for preparing and studying new materials. *Br Ceram Proc* 38:213–25.
15. Helvajian, H., ed. 1995. *Microengineering technology for space systems*. El Segundo, CA: Aerospace Corporation.
16. Pfender, E. 1988. Fundamental studies associated with the plasma spray process. *Surf Coat Technol* 34:1–14.
17. UMN. http://www.nrri.umn.edu/NLTC/SLA.htm.
18. Boothroyd, G., and W. A. Knight. 1989. *Fundamentals of machining and machine tools.* New York: Marcel Dekker.
19. Bellows, G., and J. B. Kohls. 1982. Drilling without drills. *Am Mach Special Report* 743:187.
20. Li, B., P. A. Clark, and K. H. Church. 2007. Robust direct-write dispensing tool and solutions for micro/meso-scale manufacturing and packaging. *Proceedings of the 2007 international manufacturing science and engineering conference*. Atlanta, GA.
21. Sakai, T. 1992. State of the art injection moulding of high performance ceramics. *Adv Polym Technol* 11:53–67.
22. Rak, Z. S. 1998. From technical ceramics to high-performance ceramics. *Ceramic Forum Int* 75: 19–24.
23. Belloy, E., P. Q. Pham, A. Sayah, and M. A. M. Gijs. 2000. In *Eurosensors XIV,* eds. R. de Reus and S. Bouwstra, 255–56. Copenhagen, Denmark: MIC-Mikroelektronik Centret.
24. Waterjet. http://www.waterjets.org.
25. Trotter, D. M. 1991. Photochromic and photosensitive glass. *Sci Am* 124:124–29.
26. Allen, D. M. 1986. *The principles and practice of photochemical machining and photoetching.* Bristol, United Kingdom: Adam Hilger.
27. Kern, W. 1978. Chemical etching of silicon, germanium, gallium arsenide, and gallium phosphide. *RCA Rev* 39:278–308.
28. Manos, D. M., and D. L. Flamm, eds. 1989. *Plasma etching: An introduction.* Boston: Academic Press.
29. Johnstone, R. W., and M. Parameswaran. 2004. *An introduction to surface-micromachining*. New York: Springer.
30. Diem, B., M. T. Delaye, F. Michel, S. Renard, and G. Delapierre. 1993. SOI (SIMOX) as a substrate for surface micromachining of single crystalline silicon sensors and actuators. *Transducers '93*. Yokohama, Japan.
31. Ehrfeld, W. 1990. The LIGA process for microsystems. *Proceedings: Micro system technologies '90.* Berlin, pp. 521–28.
32. Ahn, C. H., Y. J. Kim, and M. G. Allen. 1993. A planar variable reluctance magnetic micromotor with fully integrated stator and wrapped coils. *Proceedings: IEEE micro electro mechanical systems (MEMS '93)*. Fort Lauderdale, FL, February 7–10, pp. 1–6.
33. SOTEC. http://dmtsun.epfl.ch/~lguerin/sotec.html.
34. Keller, C., and M. Ferrari. 1994. In *Technical digest: 1994 solid state sensor and actuator workshop*, 132–37. Hilton Head Island, SC.
35. Pister, K. S. J. 1992. In *Technical digest: 1992 solid state sensor and actuator workshop*, 136–39. Hilton Head Island, SC.
36. Bride, J. A., S. Baskaran, N. Taylor, J. W. Halloran, W. H. Juan, S. W. Pang, and M. O'Donnell. 1993. Photolithographic micromolding of ceramics using plasma etched polyimide patterns. *Appl Phys Lett* 63:3379–81.
37. Zhao, X.-M., Y. Xia, and G. M. Whitesides. 1996. Fabrication of three-dimensional microstructures: microtransfer molding. *Adv Mater* 8:837–40.
38. Sander, D., R. Hoffmann, V. Reiling, and J. Muller. 1995. Fabrication of metallic microstructures by electroplating using deep-etched silicon molds. *J Microelectromech Syst* 4:81–86.
39. Shaw, M. C. 1984. *Metal cutting principles*. Oxford, United Kingdom: Clarendon Press.
40. DeVries, W. R. 1992. *Analysis of material removal processes.* New York: Springer-Verlag.
41. Slocum, A. H. 1992. *Precision machine design.* Englewood Cliffs, NJ: Prentice Hall.
42. Evans, C. 1989. *Precision engineering: An evolutionary view.* Bedford, United Kingdom: Cranfield Press.
43. Lambrechts, M., and W. Sansen. 1992. *Biosensors: Microelectrical devices.* Philadelphia: Institute of Physics Publishing.
44. Lebbink, G. K. 1994. Microsystem technology: Een niet te missen ontwikkeling. *Mikroniek* 3:70–73.
45. Bergveld, P. 1995. In *Micro total analysis systems*, eds. A. van den Berg and P. Bergveld, 1–4. Twente, the Netherlands: Kluwer Academic Publishers.
46. Madou, M. J., and G. J. Kellogg. 1998. *Progress in biomedical optics*. Eds. G. E. Cohn and A. Katzir, 80–93. San Jose, CA: SPIE.
47. van Lintel, H. T. G., F. C. M. van der Pol, and S. Bouwstra. 1988. A piezoelectric micropump based on micromachining of silicon. *Sensors Actuators* 15:153–67.
48. Schomburg, W. K., J. Fahrenberg, D. Maas, and R. Rapp. 1993. Active valves and pumps for microfluidics. *J Micromech Microeng* 3:216–18.
49. Nguyen, N. T., X. Y. Huang, and K. C. Toh. 2002. MEMS—micropumps: A review. *J Fluid Eng* 124:384–92.
50. Findlay, J. B., S. M. Atwood, L. Bergmeyer, J. Chemelli, K. Christy, T. Cummins, W. Donish, T. Ekeze, J. Falvo, and D. Patterson. 1993. Automated closed-vessel system for in vitro diagnostics based on polymerase chain reaction. *Clin Chem* 39:1927–33. See also United States Patent 5422271.

51. Tsao, T. R., R. M. Moroney, B. A. Martin, and R. M. White. 1991. Electrochemical detection of localized mixing produced by ultrasonic flexural waves. *IEEE 1991 ultrasonics symposium proceedings*. Orlando, FL.
52. Miyazaki, S., T. Kawai, and M. Araragi. 1991. A piezo-electric pump driven by a flexural progressive wave. *Proceedings: IEEE micro electro mechanical systems (MEMS '91)*, 283–88. Nara, Japan.
53. Jorgenson, J. W., and E. J. Guthrie. 1983. Liquid chromatography in open-tubular columns. *J Chrom* 255: 335–48.
54. Jacobson, S. C., L. B. Kounty, R. Hergenroder, A. W. Moore, and J. M. Ramsey. 1994: Microchip capillary electrophoresis with an integrated postcolumn reactor. *Anal Chem* 66:3472–76.
55. Kellogg, G. J., T. E. Arnold, B. L. Carvalho, D. C. Duffy, and N. F. Sheppard. 2000. Centrifugal microfluidics: applications. *Micro total analysis systems*. Eds. A. van den Berg, W. Olthuis, and P. Bergveld. Enschede, the Netherlands: Kluwer Academic Publishers, pp. 239–42.
56. Wittwer, C. T., and D. J. Garling. 1991. Rapid cycle DNA amplification: time and temperature optimization. *BioTechniques* 10:76–83.
57. Wittwer, C. T., G. C. Fillmore, and D. J. Garling. 1990. Minimizing the time required for DNA amplifications by efficient heat transfer to small samples. *Anal Biochem* 186:328–31.
58. Harrow, J. J., and J. Janata. 1983. Comparison of sample injection systems for flow injection analysis. *Anal Chem* 55:2461–63.
59. Shoji, S., S. Nakagawa, and M. Esashi. 1990. Micropump and sample-injector for integrated chemical analyzing systems. *Sensors Actuators A* A21:189–92.
60. Hillman, R. S., M. E. Cobb, J. D. Allen, I. Gibbons, V. E. Ostoich, and L. J. Winfrey. 1990. Capillary flow device. US Patent 4963498.
61. Sandifer, J. R., and J. J. Voycheck. 1999. A review of biosensor and industrial applications of pH-ISFETs and an evaluation of Honeywell's "DuraFET." *Mikrochim Acta* 131:91–98.
62. Thewes, R. 2002. Sensor arrays for fully-electronic DNA detection on CMOS. *IEEE international solid-state circuits conference (ISSCC)*. San Francisco, CA.
63. Madou, M., and J. Florkey. 2000. From batch to continuous manufacturing of microbiomedical devices. *Chem Rev* 100:2679–91.
64. Christensen, C. M. 2006. Hewlett–Packard: The Flight of the Kittyhawk(A). http://cb.hbsp.harvard.edu/cb/web/product_detail.seam?R=606088-pdf-ENG&ConversationId=51866.
65. Ikuta, K., K. Hirowatari, and T. Ogata. 1994. Three dimensional integrated fluid systems (MIFS) fabricated by Stereo Lithography. In *IEEE International Workshop on Micro Electro Mechanical Systems (MEMS '94)*, pp. 1–6. Oiso, Japan.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2

Nature as an Engineering Guide: Biomimetics

Outline

Introduction

Icarus Revisited?

Nature's Engineering Feats

Emergence of Life: Biogenesis

DNA and RNA: The Software of Life

Proteins: Hardware of Life and Nature's Robots

The Genetic Code

Cloning

Therapeutic Cloning, or Biomedical Cloning

Stem Cells

Reproductive Cloning and Stem Cell Research: Implications in Bioethics

Biological Cells and Their Machines

Organelles

Mitochondria and Power

Hydrophobic and Hydrophilic Forces Drive Self-Assembly in Nature

From Micelles to Liposomes and Biological Cells

Biotechnology and Biomimetics

DNA Chips

Biomimetics: Some Concluding Remarks

Appendix 2A: Uses of Extremophiles

Appendix 2B: The Twenty Amino Acids Found in Biological Systems

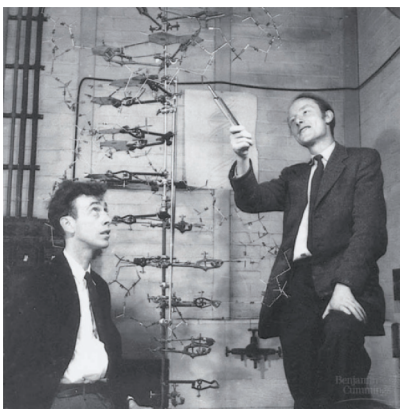
Appendix 2C: The Universal Code

Questions

References

What then is the vocation of the whole man? So far as I can make out, his vocation is to be a creator: and if you ask me, creator of what, I answer—creator of real values … And if you ask me what motive can be appealed to, what driving power can be relied on, to bring out the creative element in men and women, there is only one answer I can give; but I give it without hesitation—the love of beauty, innate in everybody, but suppressed, smothered, thwarted in most of us …

L. P. Jacks
The Education of the Whole Man



Watson and Crick with DNA model.

Fall of Icarus. Mimicking nature—biomimetics. The Flemish artist Peter Paul Rubens (1636–1638) depicts Daedalus looking on as his son Icarus is falling from the sky. The picture is sometimes used to illustrate human arrogance. In the current context, the painting illustrates an early attempt at biomimetics: the fixing of bird wings onto human arms with beeswax. When Icarus flew too close to the sun, the wax melted.

Introduction

In biomimetics, one studies how nature, building atom by atom, i.e., through bottom-up manufacturing, through eons of evolution of life, developed materials, structures, processes, and intelligence to inspire and improve the engineering and design of artificial materials, human-made structures and processes (e.g., software). Human manufacturing technology works in the opposite direction, i.e., it builds top-down; in most current manufacturing we tend to start with larger building blocks and use stiff materials (e.g., Si or stainless steel), whereas nature prefers small building blocks and mostly soft, low-Young's modulus materials (e.g., muscle or skin). Throughout history, biomimetics has been attempted but often with less than satisfactory results. Bird flight, for example, did not lead to aircraft, but mathematical expressions from aerodynamics did. As a consequence, from the middle of the eighteenth century to about thirty years ago, engineers were tempted to engineer around nature's obstacles rather than be inspired by nature itself. Today though, in fields ranging from artificial intelligence to microelectromechanical systems (MEMS), nanoelectromechanical systems (NEMS), and smart materials, the perceived advantages of bottom-up designs and manufacturing are convincing many scientists to research natural, biomimetic approaches and manufacturing methods. In this chapter, we first introduce the most important engineering feats of nature and then explain how these are applied in today's biotechnology. It can be argued that molecular scientists and genetic engineers were practicing nanotechnology long before the name became popular with electrical and mechanical engineers. Actually molecular biology or "wet nanotechnology" has been called by some "nanotechnology that works." To be fair, molecular biology discoveries were rarely seen in the light of improved manufacturing techniques, as the discoverers were usually aiming for improved diagnostics and therapeutics. Within this context we introduce the use of natural polymers and their mutants not only for optimized sensing and realizing new drugs but also for the fabrication of actuators and new manufacturing techniques in general.

We start this chapter on a philosophical note, inquiring into the wisdom of biomimetics. We then detail a long list of nature's engineering feats. We try to define life on Earth and other planets, list the many ideas about the emergence of life, introduce DNA and RNA as the software of life and proteins as its hardware, and explain the build of a biological cell and the genetic code. We dwell on how nature generates power, while biomimetic approaches to packaging of nanocomponents are covered in Chapter 4, this volume, on packaging, assembly, and self-assembly, and on implementation of artificial intelligence in Chapter 9, on power and brains in miniature devices.

After introducing a definition of biotechnology we present examples of the use of MEMS and NEMS in biotechnology. We finish this chapter with a comparison of human- and nature-engineered structures.

This chapter is also meant as a bridge from *dry* engineering, in air and with rigid materials (mechanical and electrical engineering), to *wet* engineering, with flexible materials in aqueous solutions (biology and molecular engineering). Miniaturization science in 2011 is still practiced predominantly by electrical and mechanical engineers, but as applications are becoming more and more biological in nature, biologists, materials scientists, chemists, and physicists must work together with electrical and mechanical engineers to develop new miniaturization solutions. We hope that this chapter will further stimulate such collaborations.

Icarus Revisited?

Biomimetic-oriented manufacturing is a hot topic today, but some caution might be in order. As the legend of Icarus reminds us, at least some previous attempts at biomimetics have resulted in failure. The chapter-opening *Fall of Icarus* illustrates an early attempt at biomimetics. Icarus and his father Daedalus fixed bird wings onto their arms with beeswax. When Icarus flew too close to the sun, the wax melted and he fell to his death in the Mediterranean.

Smallness is the ancestral condition of life. Large organisms are built bottom-up from cells rather than divided into cells. Cells are, in turn, fabricated from yet smaller entities. A single *Escherichia coli* bacterium,

for example, is shaped like a cylinder about 1 μm across and 3 μm long. It contains at least 3000 different molecular parts and will divide every 20 min when nutrients are available. No wonder scientists are intrigued by the miniaturization skills displayed by nature! But zoologist Steven Vogel, in his book *Cats' Paws and Catapults*,[1] asks a question very relevant to this chapter: "Who then is the better technologist? Mother Nature or the human engineer?" Natural design and biomimetics are very popular today and, according to Vogel, in danger of being overused or misinterpreted. Vogel presents a balanced view on the relative merits of natural and human mechanical design. He points out that although nature often exemplifies desired features, mimicking it has proved useful on surprisingly few occasions (as the Icarus legend attests). Vogel's explanation is that human and natural mechanical designs are two individually well-integrated technologies, but they are within very separate contexts. Each might be uniquely integrated by its own elements of internal harmony and consistency, but an impressive aspect of one may have little relevance to the other. A prominent example is metallurgy. In human technology, metals enable stamping, forging, casting, grounding, slicing, and sawing. No known organism uses pieces of pure metal for any mechanical purpose; they make stiff materials into artifacts by internal growth and surface deposition. So humanity's diverse array of metal manufacturing techniques is of no value in nature. Nature, on the other hand, does not use steel or favor the production of flat surfaces and sharp corners—all very useful in human manufacturing. Nature builds with proteins and produces mostly curved surfaces and rounded corners, resulting in such masterfully engineered objects as biological cells. Both natural and human manufacturing approaches have their merits within their own proper frame of reference. As a consequence, Vogel maintains, most attempts at mimicking nature have failed; we do not fly in aircraft with flapping wings or use wood to build skyscrapers. Vogel's comparison of human and nature's engineering feats principally deals with large designs, where he finds them competitive.

However, it is at the nanoscale level that nature truly excels, and perhaps Vogel's contention might not hold true in nanoengineering. In the nanoworld, nature is way ahead of human engineering, as it works with much smaller, more versatile building blocks and has mastered the self-assembly and multiparallel construction of those building blocks to create large, hierarchically arranged organisms that are much more complex than even the most complex human machines. Today, using human engineering techniques such as MEMS and NEMS, we are not able to build a single micromachine displaying anywhere near the complexity of even the simplest biological cell.

With biomimetic bottom-up manufacturing techniques, we are able to synthesize natural and mutant nucleic acids and proteins. Using well-established approaches such as recombinant DNA, gene shuffling, and knockout mutations, we can alter the genome of a cell, but there has been little progress made assembling an artificial cell or even the simplest organelle. But all this might be changing soon: at the end of 2007, a team of 20 scientists constructed a synthetic chromosome that is 381 genes long and contains 582,970 base pairs of genetic code, using lab-made chemicals at the J. Craig Venter Institute.[2] Prior to this work, the largest synthesized DNA contained only 32,000 base pairs. The J. Craig Venter team achieved its technical breakthrough by chemically making DNA fragments in the lab and developing new methods for their assembly and reproduction. After several years of work perfecting chemical assembly, the team members found that they could use homologous recombination (a process that cells use to repair damage to their chromosomes) in the yeast *Saccharomyces cerevisiae* to rapidly construct the entire artificial bacterial chromosome from large synthesized subassemblies. This team then started working toward the ultimate goal of inserting a synthetic chromosome into a cell and booting it up to create the first synthetic organism. The artificial chromosome was to be transplanted into a living bacterial cell, and in the final stage of the process, it was anticipated to take control of the cell and in effect become a new life form. The new life form would still depend for its ability to replicate itself and metabolize on the molecular machinery of the cell into which it has been introduced, and in that sense, it would not be a wholly synthetic life form, rather one more in tune with a parasitic

life form. In May 2010, Venter's team did indeed create such a "synthetic cell," i.e., a normal bacterial cell controlled by DNA made in the lab. The Venter team first decoded the genome of the bacterium *Mycoplasma mycoides*. Then they purchased about 1000 bits of DNA about 1000 letters long that make up that bacterial genome. These bits were stitched together in the right order by putting them into yeast cells. The completed synthetic genome, containing almost 1000 genes, was then put into a different kind of bacterial cell, *Mycoplasma capricolum*, where it replaced the native DNA. The man-made DNA booted up the host cell, which began to replicate, making millions of *M. mycoides* cells.[70] Possible applications of this technology are unlimited. For example, in medicine, imagine that synthetic genes could be introduced into individuals suffering from a genetic disease, where they are missing genes or have damaged ones. The ability to correct the original faulty gene could potentially constitute a cure for a number of genetic diseases. Since building large genomes is now feasible and scalable, other important applications, such as biofuels, can be envisioned. Along this line, Venter's team is working already with Exxon Mobil to develop algae that soak up carbon dioxide and convert it into hydrocarbon fuel, and on new vaccines with Novartis. The hope is that many other complex, engineered microbes will be developed to clean up water, produce new forms of food, textiles, pharmaceuticals and industrial chemicals in ways that naturally occurring microbes cannot achieve.

The reason that nature is leading in nanotechnology may be sought in the fact that evolution has worked much longer on developing a single biological cell (3.5 billion years for the first life-forms to evolve) than on the larger life-forms (*Homo sapiens sapiens* first appeared only about 120,000 years ago). In other words, many designs were experimented with over a much longer period of time to come to the optimally designed single-cell life forms than for the larger living creatures such as trees and humans. It will perhaps be through nanochemistry that humans can aspire to reach the same level of sophistication in the manufacture of nanostructures as nature does. Moreover, the flight of birds did inspire the Wright brothers' first successful fixed-wing aircraft. But, as we are cautioned to do, the Wright brothers did not blindly mimic nature: they used fixed-wing aircraft, not the flapping bird wings of Icarus. That is, one should learn from nature's manufacturing techniques and adapt them to human/societal needs rather than blindly mimic king nature's approaches. After all, nature's manufacturing techniques are geared toward maximizing procreation, not toward maximizing speed and efficiency of product manufacturing techniques in a modern human societal context.

As humans learn to build with the same construction set as nature does, we are bound to challenge nature in nanoengineering. Because it uses relatively large building blocks, human manufacturing is rapid, and the expectation is that nature, because it uses much smaller building blocks (for example, atoms with a diameter of 0.3 nm), must be very slow. To offset the time it takes to work with small, basic building blocks, nature, in growing an organism, relies on an additive process featuring massive parallelism and self-assembly, the latter of which is introduced as one of the key characteristics of bottom-up manufacturing, detailed in Chapter 3, this volume.

Nature's Engineering Feats

Life?

The question of how exactly to define life has been a constant dilemma throughout humankind's history. Scientists of the stature of Schrödinger, Fermi, Arrhenius, Hoyle, von Helmholtz, Eigen, Haldane, von Neumann, Urey, Oparin, Monod, Dyson, de Duve, and Crick, to name a few, have been preoccupied with this question. From the plethora of ideas and the many attempts to characterize life that emerged from these thinkers, we select a few that promise the longest staying power. We refer to "ideas" and "characterizations" of life rather than "definitions," as no definition has yet been presented that is accurate enough.[3]

One of the better ideas comes from the book *Life Itself: Its Origin and Nature* by Francis Crick, who describes life as a process that is able to directly replicate its own instructions and indirectly replicate any machinery needed to execute those instructions.[4] Others consider something to be alive if it exhibits a minimal set of features that include,

Table 2.1 Characteristics Associated with Living Things

Characteristic	Remarks
Typical chemistry	Proteins, nucleic acids, fatty acids, sugars and water. Several of these chemicals can also be found in nonliving matter.
Very high degree of complexity	The incredible complexity guarantees the unpredictability of organisms. The weather is also very complex though and few would consider it alive.
Self-organization	Perhaps more important than complexity is organized complexity. Then again, there are also self-assembled monolayers and micelles that we would not call alive.
Chirality	Life on Earth almost exclusively uses left-handed amino acids and right-handed DNA. Experiments show that proteins cannot form if both right- and left-handed DNA are present.
Self-repair	The object is capable of rectifying errors in either its metabolic or genetic machinery.
Autonomy or self-determination	Spontaneity, inner freedom—perhaps the most enigmatic characteristic of life to explain, and one that sets life apart from nonlife.
Decrease in entropy	The increasing complexity of a living organism goes hand in hand with a decreasing entropy. Unfortunately, such a definition also characterizes a refrigerator.
Metabolism and homeostasis	Taking up substances (nutrition) to garner energy and excrete others reaching an equilibrium body shape. But what does this characterization make of a dormant microorganism?
Self-recognition	Distinguish foreign from self and reject foreign.
Sensitivity	Response to environmental stimuli.
Information content	The information needed to replicate an organism is encoded in the genes.
Replication	Identical offspring are generated. Then again, many molecules and crystals form templates for replication.
Evolution (natural selection)	Populations change over time.
Reproduction	Offspring is more than a facsimile; it also must incorporate a copy of the replication mechanism. Slight differences in the offspring are implied. Clays can replicate as well as reproduce (see text).
Growth and development	Growth and gradual evolutionary development, often from a single cell, to more complex structures. Crystals also grow, but there is no actual evolutionary adaptation.
Heredity	Transmit information from generation to generation with some slight changes embedded.

for example, metabolism, self-repair, and replication. A more complete set of characteristics associated with living "things" is summarized in Table 2.1. Not all organisms exhibit all these features all at once, but a combination of some of these characteristics can describe almost anything that can be defined as *alive* on Earth.

From Table 2.1, we learn that thermodynamics offers one of the many descriptions of life. In a living system, materials and free energy (ΔG) flow into the organism, and waste and heat (ΔH) flow out of it. Living things, from the lowly bacteria to the more sophisticated, intelligent humans, all are highly ordered compared with their constituent molecules and seem to violate the law of increasing entropy (second law of thermodynamics). The second law of thermodynamics states that in any spontaneous process, there always is an increase in entropy. Living creatures have spontaneously developed into more ordered and sophisticated systems through evolution with an accompanying decrease in entropy (ΔS). This apparent violation of the second law of thermodynamics is easily explained if one recognizes that the evolutionary process takes place within the broader context of a universe in which chaos keeps on increasing. In more rigorous thermodynamic terms, the argument goes as follows: the entropy of the universe may be expressed as the sum of the system under study (life) and its surroundings, as follows.

$$\Delta S_{\text{universe}} = \Delta S_{\text{system}} + \Delta S_{\text{surroundings}} \qquad (2.1)$$

In life, large molecules assemble from simple, small ones into those that are increasingly complex; that is, ΔS_{system} is negative. To reconcile this with Equation 2.1, we must remember that $\Delta S_{universe}$ must be positive, which can be achieved only if $\Delta S_{surroundings}$ is large and positive.[5*]

To illustrate how weak some of these "definitions" of life are, just consider a thermodynamic definition, often presented in the literature: a living system is one that can maintain itself in a state of low entropy. However, a definition in which a system reducing its entropy through interaction with its environment is considered alive would also qualify your refrigerator—or the weather, for that matter.

Computer science (through modeling of artificial life), mathematics (through complexity theory), and information theory all contribute their own definitions. The following are some of the basic tenets. In artificial life, or a-life, computer-virus-like entities are created in an artificial computer world (e.g., Tom Ray's *Tierra* and *Network Tierra*);[6] these creatures reproduce, mutate, evolve, and die. Some scientists actually have considered these digital entities alive. Critics of a-life are skeptical because of the many simplifications and the arbitrariness of its rules. Independent of the issue of a-life being alive or not, important information about what life is and how it evolved might be gathered in days rather than over millions or billions of years. Indeed, because of the simulation power of today's computers, Darwinian evolution scenarios, involving creatures of varying complexity and different degrees of environmental pressures, can be played out over short periods of time.

The continuing study of complex systems may result in a better definition of *life*. Complexity scientists suggest that life is an "emergent" or "self-organizational" property that arises spontaneously when a broth of different chemicals reaches a certain level of complexity.[3] Others are adamant that it is the principle of information economy in self-developing systems that provided the guiding principle of biological evolution. Artificial life and complexity theory is touched upon in Chapter 9, on power and brains in miniature devices, and in "Emergence of Life: Biogenesis" below. Some artificial life Web sites can be found at http://alife .org and http://www.alcyone.com/max/links/alife .html. It is perhaps best at this point to admit that, today at least, because of its complexity, we cannot define what life really is and that describing life obviously is easier than defining it.

Life on Other Planets?

To garner a better understanding of what life is, we might look for it on other planets. The search for extraterrestrial intelligence (SETI) (http://setiathome.ssl.berkeley.edu) is privately funded and kept somewhat at arms' length by NASA out of fear of criticism by overzealous politicians, who would rather spend tax dollars finding weapons of mass destruction. Use of miniaturized systems figures prominently in this quest, and NASA is an important funding source for miniaturization science. It is beyond the scope of this book to dwell on SETI, but a brief introduction of the Drake equation, which predicts the probability of finding intelligent life on other planets, might motivate interested readers to explore the topic further. The Drake equation has been a starting point for almost all theoretical speculation on the existence of extraterrestrial intelligence since 1960.[7] The equation is a product of a large number of elements, most of which are only vaguely known:

$$N = R^* \times \underbrace{f_p \times n_e}_{\text{Physical}} \times \underbrace{f_l \times f_i}_{\text{Biological}} \times \underbrace{f_c \times L}_{\text{Cultural}} \qquad (2.2)$$

The elements forming the Drake equation are as follows.

Physical Elements

R^* = the rate at which stars are formed in our galaxy per year; f_p = the fraction of stars, once formed, that will have a planetary system; n_e = the number of planets in each planetary system with the right environment suitable for life.

* But where did the order of the universe as a whole come from? The second law of thermodynamics must have been violated during creation! Pointing out that the expanding universe allows increasing room for order to form could deflate this point, but then one may further ask where did the energy and mass of the universe originate? The first law of thermodynamics (conservation of energy) must have been violated at creation. This point may perhaps be refuted as well since the positive energy of the universe is exactly balanced by the negative gravitational potential energy: E = k(inetic) + mc^2 (matter/energy) – V_g (gravitational) = 0. So perhaps no energy was required to produce the universe?

Biological Elements

f_l = the probability that life will develop on a suitable planet; f_i = the probability that life will develop to an intelligent state.

Cultural Elements

f_c = the probability that intelligent life will develop a culture capable of communicating over interstellar distances; L = the time (in years) that such a culture will spend trying to communicate.

We know that *N* is no less than 1, as manifested by life on Earth. Beyond that, the widest range of *N* values has been proposed—from 1 all the way up to >100 million! *Paradigms Lost* by John Casti (1989)[7] presents an excellent treatise on the subject. Casti revisited the topic in 2000[8] and points out that, since his *Paradigms Lost* of 1989, the evidence for planetary systems outside our own solar system has become overwhelming, increasing n_e in Equation 2.2.

Most of the "life" sensors that NASA has developed so far are based on the assumption that life on other planets is similar to that on Earth and are based on sensors for flow of typical Earth-type feed materials (e.g., oxygen) and Earth-type waste materials (e.g., carbon dioxide), or they may be heat flow detectors. A few such "life" detectors were tested on Mars. Oyama, for example, in a Mars *Viking* experiment, wetted a soil sample with amino acids and other nutrients and then monitored common metabolic gases, such as carbon dioxide, oxygen, and hydrogen.[9] Perhaps too many experiments designed to find life on other planets are based on the expectation of Earth-like life and a specific molecular structure rather than on an abstract system featuring properties such as adaptability, flexibility, reproduction, and self-organization. Life elsewhere might well involve molecules other than DNA.

Crick, in his 1981 *Life Itself: Its Origins and Nature,*[4] details the Panspermia theory, quite controversial at the time, in which life on Earth came from somewhere else in the universe. This, of course, puts the question about what life is, or how it emerged, on another planet and does not answer the underlying question. Interestingly, the Panspermia theory has found new supporters. In August 1996, it was announced that features that were explained as life forms more primitive than those on Earth were found in a 4-pound meteorite retrieved on Earth but originating from Mars* named ALH84001[8] (Figure 2.1). A mineralogical study concluded that the temperature of the rock never rose much above that of boiling water, and a living microbe or spore deeper inside would have survived the voyage. There have also been new suggestions that extremely sturdy life forms might reach Earth unscathed via dust or small objects such as comets.[10,11] We do, indeed, know now that some types of meteorites contain carbon and organic chemicals, including amino acids.[11] In this context, the amino acid glycine was found inside the galactic cloud Sagittarius B2 in 1994.[8] An issue that the cosmic seed theory does address is how life on Earth could get off to such a quick start; it might have arrived nearly fully developed!

FIGURE 2.1 ALH84001. A rock from Mars retrieved on Earth.

Emergence of Life: Biogenesis

Introduction

Religious debates aside (Genesis with creationism, vitalism, reincarnation, etc.), and ignoring earlier suggested life forces such as psyche, air, lifeblood, phlogiston, electricity, and radioactivity, the question about the "emergence" of life has been even more controversial than the question about what life itself is. The two questions, as suggested in the preceding section, are intertwined: if we knew how

* The 1996 NASA announcement of potential proof for life on Mars is still mired in controversy relating to life originating from inorganic or organic sources.

life evolved, we might be able to define it better. There are several schools of thought on the issue. We briefly review the most prominent ones.

DNA, RNA, and Proteins: Who Was on First?

Four billion years ago, there was so much cratering on the moon and on Earth's surface that no life could have taken hold. The oldest fossilized bacteria date the start of life on Earth to 3.8 billion years ago. The starting point of "life," in a "narrow" window between 3.8 and 4 billion years ago, might have been a simple molecule with the ability to catalyze its own formation from raw materials found in its immediate surroundings—*replication through autocatalysis*. The period of "competing chemistries" before terrestrial life developed is known as the *prebiotic* phase. One or several self-replicating "mother molecules" must have led to the molecules of life we have come to know: deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. In Table 2.2 we summarize some of the most common terms/definitions used in the context of the emergence of life (biogenesis) and genetics.

Many of the debates have focused around the primacy of DNA versus RNA versus proteins. We discuss life's functional polymers—DNA, RNA, and proteins—and the genetic code in more detail below. For now, it will suffice to expose the purported roles of DNA, RNA, and proteins in the emergence of life.

1. *DNA first.* DNA is the repository of genetic information and would seem like a logical candidate to be the first molecule of terrestrial life to follow one or more of the mother molecules of prebiotic times. However, this is not so. It is widely accepted that the carrier of the genetic code could not have been at the cradle of life, as it cannot replicate without the presence of certain proteins. To be specific, it needs enzymes unzipping its double helix structure before it can replicate. Some theories have RNA emerging first, others proteins, and yet others have both of them arising separately and developing a symbiotic relationship. In all instances DNA would have emerged later as a more reliable repository for genetic information.
2. *RNA first.* RNA is involved in the copying of the DNA genetic code and the assembly of amino acids into proteins on the basis of the copied DNA blueprint. As told by Casti, Eigen et al., in 1974, showed that from a broth containing the four RNA bases and an enzyme, RNA could self-assemble.[8] Thomas Cech (1989 Nobel Prize winner for Chemistry for the codiscovery of ribozymes, and now director of the Howard Hughes Medical Institute), when at the University of Colorado, demonstrated that RNA exhibits enzymatic properties and can catalyze the replication of other RNA molecules (i.e., ribozymes).[8] According to Matt Ridley, RNA was Greece to DNA's Rome, Homer to its Virgil.[12]
3. *Proteins first.* Proteins are the basis of a wide variety of functions (e.g., catalysis of reactions and immunoreactions to combat disease) and structural elements (e.g., in skin and hair). There is also evidence for the protein-first model. Reza Ghadiri, for example, showed that in homogeneous solutions, certain small peptides can catalyze the formation of other peptides, form symbiotic relationships, and spontaneously develop error-correcting mechanisms that prevent mutant forms from replicating.[8] In 1997, Lee et al.[13] published results in *Nature* on a helical peptide that is able to replicate itself, so the debate of what came first, RNA or protein, remains unsettled.
4. There also are serious problems with the pre-DNA models. The RNA-first model, for example, is thought to be too prone to errors. Coding errors are induced by mutations, and they are quite frequent in RNA, as there is only one strand of the molecule, and there would have been no "proofreading" enzymes available. As we will see below, DNA features two strands and special enzymes that repair coding errors before the code is translated into proteins. The RNA error rate explains why nature eventually switched to the more secure DNA molecule in order to protect its software. In the case of the proteins-first model, it is not easily understood how proteins could have passed on their heritage. An example of a symbiotic relationship theory is the one postulated by

Table 2.2 Common Terms Used in Biogenesis and Genetics

Term	Definition
Allele	Slight variant of the same gene in a given species.
Archaea	Archaeans are single-celled creatures that join bacteria to make up a category of life called the prokaryotes (pronounced "pro-carry-oats").
Chromosomes	A complex structured set of DNA with proteins inside the cell nucleus. It stores genetic information.
DNA	Biological chain molecule built from four kinds of building blocks, the nucleotides. DNA, normally found as a double helix spiral of two complementary strands, is the carrier of genetic information in all cellular life forms and many viruses.
Eukaryotes	Creatures whose cells have a nucleus, which contains the genetic information (pronounced "ew-carry-oats").
Gene	A segment of DNA containing the necessary information to produce a particular protein.
Genome	All of the DNA possessed by an individual member of a species.
Genotype	The particular collection of various possible gene alleles present in the genome of a given individual.
Germ cells	The cells of the body involved in reproduction; i.e., the sperm of the male and the eggs of the female are formed from germ cells.
Heterozygous	If there is more than one allele of a given gene within the population, an individual might have two different alleles for a given gene, in which case that individual is said to be heterozygous for that gene.
Homologous chromosomes	Chromosomes containing the same gene sequence.
Homozygous	Having the same alleles at a particular gene locus on homologous chromosomes.
Mutation	Modification of a DNA sequence extending from a single base change to the addition or elimination of an entire DNA fragment.
Phenotype	The physically discernible features resulting from a particular genotype.
Polymorphic genes	Genes that have allelic forms.
Prokaryotes	Prokaryotes' genetic material, or DNA, is not enclosed in a central cellular compartment called the nucleus. Bacteria and archaea are the only organisms of this class.
Protein	Molecules made up of long chains of amino acids. They build tissues, organs, etc., and carry out many critical functions in the body.
RNA	A nucleotide chain that differs from DNA in having the sugar ribose instead of deoxyribose and having the base uracil instead of thymine. RNA helps translate the instructions encoded by DNA in order to synthesize proteins.
Single-nucleotide polymorphisms	Differences (polymorphism) of individual bases within a genome from different individuals.
Somatic cells	All of the cells in the body except the germinating cells, such as those in the eggs and sperm.

Freeman Dyson. In this attractive model, a vesicle or cell (the hardware) metabolizes proteins but has no genome (the software) until a primitive nucleic acid replicating molecule invades the vesicle and finds it advantageous to replicate not only itself but also all the helpful proteins it finds in the cell.[8]

Made from Clay?

A rival theory to the DNA-, RNA-, or proteins-first models has life originating from inorganic materials such as clays. (Life originating from clay! Where have we heard this before?) This concept was first advanced in 1982 by Cairns-Smith,[8,10] a chemist at the University of Glasgow, Scotland. In a test tube,

small clay plates (platelets), less fragile than nucleic acids, given the right feed stock, will replicate—that is, identical offspring are generated (see Figure 2.2a). In a natural, more challenging setting in which a small change might induce the platelets to change shape (evolution), they may even reproduce—that is, their offspring become somewhat different, and these offspring may in turn replicate (identical offspring) or reproduce (offspring changed in shape). Microcrystals of clay, typically flat platelets with two large plane surfaces exposed to the surrounding medium, consist of a silicate lattice with a regular array of ionic sites but an irregular distribution of metal ions (e.g., Mg^{2+} and Al^{3+}) occupying these sites. The precise arrangement of metal ions in the clay may be capable of catalyzing specific reactions of organic molecules on the clay surface. These organic molecules in turn can bestow certain benefits onto the clay crystals: perhaps a more advantageous platelet shape or cementing the crystal planes together better—anything that protects them against the environment and speeds up (catalyzes) their replication. During crystal growth, silicate and metal ions from the surrounding medium will tend to carry the same pattern of ionic charges as the clay layer below it. In this way, the metal ion patterns, the carriers of information, are replicated and selected by environmental pressures. Replication and reproduction of clay are illustrated in Figure 2.2a. In a clay such as kaolinite, many cross-sectional shapes of crystallites are possible. The electron micrograph of the structure of kaolinite in Figure 2.2b illustrates the basic idea. The distribution of metal ions on these platelets, in association with organic molecules from the surrounding medium, helps favor one platelet out-replicating another. Physical aspects of the crystal act as the phenotype (what an organism—in this case a crystal—looks like as a consequence of its genotype) and cause selection to choose among the genotypes (the genes that an organism possesses).

Cairns-Smith suggests that clays came first, enzymes second, cells third, and genes fourth. After natural clay *learned* to direct synthesis of organic molecules, clay and organics *learned* to make a cell membrane. These cells containing clay as the hereditary unit eventually *discovered* RNA, which "piggybacked" at first on clays as they fitted snuggly into clay lattices. Clays might then have found a way of fabricating these self-replicating organics and, consequently, they would have quickly out-replicated other clays. By piggybacking on clays, RNA could have existed for a long time before becoming self-replicating. Once the organic molecules on the clay started replicating faster and faster by absorbing organic material from their surroundings, "organic life" might have taken over from "inorganic life." The latter is called a genetic takeover. For Cairns-Smith's bibliography and work, see http://originoflife.net/cairns_smith/index.html. There was not much experimental evidence for the Cairns-Smith's model until 1996, when Ertem and Ferris discovered that long strands of RNA, given the right feedstock, form spontaneously on the surface of montmorillonite (a type of clay).[14]

Life's Origin beneath the Earth's Surface?

The Archaea (also Archaeabacteria) constitute the earliest and most basic life form: relatively simple single-celled organisms that are chemotrophic; i.e., they do not need light, as they make biomass via chemical energy rather than light. Today, these organisms are still often are found in "extremophiles" living in harsh, toxic environments. Archaea use

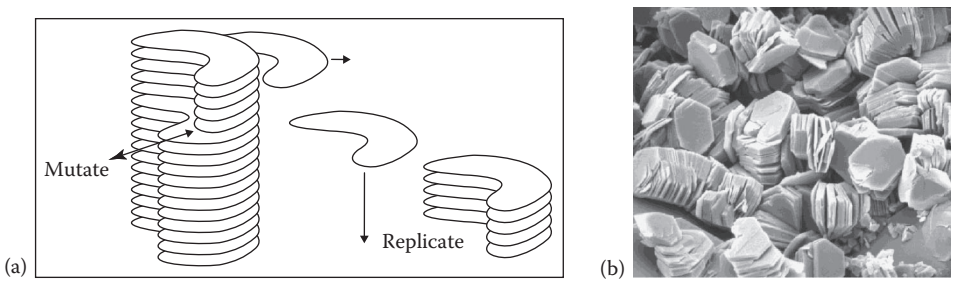


FIGURE 2.2 Replication and reproduction. (a) Replication and reproduction in a stack of clay platelets. (b) Electron micrographs of the structure of kaolinite clay.

a remarkable diversity of metabolic tricks and adaptations to survive under apparently impossible conditions. A lot of information has been gathered on these "superbugs" in the last twenty years; they have been found in oil wells, the Dead Sea, the Antarctic, near volcanic deep-sea vents, and deep down below the Earth's surface. From the Archaea two quite different groups developed, namely, the Eubacteria (bacteria) and the Eukarya (eukaryotes). Archaeans and Eubacteria (bacteria) make up a category of life called the prokaryotes. In prokaryotes, the genetic material, or DNA, is not enclosed in a nucleus, a central cellular compartment. The Eukarya include all organisms made up of cells with nuclei; that is, their DNA is walled off in a separate compartment of the cell (see also below under "Cells," Figure 2.20). It should be noted that extremophiles also include bacteria, and these hardy organisms can be grouped as acidophiles (acid loving), halophiles (salt loving), thermophiles (heat loving), hyperthermophiles (the present record is 121°C), psychrophiles (cold loving), alkaliphiles (thriving at high pH), barophiles (surviving under high-pressure levels, especially in deep sea vents), osmophiles (surviving in high-sugar environments), xerophiles (surviving in hot deserts where water is scarce), anaerobes (surviving in habitats lacking oxygen), microaerophiles (surviving under low-oxygen conditions), endoliths (dwelling in rocks and caves), and toxitolerants (able to withstand high levels of damaging agents, for example, living in water saturated with benzene or in the water core of a nuclear reactor).

Carl Woese, from the University of Illinois, in 1977, had a hard time convincing the world about the existence of Archaeans.[15] Woese found that Archaeans clustered together as a group well away from the usual bacteria and the eukaryotes. Because of their vast differences in genetic makeup, he proposed that life be divided into three domains: Eukaryota, Eubacteria, and Archaebacteria (Figure 2.3). Craig Venter, in 1996, completed the genome sequence of the first Archaean, *Methanococcus jannaschii*. The genome data of this hardy bacterium, thriving at a balmy 85°C and 200 atm of pressure,

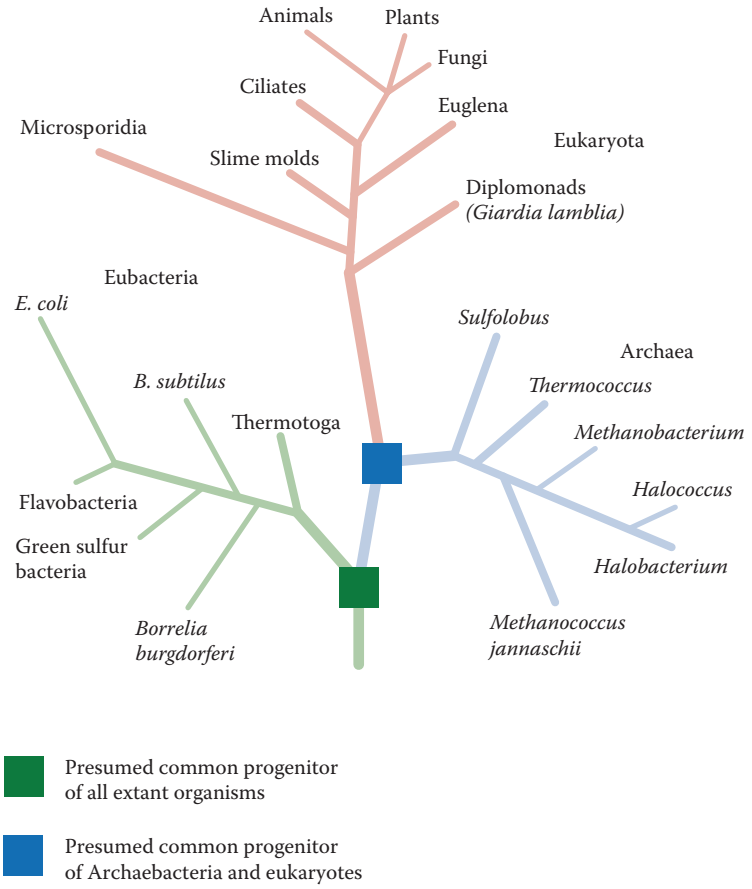


FIGURE 2.3 The tree of life.

includes the sequence of more than 1700 extremely heat-stable proteins never before sequenced.

The idea that chemicals beneath the Earth's surface produced these prebiotic replicators is attractive to many researchers. Early Earth, with high concentrations of $CO_2/H_2S/H_2$ etc., low oxygen levels, and high temperatures, was largely inhospitable to life as we know it today. Life forms that could evolve in such an environment needed to adapt to these extreme conditions. Many hyperthermophiles and Archaea are H_2 oxidizers. Thus, it is widely held now that extremophiles represent the earliest forms of life, with nonextreme forms evolving after cyanobacteria had accumulated enough O_2 in the atmosphere.

The recognition of life in those harsh environments on Earth makes the prospect of life on a harsh planet like Mars a bit more likely, and Archaea are favorites of the adherents of the Panspermia theory. A Web site with good information on extremophiles is http://www.astrobiology.com/extreme.html.

Extremophiles present a tremendous opportunity for more robust biotechnology processes and nanochemistry. Ordinary natural macromolecules cannot withstand extreme environments, making the prospect of widely used bottom-up, biomimetic devices less bright unless proteins from these superbugs are used to implement more robust nanochemistry. A well-known example of this approach is that of the hyperthermophile *Thermus aquaticus* used as the DNA polymerase in the very useful polymerase chain reaction (PCR) technique, which amplifies DNA samples by heat cycling the DNA sample together with the required chemistry cocktail (see "Polymerase Chain Reaction"). In Appendix 2A, we list other applications of extremophiles.

Complexity Theory

A groundswell of opinion has it that the difference between life and nonlife is only one of degree of organization, not of the type of components involved, and that life is how things behave, rather than of what they are made. Along this line, complexity scientists suggest that life is an "emergent" or "self-organizational" property that arises spontaneously when a large set of chemicals reaches a certain level of complexity (see also Chapter 3, "Self-Organization"). The background here is quite profound: humans are reasonably capable of calculating behavior of inorganic materials and the structures built from them, but we have had much less success with the more complex organic systems. This is where the science of complexity comes in—just like scaling and artificial life, complexity is a subdiscipline of biomathematics and represents a general attempt to elucidate fundamental biological principles on the basis of mathematics.

The Belgian Nobelist Ilya Prigogine (1917–2003) models complex systems that are far removed from thermodynamic equilibrium and describes their self-organizing* properties. These self-organizing properties have a tendency to reach critical "bifurcation" or "indecision points," where their behavior may leap into chaos or greater complexity and stability.[16] The latter scenario, Prigogine envisions, is how life might have started, that is, on the precipice between order and chaos.

An outstanding example of work in complexity is the theoretical work by Kauffman, from the Santa Fe Institute, who showed that an autocatalytic set of reactions leads to "emergent" and "self-organizational" behavior when a certain level of interconnectedness is reached between the reactions.[17] Kaufmann's Web site can be found at http://wn.com/Stuart_Kauffman. There is plenty of experimental evidence gathering for his theory in the laboratory. We described Ghadiri's protein work already (see "DNA, RNA, and Proteins: Who Was on First?"). In similar work, Rebek has demonstrated that synthetic chemicals can also reproduce, mutate, evolve, and combine with other reproducing chemicals.[18] Skeptics not only point to the simplifications and the arbitrariness of the rules of complexity theory but also point out that life is actually not an example of self-organization but one of genetically directed or specified organization. In other words, in life there is a software code directing the organization that emerges from complexity; in the complexity theory, on the other hand, there is no underlying software. In complexity theory, the physical boundaries of the system lead to the

* In Chapter 3, we clarify the distinction between self-assembly (near equilibrium) and self-organization (far removed from equilibrium) (see "Self-Assembly and Self-Organization").

emergence of organization. Despite the skeptics, this discipline is proving itself. We need to know how the system's software came about before we can understand how life emerged.

A Model from Information and Communication Theory

Yet another way of looking at the emergence of life comes from the theory of information and communication. Considering that DNA stores the information needed to construct and operate an organism, noise, a form of disorder, may then be regarded as the encroaching entropy. One then has to explain the creation of information (negative entropy) from noise, which again seems to violate the second law of thermodynamics until we realize that we are dealing with an open system (see above). Mutations in this model generate the noise, and it is the environment, via natural selection, that selects which information out of that noise ends up in the genetic message. The ultimate question in this model is, from where did the information content of the universe come?[10] The above considerations can be expressed in some simple mathematical formulations. There is a direct connection between information content (I) in a system and its entropy (S); an increase in entropy implies a decrease of information, and vice versa. Boltzmann linked entropy, S, to the computed probability of sorting objects into bins—a set of N into subsets of sizes n_i as follows:

$$S = -k\sum_i p_i \ln p_i \qquad (2.3)$$

where p_i is n_i/N, and k is the Boltzmann constant ($3.2983\ 10^{-24}$ calories/°C). On Boltzmann's tombstone, Equation 2.3 is written out as follows:

$$S = k \log W \qquad (2.4)$$

where W is the probability of a given state. One of the classic equations of information, derived by Claude Shannon in the 1940s, is formally similar to Equation 2.3.

$$I = \sum_i p_i \log_2 p_i \qquad (2.5)$$

Information, I, is a dimensionless entity, but its partner entity, entropy, has a dimension; comparing Equations 2.3 and 2.5, one notices that S and I have opposite signs and otherwise differ only in their scaling factors, and to convert from one to the other expression, one may write $S = -(k \ln 2)I$, or an entropy unit equals $-k\ln 2$ bits. Information is thus a concept equivalent to entropy, and "life" can be described in terms of one or the other.* Lowenstein promotes the idea that the first molecular mutants that came about were those that made the most efficient information loops, an information loop where the product promoted its own production.[19] As with digital computers, it is in the resetting of the memory register where the unavoidable thermodynamic price is paid. Each time a molecule has been produced on its template, the molecule needs to be reset for spinning out the next "clone."

For further reading on the origins of life in general, good starting points are http://www.seti.org/seti and http://www.panspermia.org/.

DNA and RNA: The Software of Life

In all known life forms on Earth, heredity, or genotype, is encoded by deoxyribonucleic acid (DNA). This molecule first appeared on Earth more than 3.5 billion years ago. Sections of DNA, called genes, constitute the chemical recipe expressed in the actual form of an organism or its phenotype. DNA—a master "read-only" memory—guides, records, and passes on every evolutionary development of all carbon-based life on Earth. Its building blocks are sugar (deoxyribose), phosphate, and four different nitrogen-containing bases, namely guanine (G), cytosine (C), thymine (T), and adenine (A). A unit of a base, sugar, and phosphate is a *nucleotide*, and since there are four different bases in DNA, there are also four different nucleotides in DNA. A model of a DNA molecule is shown in Figure 2.4b.

In 1953, James Watson and Francis Crick (see photo on chapter-opening page), building on experimental results by Rosalind Franklin (Figure 2.5), Maurice Wilkins, and many others, elucidated the

* William Dembski, a proponent of intelligent design, has claimed that natural causes are incapable of generating information. He calls this the *law of conservation of information*. From Equations 2.3, 2.4, and 2.5, information gain is equivalent to loss of entropy, and Dembski is confusing entropy and energy: entropy is not a conserved quantity like energy.

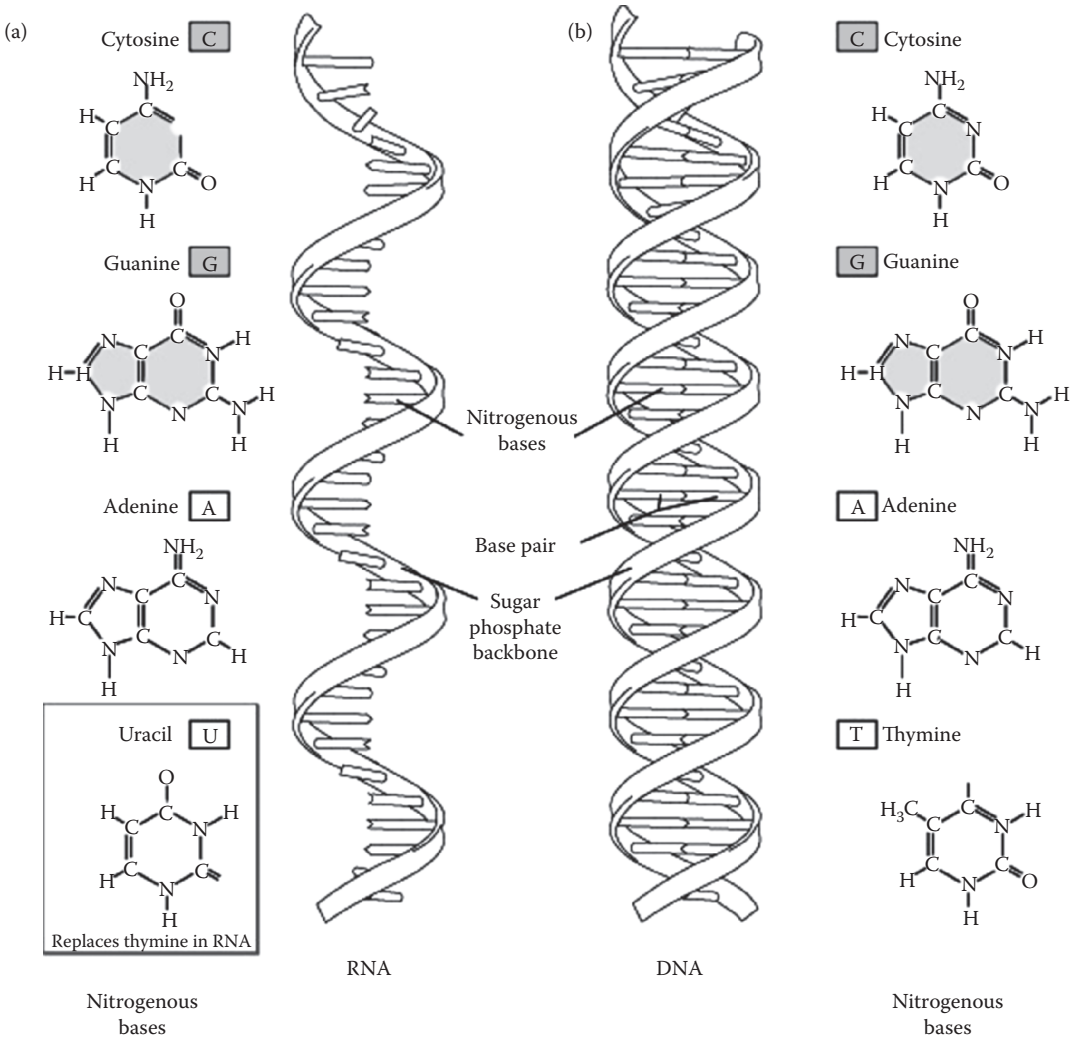


FIGURE 2.4 (a) Single-stranded RNA. (b) Double-stranded DNA.

double helix motif of DNA. Within a few years, it became clear how DNA, with the help of some other more or less complex molecules, is able to store and pass on hereditary information. This insight, among others, earned a Nobel Prize for Watson, Crick, and Wilkins in 1962. A fascinating historical, scientific, and ethical perspective on the discovery of the human blueprint is that of Robert Shapiro, *The Human Blueprint: The Race to Unlock the Secrets of Our Genetic Script*.[20] While scientific, Shapiro's book manages an entertaining account of the personalities involved in this greatest of scientific discoveries. Watson's autobiography *Double Helix* reads like a detective story and has become a classic bestseller.[21] Contrasting it with Crick's autobiography, *What Mad Pursuit*,[22] reveals how differently these two Nobel laureates measured their own contributions and that of others leading to that same monumental discovery (Crick passed away in July 2004).

The double helix DNA molecule consists of a right-handed spiral coil of two linked strands and is about 20 Å in diameter. Watson, during a dinner talk, perhaps after a bit too much wine, could muster only, "It is beautiful."* A model of the molecule as shown in Figure 2.4 is a bit misleading, since DNA is in reality a very narrow but extremely long molecule. From an

* It has been noted that the height of one unit of the DNA helix is related to its width in a proportion equal to the *Golden Ratio*. Throughout history, the ratio for length to width of rectangles of 1.61803398874989484820 has been considered the most pleasing to the eye. The Greeks named this ratio the Golden Ratio. In the world of mathematics, the numeric value is called "phi," named for the Greek sculptor Phidias. Many artists who lived after Phidias have used this proportion. Leonardo Da Vinci called it the "divine proportion" and featured it in many of his paintings. Long before Phidias, the Great Pyramids were proportioned based on the Golden Ratio.

FIGURE 2.5 Rosalind Franklin.

engineering point of view, DNA is a highly asymmetric molecule, with a width on the atomic scale and a length on the macroscopic scale, that is, from millimeters to centimeters and up to meters for a human.

This asymmetry makes DNA quite susceptible to breakage by shearing forces. The winding staircase of the DNA molecule in Figure 2.4 has railings made up by alternating molecules of sugar and phosphate; the steps are made up by pairs of bases. The four bases guanine (G), cytosine (C), thymine (T), and adenine (A) form only two types of pairs. An A base always joins to a T base via two hydrogen bonds (in shorthand, A = T), and a G always joins a C via three hydrogen bonds (G ≡ C). Walking up the staircase of the double helix, staying close to one side of the steps, the climber might read at his or her feet a base sequence such as AGCTAT; coming back down, on the other side of the steps, the descending person will read TCGATA. DNA often operates in conjunction with another nucleic acid, namely, ribonucleic acid (RNA). RNA is involved in the copying of the DNA genetic code and the assembly of amino acids into proteins based on the copied DNA blueprint. RNA uses a base called uracil (U) in place of thymine (T) and also employs a slightly different sugar ring (ribose instead of deoxyribose) (Figure 2.4a). Apart from these two differences, the components of RNA are identical to those of DNA.

The joining of two single-stranded DNA (ssDNA) strands into one double-stranded DNA (dsDNA) is called *hybridization*. The Watson-Crick pairing or hybridizing of DNA bases is shown in Figure 2.6.

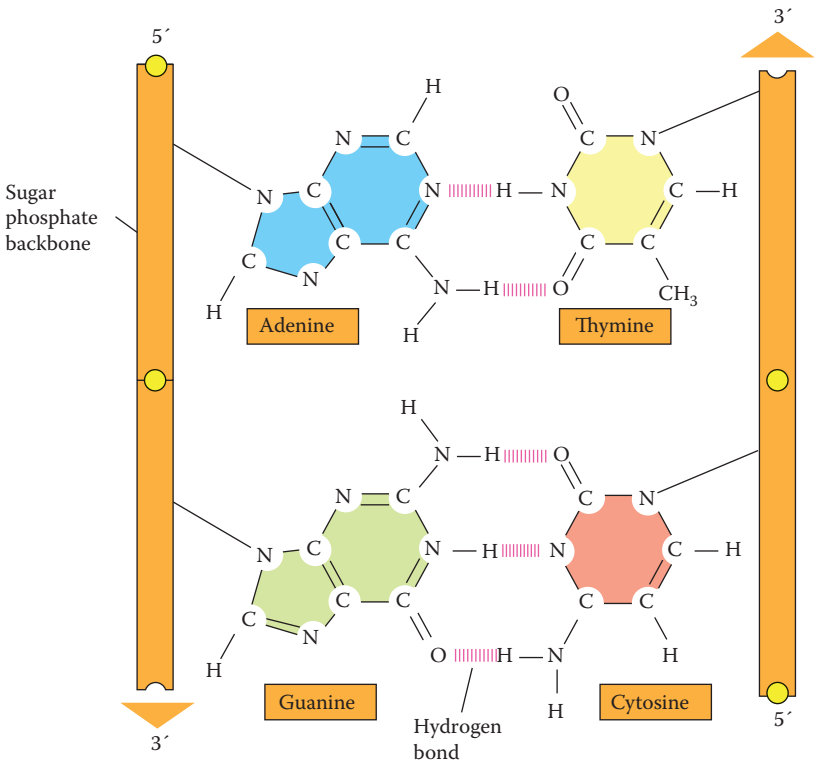


FIGURE 2.6 Base pairing in DNA.

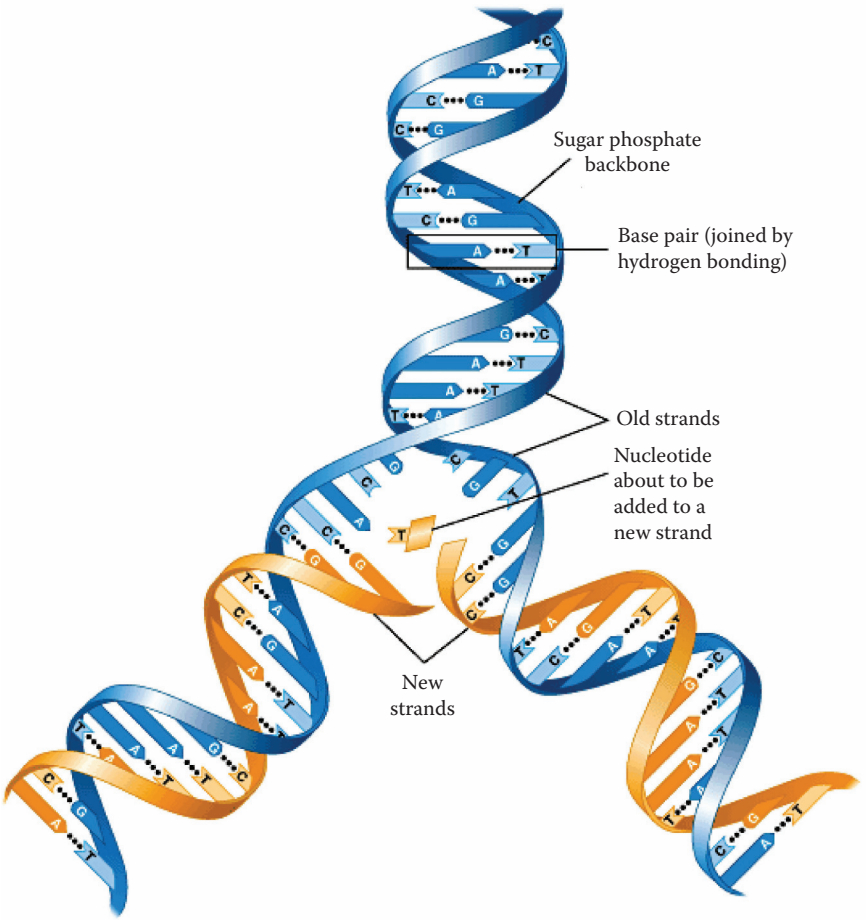


FIGURE 2.7 The DNA replication process.

To pass on information, a DNA molecule simply splits apart down the middle and then rebuilds matching parts to each half. The two strands in the ladder are complementary to one another, and, if the strands separate, each is fully capable of recreating the base sequences of the other with the help of a specialized enzyme. The enzyme, DNA polymerase, links up the appropriate nucleotides to make two daughter strands. The replication process is demonstrated in Figure 2.7. From this figure, we see that once the original double helix "unzips," the two separated strands act as templates (molecular molds), directing the production of new matching strands. If a double-stranded DNA is heated above its so-called "melting temperature," T_m, the two strands separate into two single strands. The melting temperature is a function of the ion concentration of the solution and the G-C content in the specific DNA sequence. When the temperature is reduced again, the two strands will eventually pair again by diffusion, and consequently "rehybridize" or "renature." The process just described is the one taken advantage of in PCR to replicate DNA (see above).

As mentioned earlier, one must distinguish between DNA replication as shown in Figure 2.7 and reproduction. Reproduction involves making modified copies of itself that can reproduce in return. If reproduction were flawless, evolution would never have occurred. DNA not only replicates and allows for reproduction, it also is the blueprint* for the synthesis of proteins. Further below, we introduce

* Not everyone sanctions the blueprint analogy. Ian Steward in *Life's Other Secret* correctly points out that genes are actually more like recipes in a cookbook rather than engineering blueprints; they tell us what ingredients to use, in what order, and in what quantities, but they do not provide a complete, accurate plan of the final result. Matt Ridley in *Genome* prefers the analogy of an immense book, a recipe of extravagant length; he also considers the term *blueprint* a poor analogy for a one-dimensional digital code. Then again, Robert Shapiro titled his book *The Human Blueprint: The Race to Unlock the Secrets of Our Genetic Script*.

some important characteristics of proteins, but we must first learn how a reading of the DNA blueprint results in their synthesis or production; i.e., we need to understand the mechanism of the genetic code.

Proteins: Hardware of Life and Nature's Robots*

Introduction

Virtually every process and product in an organism depends on proteins. In addition to their roles as enzymes (catalysts), messengers between cells, and defenders against infection, they are the principal constituents of cell membranes, tendons, muscles, skin, blood, bone, silk, hair, and other structural materials. The human body alone contains more than 100,000 distinct types of proteins, and all are synthesized in the same basic way. Just like nucleic acids, they are macromolecules, long polymeric chains. In the case of proteins, the chains are composed of just twenty different types of amino acids that are always joined in the simplest possible manner: linearly, in long chains (see Appendix 2B for a listing of their names and chemical formulas). There are many more amino acids, but nature only uses the twenty listed.

Biomolecules can process electrical and/or optical signals, but what proteins excel in is at recognizing and reacting to three-dimensional shapes. Proteins have twenty letters in their alphabet rather than the four of DNA, and they have the ability to adapt an endless number of three-dimensional (3D) shapes, making them much more versatile molecules than DNA. Information in biology, we shall see, is stored in the 3D shape of molecules.

Making Proteins

Amino acids are the basic components that make up all proteins. They are left-handed chiral molecules (except for the simplest, glycine) containing both a carboxylic acid group and an amine group. Chiral molecules are molecules that come in two forms, enantiomers, which are structurally identical except that one is the mirror image of the other. Like left- and right-hand gloves, these molecules cannot be superimposed (Figure 2.8). All chiral amino acids, forming proteins, in nature are "left-handed," and nucleotides, forming the genetic material, are "right-handed." It is one of the great unifying principles of biochemistry that the key molecules of life have the same hand in all organisms. The twenty common amino acids can be combined in an almost infinite number of ways to produce proteins. The reaction involved is the formation of a peptide bond (amide bond) in a so-called "condensation reaction" between the amine group of one amino acid and the carboxylic acid of another amino acid. In this reaction, a water molecule is released, as shown in Figure 2.9.

Each type of protein differs in its sequence and number of amino acids; therefore, it is the sequence of the chemically different side chains that makes each protein distinct. The two ends of a polypeptide chain are chemically different: the end carrying the free amino group (NH_3^+ or NH_2, depending on pH) is the amino, or N, terminus, and the one carrying the free carboxyl group (COO^- or COOH, depending on pH) is the carboxyl, or C, terminus. The amino acid sequence of a protein is always presented in the N to C direction, reading from left to right.

When only a few amino acids are linked, the product is called a *peptide*, the difference between a peptide and a protein being one of size. A protein with a molecular weight of 50,000 is also referred to as a protein of 50,000 daltons or 50 kilodaltons (kD). One dalton is equal to one atomic mass unit.

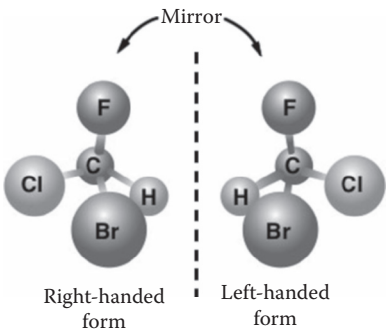


FIGURE 2.8 Right- and left-handed bromochlorofluoromethane. The left-handed form of a chiral molecule rotates a beam of polarized light to the left, while the right-handed form rotates it to the right. The example shown in this figure is the CHFClBr molecule.

* From the title of the book *Nature's Robots: A History of Proteins* by Charles Tanford and Jacqueline Reynolds.

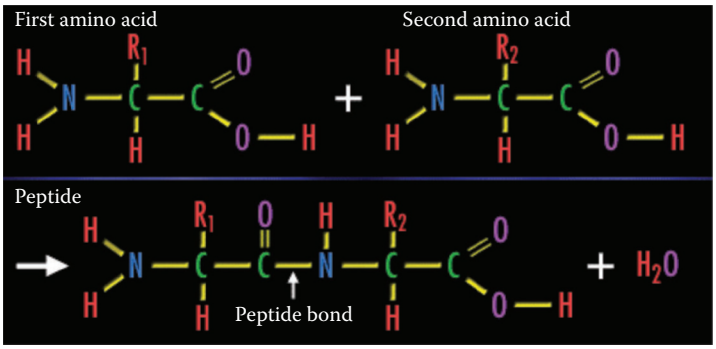


FIGURE 2.9 Water release in protein formation: proteins consist of a polypeptide backbone with attached side chains.

In Table 2.3, example proteins and their function are reviewed.

The Odds against the Emergence of Proteins

All models of the emergence of life eventually rely on the theory that a chance combination of chemicals leads to increasingly complex molecules and eventually to "life." Some simple organic molecules were already present in Earth's primeval atmosphere, as first demonstrated in 1953 in the now famous Miller-Urey experiments.[7] Miller, inspired by a suggestion of his advisor, Urey, reproduced a potential primeval atmosphere in laboratory glassware. When heating methane (carbon source), ammonia, and hydrogen with water, and passing an electrical discharge, to simulate lightning, through the mixture, he produced significant quantities of organic compounds, including some amino acids (e.g., glycine and alanine). He obtained this result after running the experiment for one week. Earth's early atmosphere was at first believed to have been reducing. Although Miller used hydrogen, ammonia, and methane, today it is believed that the atmosphere was more of a neutral mixture, consisting mainly of carbon dioxide and nitrogen and only small amounts of methane and ammonia. Since 1953, the Miller-Urey experiments have been reproduced with different simulated environments, such as the electric spark being replaced by a furnace, an ultraviolet lamp, shock waves, and energized chemicals. It turns out that amino acid synthesis is rather easy; as we saw before, they actually have been found even in meteorites.[3,7,10] The issue is not so much how to make individual amino acids but how to string them into the three-dimensional complicated proteins that life needs. The odds are stacked against making even the smallest protein this way. Take a small protein of 100 amino acids; with 20 different amino acids, the number of differently arranged proteins corresponds to 20 multiplied by itself 100 times (which corresponds to 1.26×10^{130}). Even the total number of stars in the observable universe, on the order of 10^{21}, is trivially small compared with the gigantic odds of stitching together a simple protein properly.[10] The latter is the major reason why scientists such as Crick and Davies came to believe that life on Earth was seeded from interstellar space, as there was not enough time to reach that degree of complexity on Earth in such a short amount of time.[23]

TABLE 2.3 Examples of Proteins

Structural	Functional
Collagen (found in bone and skin)	Hormones (control body functions)
Keratin (principal component of hair and nails)	Antibodies (fight infections)
Fibrin (helps blood clot)	Enzymes (catalyze chemical reactions)
Elastin (major part of ligaments)	Hemoglobin (carries oxygen and carbon dioxide in the blood)
	Myoglobin (transports oxygen in muscles)

Protein Structure

Proteins begin their life as linear chains of linked amino acids; however, most of them do not stay that way long, as is clear from Figure 2.10. The floppy protein chains quickly fold to minimize free energy. This minimization involves hydrophilic groups (groups that like water) orienting themselves toward the aqueous environment and hydrophobic groups (groups that dislike water) orienting themselves inward, that is, away from the aqueous or hydrophilic environment. Similarly, there are groups attracted to one another and groups that repel one another. The sequence of amino acids constitutes the primary structure of the protein. The secondary protein structure can follow several motifs—an α-helix (a right-handed spiral), a β-pleated sheet, or a ribbon. The tertiary protein structure describes how secondary structural motifs are packed together to give the molecule its three-dimensional shape. The tertiary structure of an enzyme molecule creates cavities and grooves on the molecule surface into or onto which only selected partner or ligand molecules fit. Those select ligand molecules, while bound to the enzyme, may in turn react as they present the right functional chemical groups to the immediate environment. The quaternary structure refers to the grouping of individual proteins into protein clusters that carry out a task collectively. The final assembly of any protein or protein group is all programmed by the original amino acid sequence in the linear primary structure. The seamless differentiation of proteins into different components results in one smooth and uniform manufacturing process. Every living body is thus a unique molecular machine in which function at the nanoscale dictates the cellular and whole organ behavior at the macroscale.[24] The various protein structural organizational levels are illustrated in Figure 2.11.

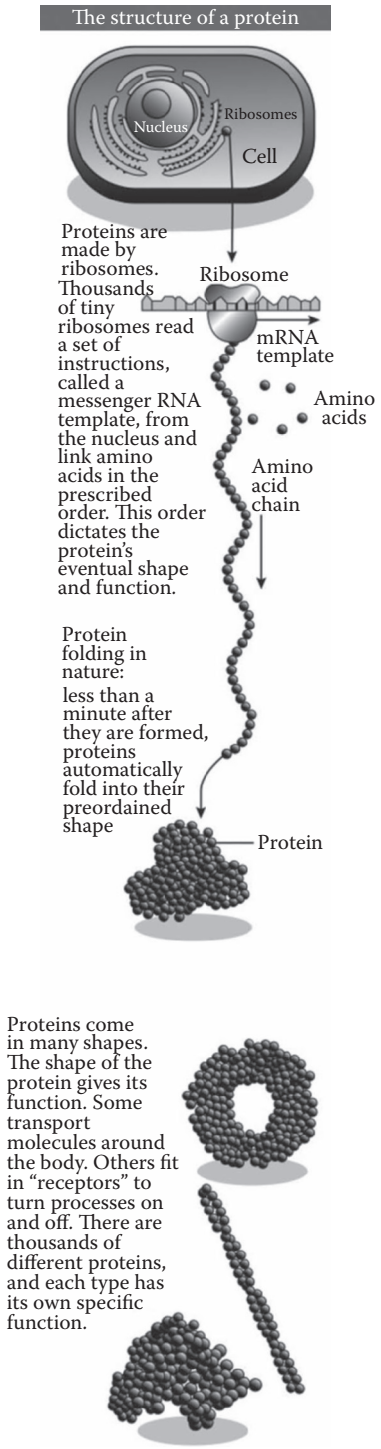


FIGURE 2.10 A protein emerges from a ribosome.

Today's computers can barely simulate a fraction of the fascinatingly complex protein-folding process. With that task in mind, IBM developed a computer called Blue Gene. Blue Gene is 500 times faster than other supercomputers. It was the first petaflop computer (10^{15}), and it is able to simulate interactions of certain proteins with their natural ligands and potential drug candidates and thereby potentially aid in the design of more efficient and potent drugs.

Protein Self-Repair and Recycling

Proteins and derived structures (e.g., tissues such as skin) are of use only in a small operational range of temperature and pH. Despite this narrow dynamic

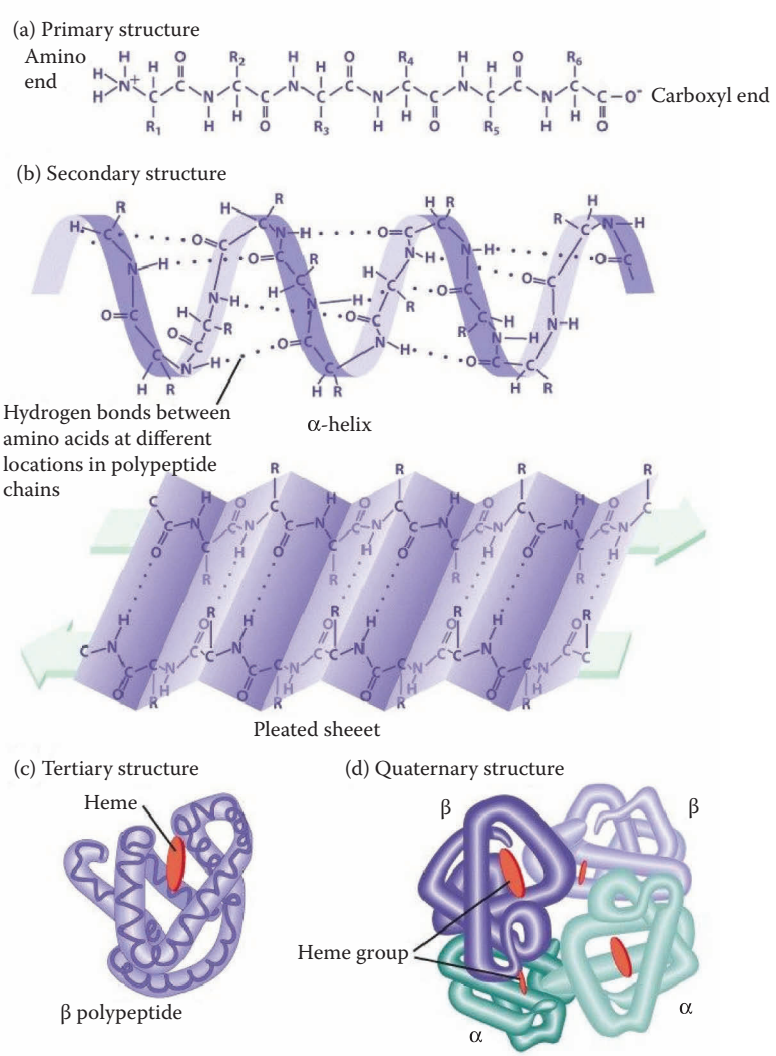


FIGURE 2.11 The various protein structural organizational levels.

range, the synthesis of natural organisms—starting with atoms and molecules, through intermediate structures such as fibers and grains, to components such as skin, bone, and teeth, has led to marvels of enduring engineering such as trees, birds, and humans. Humans usually use stronger (higher Young's modulus) building materials than proteins (say steel or aluminum), but nature has another major advantage, namely, self-repair. The external physical observable form, function, and behavior of living organisms (i.e., the phenotype) are not permanent constructs; they derive from the entire genetic information (i.e., the genotype). Some structures, such as skin and brain, are in a state of constant regeneration; proteins are continually synthesized de novo. Other structures, such as the eyes, are made up by proteins with some of the slowest turnover rates. This type of self-repair compensates for many of the deficiencies of life's construction materials. In human engineering, we have nothing of this kind yet, and together with self-assembly, self-repair is a must for the bottom-up approach in human engineering to succeed.

Cells of nearly all life forms possess systems for recycling proteins that have been damaged or are no longer needed. Ubiquitin tags proteins for degradation in a complex protein system, the proteasome, a molecular abattoir, in which only those proteins that are unfolded and marked with ubiquitin are destroyed. The proteasome also plays an important role in the immune system as it cuts foreign proteins down into chunks that the immune cells then present on their surfaces to signal the start of the attack by antibodies.[25]

The Genetic Code

In this section, we address the question of how nature, using DNA as a template, builds a plethora of different proteins from just twenty different amino acids. From our description of DNA, we learned that there are four different bases and four different nucleotides in DNA. All cells contain the same DNA, but different genes are expressed depending on the local need for an enzyme, a muscle cell, a neuron, or a bone. In this process, DNA is the repository for the construction blueprint, but it is not responsible for building the proteins. This hard work is done by another nucleic acid, namely, RNA. RNA uses a base called uracil (U) in place of thymine (T), and it also employs a slightly different sugar ring (ribose instead of deoxyribose). Apart from these two differences, the components of RNA are identical to those of DNA.

The details of the protein-building process are sketched in Figure 2.12. The process of protein manufacture (e.g., of insulin) may be compared to that of the manufacture of a car (such comparison was first made by Casti).[7] It all starts with copying the car's blueprint at headquarters, corresponding to transcription or copying of DNA in the cell nucleus. Next is a quality control step in which the copy of the blueprint of the car is inspected, a process that has its analog in the editing, or snipping out, of introns in protein manufacture. This step is followed by mailing of the copies to the various manufacturing plants—the ribosomes in the case of a cell. After a new car, or protein, is finished, it is packaged and shipped, and, finally, if the car or protein is no longer needed, it is disassembled and recycled.

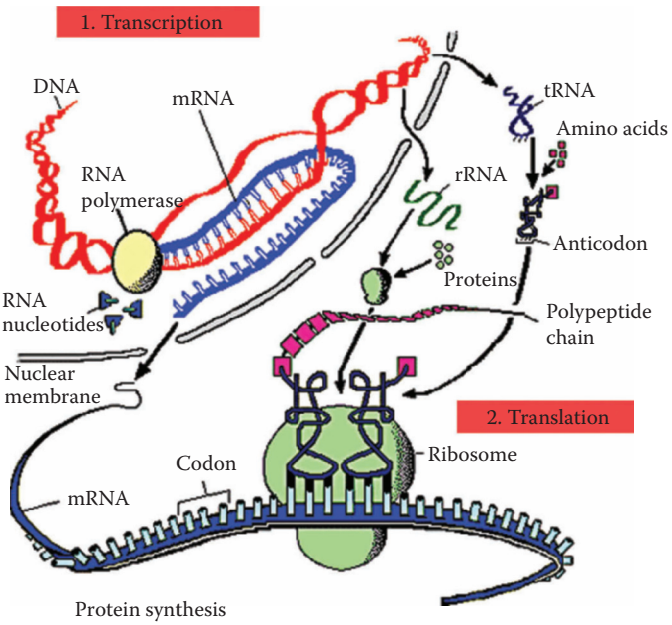


FIGURE 2.12 Protein manufacture: transcription and translation.

A bit more slowly, and in more detail, the story unfolds as follows. In protein manufacture, an enzyme first unzips a section of DNA, and the gene (DNA segment) that codes for the protein (in this case insulin) is freed for expression. All genes use the ATG codon to serve as the marker to start protein synthesis and TAA to stop it. Expression of a specific gene is switched on and off by proteins physically attaching themselves to promoter and enhancer sequences near the start of the gene text. In other words, the blueprint is unfolded for reading and copying one specific part of the original only (one car or protein is copied). Messenger RNA (mRNA) interprets the freed construction blueprint from the DNA inside the nucleus of the cell in a process called *transcription*. In transcription, an mRNA strand is synthesized by complementary base pairing between the nucleotide bases of the DNA and the mRNA. In Detroit lingo, this translates to the photocopying of the blueprint for a certain type of car. In eukaryotic cells, the mRNA now leaves the nucleus and acts as a template for protein synthesis in ribosomes. In our car analogy, copies of the car's blueprint are being distributed to different factories in Ohio and Indiana. The code to make proteins on the mRNA template is a triplet code—a sequence of three bases, or a *codon*, describes one amino acid.

If each individual base is coded for only one amino acid, then the four bases would only code for four amino acids. A pair of bases could code for sixteen different amino acids (4^2), and a triplet could handle sixty-four (4^3). It is the latter code, the one with the largest bandwidth available to call on the twenty amino acids, that nature has chosen to build proteins (see Appendix 2C for a list of the twenty amino acids). Actually, an amino acid may be encoded by between one and six codons, several different triplets may code for the same amino acid, and three triplets are assigned to the instruction for

a stop. In 1967, Har Gobind Khorana and Marshall Nirenberg were able to correlate each one of the twenty amino acids with the sixty-four possible codons. The resulting universal decoder chart is presented in Appendix 2C.

In many cases, stretches of DNA coding for polypeptide chains are not continuous but are interrupted by large stretches of "nonsense" sequences. Even within genes, there are stretches of nucleotides, called *introns*, that don't code for anything; Casti calls them *commercial breaks*.[8] Exons within the gene are the only segments where protein coding occurs (see Figure 2.13). In humans, there is only about one gene per 100,000 base pairs, and the mean gene extends over 27 kilobases, whereas the mean transcript is 1.3 kilobases.

The validity of the mRNA copy is checked by verifying the integrity of all base pairs (also called *editing*). A team of enzymes snips out the introns and reconnects the parts that contain useful information called *exons*, which are the expressed genetic sequence. This editing process reduces the error rate to about one in 1 billion base pair replications. Although most editing is done by enzymes, some RNA molecules can do it unassisted. These are called *ribozymes*, reflecting the

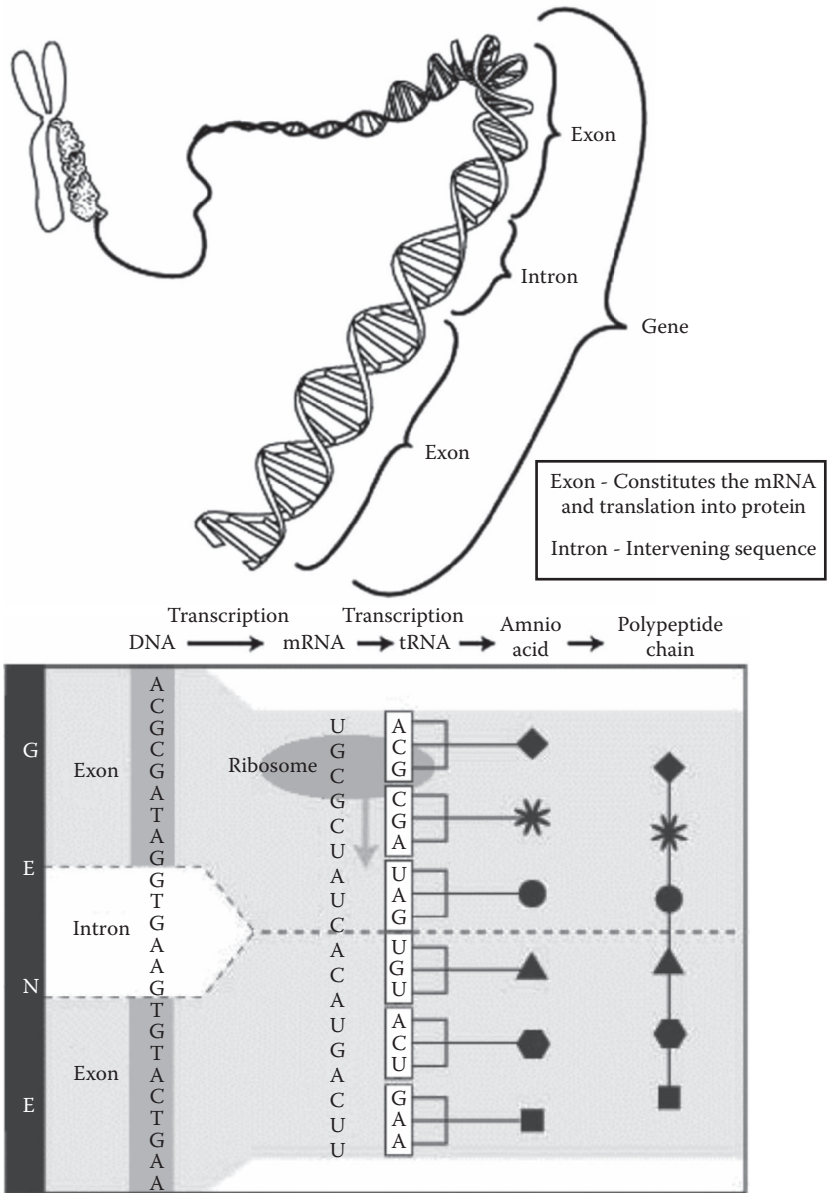


FIGURE 2.13 In the expanded gene, only the exons actually code for the protein corresponding to the gene.

fact that they show -enzyme-like tendencies. The nonsense DNA—perhaps up to 95% of the human genome—has been called *junk* DNA in the past but more recently has been upgraded to *selfish* DNA, as scientists recognized that it might play some role in heredity after all. Crick speculated that this selfish DNA might be involved in the rapid evolution of some of the complex genetic control mechanisms essential for higher organisms (lower organisms have none or much less of this selfish DNA). Crick believed selfish DNA might be there for its own sake rather than that of the organism itself. The same claim is made by Richard Dawkins[26] in his book *The Selfish Gene*.*

We continue now our story about the genetic code and car manufacture and study the process that follows transcription and editing—that is, translation of the blueprint into proteins, or how to go from a blueprint copy to a car.

Living organisms feature armies of three-dimensional protein construction sites called *ribosomes* (see also Figure 2.12) in which any conceivable protein can be manufactured. Small RNA molecules called *transport RNA* (tRNA) transport the proper amino acids to the ribosomes, which are cell organelles located in the cytosol outside the nucleus, to build new peptide chains on the mRNA template in the so-called *translation* process. Each of the twenty amino acids has its own tRNAs equipped with a codon, complementary to the triplet of bases that encodes the tRNA's amino acid on the mRNA template. In the car analogy, individual car components are being shipped to the car assembly plant. The ribosomes bind together the amino acids, docked in place by the tRNAs, in a growing protein chain at a rate of five amino acids per second. Many ribosomes can be lined up on one mRNA strand so that a full copy of a new protein—for example, insulin—can be delivered every few seconds. The insulin protein just fabricated in the ribosome contains a short sequence specifying the part of the cell to which it should be sent. This is the equivalent of a bar code on a new car. When this signal sequence (the bar code) has emerged from the ribosome, the cell's transport mechanism is activated, the bar code is read by a signal recognition particle, and in the case of an emerging insulin it is, together with its ribosome, directed toward the membrane of the cell's shipping dock, the endoplasmic reticulum (ER). The ribosome attaches to the ER membrane, and the protein is threaded through a pore in the ER. Once inside the ER, the signal sequence or bar code is removed, the protein—assisted by molecular chaperones—folds into its native shape and is packaged into a small sphere (vesicle) that buds out of the ER membrane. The vesicle then transports the insulin to the Golgi apparatus. The Golgi apparatus, similar in appearance to and perhaps continuous with the ER, is a region of smooth, stacked membranous sacs. Cell biologists have shown that the Golgi apparatus modifies proteins, which it receives from the ER in vesicles, and prepares them, in a series of steps, for excretion to the exterior. Just as a letter is taken to the post office to be mailed to another location, proteins are transported to the Golgi apparatus for shipment to other locations in the body. Within the Golgi apparatus, the proteins and lipids are labeled with sequences of molecules (an address), which tells the body where the product should be delivered.

The French biochemists Francois Jacob and Jacques Monod showed in the 1960s that genes regulate one another, switching each other on and off via the proteins they encode. For example, some genes that encode proteins used in the cell (called *structural genes*) are partnered with regulatory genes that encode repressor genes. With the regulatory gene switched on, the repressor protein is synthesized and binds to the structural gene, preventing it from being *expressed* (transcribed and translated into its protein). Jacob and Monod called these regulated stretches of DNA *operons*.

The fact that information flows from DNA to RNA to protein is called the *central dogma* of molecular biology. This "dogma" also implied that there was only one protein per gene. A first serious attack on the first part of this central dogma came in 1970 when David Baltimore and Howard Temin discovered an enzyme that does just the opposite: it transforms the RNA genome of a group of RNA viruses, now called *retroviruses*, into a DNA copy that can be

* *The Selfish Gene* is an elegant account of neo-Darwinism in which Dawkins argues the view that DNA rules.

inserted into a genome. The family of retroviruses includes HIV. The enzyme in question is a reverse transcriptase. In comparison with DNA polymerase that replicates DNA chains, reverse transcriptase is very error-prone. The latter is what makes it so difficult to make an HIV vaccine. Baltimore and Temin received a Nobel Prize in 1975 for this discovery.

Gene Copying Errors: The Error Catastrophe

A higher organism has about thirty thousand genes capable of storing about one hundred million bits of information, each of which may be subject to copying errors. Too many copying errors can cause the reproduction machine itself to come to a halt. The German biochemist Eigen called this the *error catastrophe*.[27] Eigen, who believed in an RNA-first world, introduced a simple mathematical formulation for characterizing this error catastrophe. He assumed that a self-replicating system is specified by N bits of information and that each time a single bit is replicated from parent to offspring the probability of making and error is given by ε. He also accepted that natural selection penalizes copying errors by a selection factor S. In other words, an error-free replication has a selective advantage of S over a system with one error, and so on. Then according to Eigen the criterion for survival of a replicating organism is:

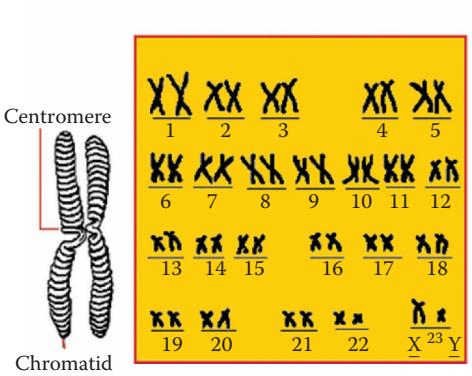
$$N\varepsilon < \log S \qquad (2.6)$$

If the above condition does not hold, the species will die out, but with the inequality satisfied, the selective advantage of the error-free system is great enough to maintain a population alive with few errors. To understand the deeper meaning of Equation 2.6, we may look back to the information theory of life (see above) for some further insights. The left side of the inequality represents the information loss through copying errors in each generation. The right is the number of bits supplied by the selective action of the environment. If the information supplied is less than the information lost in each generation, a progressive degeneration is inevitable. The selective advantage of an error-free system cannot be too large, so that log S is probably not much greater than 1 (life always survives on the edge of chaos). The greater the number of genes an organism possesses, the lower the error rate must be to avoid error catastrophe. Human beings accumulate about 100 mutations per generation, with N of the order of 10^8, to satisfy Equation 2.6, ε must be of the order of 10^{-8}. In other words to avoid error catastrophe, the copying error in human DNA should be less than one in a hundred million per replication. In human cells, an editing process weeds out errors, and the remaining error rate is actually one in a billion. Bacteria with considerably fewer genes than humans can get away with many more errors (they can tolerate up to one in a million), and a virus can survive an even greater number of errors. The latter is the reason why drug-resistant mutations emerge with such rapidity, always keeping researchers one drug behind. This will continue to challenge chemists and the pharmaceutical industry, keeping the latter happy and rich for a long time to come.

Genes and Chromosomes

A gene, as we learned above, is a segment of DNA with a unique sequence of nucleotides that encodes information—the genetic code—for assembling amino acids into a particular protein. Genes are located on chromosomes and humans have twenty-three chromosome pairs (Figure 2.14). Genes are isolated from chromosomes through purification of RNA molecules. Since RNA is not very stable, this purified material is first converted back into DNA known as *complementary DNA* (cDNA). To store and make copies of genes, the cDNA is stitched into a small circular bacterial chromosome. Whenever more of the gene is needed, the bacterial colony containing the gene of interest is grown, and the required amount of cDNA is isolated.

The biggest surprise of the rough analysis of the first sequencing of the human genome (see below) was the low number of genes it contains. For years scientists had been predicting that human DNA would contain somewhere between 100,000 and 140,000 genes. It turns out we may have as few as 26,000—a genome about the size of a rat or a corn plant, with roughly a third more genes than the fruit fly. We are coming to the somewhat uncomfortable conclusion that there is no relation between the complexity of an organism and the number of chromosomes or genes or the length of the DNA strand. *Lysandra atlantica* (a butterfly) has 250 chromosomes



Organism	Genome Size (Bases)	Estimated Genes
Human (*Homo sapiens*)	3 billion	30,000
Laboratory mouse (*M. musculus*)	2.6 billion	30,000
Mustard weed (*A. thaliana*)	100 million	25,000
Roundworm (*C. elegans*)	97 million	19,000
Fruit fly (*D. melanogaster*)	137 million	13,000
Yeast (*S. cerevisiae*)	12.1 million	6,000
Bacterium (*E. coli*)	4.6 million	3,200
Human immunodeficiency virus (HIV)	9,700	9

FIGURE 2.14 Human chromosomes and genome size (bases) and estimated number of genes for some different organisms.

versus our meager 46. Nor is there a correlation with the number of genes; the fruit fly has 13,700, whereas *Caenorhabditis elegans* (a 1-mm nematode worm) has 20,500. Despite the fact that the eye of the fruit fly is composed of more cells than are found in the entire *Caenorhabditis* worm (profoundly more complex in structure and behavior than the microscopic roundworm), the fruit fly has 5000 fewer genes! Finally, there is also no correlation with the length of the DNA, with, for example, the *Amoeba dubia* at 670,000,000,000 bases and *Homo sapiens* at 3,000,000,000. Obviously, the complexity of organisms is not reflected in the complexity of its genes. Some people are referring to this dilemma, the lack of correlation between the genome and the complexity of the organism, as the cosmic joke. The cosmic joke violates a second tenet of the central dogma, i.e., one gene to one protein. Ron Evans, Michael Rosenfeld and others clearly demonstrated that the same RNA gene transcript can be spliced in many ways to encode multiple proteins.[28] This discovery shed much-needed light into the low number of genes–large number of proteins paradox.

Only 40% of a chromosome consists of DNA. The rest is made up of packaging protein materials. Figure 2.15 illustrates an exploded view of a chromosome as well as some other important cell components. The total human genome, that is, the full set of chromosomes, comprises about 3 billion DNA base pairs. One strand of human DNA stretched out would measure up to 5 cm. Since there are forty-six chromosomes in a human cell, this means we have about 230 cm, or 6 feet, of DNA for each cell. Within the chromosome, the DNA is obviously very well coiled and packed. This remarkable feat of packaging is described in more detail in Chapter 4 on packaging (see Figure 4.1). In this packed state, the DNA is so tightly bundled that it is entirely inaccessible to gene-activating enzymes. Only as stretches of chromosome move out of this condensed state are genes exposed and available for expression. Inside the nucleus of each cell there are two complete sets of the human genome. Egg and sperm cells contain only one copy, and red blood cells do not have any DNA.

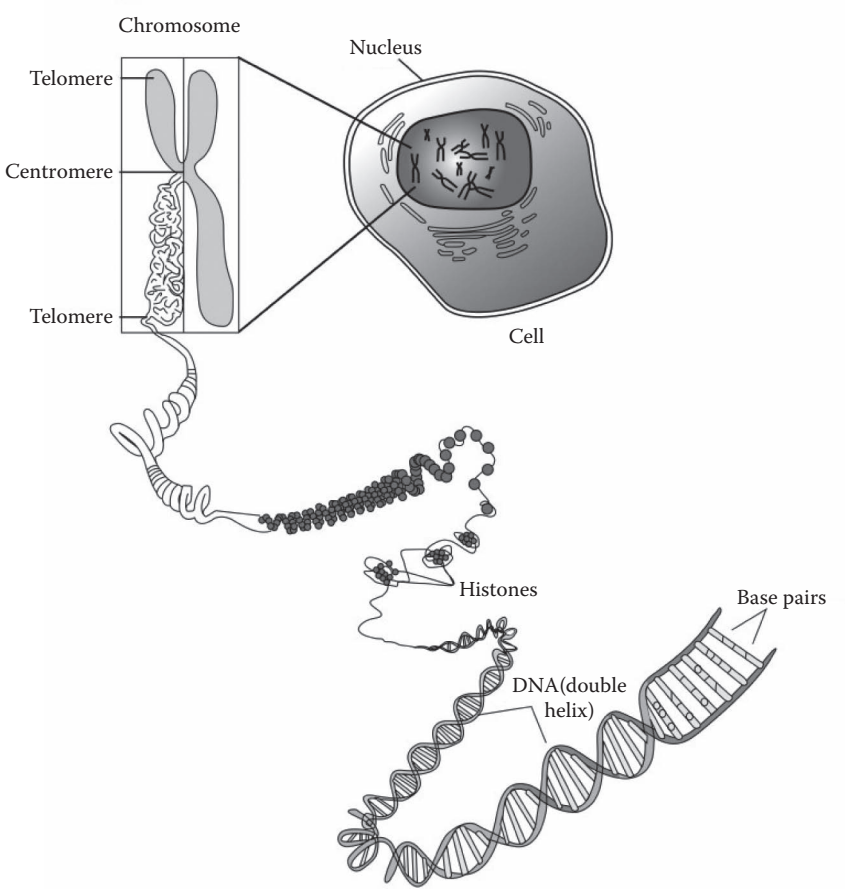


FIGURE 2.15 From chromosome in the nucleus to DNA.

Evolution

The driving force for the sequence of events that led to the plethora of life forms we know today comes from the ever-changing challenges presented by the environment to living organisms in a process called *evolution* or *natural selection*. Evolution theory tells us that over long periods of time, life forms change. Perfect replication of DNA would only have led to a planet with very little diversity among its life forms. Errors during DNA replication do occur, and when errors occur, they can manifest themselves as mutations in the organisms that inherit them. Mutations not only arise from copying errors but may also be caused by physical damage to the DNA in the genes, such as that caused by radiation and certain chemicals. If an alteration of a gene results in an observable change in some aspect of an individual's phenotype, and if this change is passed on to his or her offspring, then we say that a mutation has occurred or a new allele of a gene was created. Only mutations in germ cells are passed on to the offspring: mutations in somatic cells (all the cells of an organism except for the germ cells) do of course occur, but they are not passed on. An allele is defined as one of the two alternative states of a gene that occupy the same locus on homologous chromosomes (chromosomes containing the same gene sequence) and that are responsible for alternative traits. Not all genes have alleles; many genes within a given species are invariant among individuals. If there is more than one allele of a given gene within a population, an individual might inherit two different alleles for a given gene (one from the father and one from the mother), in which case that individual is said to be heterozygous for that gene. Some alleles are dominant over others, and with a heterozygote the dominant allele will be the one that is expressed in the phenotype. With two identical alleles, the individual is called monozygotic. Random errors in the DNA coding are mostly disadvantageous, sometimes neutral, and very rarely beneficial, conferring

an advantage on the mutant. Over long periods of time, favorable traits will be transmitted more efficiently to the offspring, and unfavorable ones will occur with decreasing frequency, until they are eliminated. For any two unrelated individuals, there is roughly one alteration every 1000 bases, or 3 million sequence variations altogether. These sites of sequence variations are called single-nucleotide polymorphisms (SNPs, pronounced *snips*). Those "snips" are critically important in the development of new molecular diagnostics (see applications in Chapter 10).

The concepts of evolution and inheritance were introduced in the second half of the nineteenth century with Charles Darwin's *On the Origin of Species* (1859). Gregor Mendel published in *Transactions of the Brunn Natural History Society* on heredity of peas in 1865, but his work was not widely known until 1900, when it was rediscovered. Darwin had predicted a continuous variation within a species with gradual change over long periods of time. The perpetuation of any life form requires a fine balance between fidelity of replication and variation from generation to generation. A slight imbalance, and extermination certainly follows. Among Mendel's pea plants, a more discrete variation manifested itself, with the plants exhibiting either white or purple flowers with no continuous variation in between. In the early twentieth century, evidence pointed toward gene mutations having a big step effect rather than a continuous effect. Only later did it become apparent that although the units of heredity—genes—are discrete, many mutations do have small-scale effects, and that combinations of genes can produce continuous variations, such as height in humans. The mathematical foundation of Darwin's theory was on solid footing by the 1930s. Mendel's very discrete heredity results were found to be coincidental, because the traits he studied were controlled by single genes only. If this had not been the case, he would not have obtained such clear-cut results.

There is some controversy as to whether natural selection explains all features of evolution. One extreme position is that of Dawkins, who, in his books *The Selfish Gene*[26] and *The Blind Watchmaker,*[29] contends that everything is shaped by natural selection alone. The opposite extreme, much more controversial, is represented by, for example, Micheal J. Behe with his intelligent design who claims that Darwin's mechanism—natural selection working on variations—can explain small changes but does not explain how complex molecular systems, such as vision, might have been gradually produced (for an introduction visit http://www.talkorigins.org/faqs/behe/review.html). Intelligent design is a criticism of Darwinism, not a theory, as it does not logically follow from experimental evidence. In between, there are scientists such as the late Stephen Jay Gould, who believed that there must be other factors involved and that selection alone is not capable of explaining all patterning forces in the history of life (http://www.stephenjaygould.org/library/gould_structure.html). This view supports complexity theory, which sees evolution as a complex system, not to be studied by its individual parts alone. This implies that self-organizational properties typical of complex systems would go hand in hand with selection. Artificial life simulations also have the potential to solve some of the major problems that evolutionary biologists have struggled with for decades. The first one has to do with the concept of predeterminism in evolution. Evolution is not supposed to have purpose or a predetermined goal in mind—mutations are random, which makes it difficult to explain how life organized itself into increasingly complicated organisms. However, computer simulations of artificial life do suggest a tendency for systems to organize themselves into increasingly complex forms for purely mathematical reasons. A second problem is that of mass extinctions, which computer simulations suggest are the norm, not the exception.[6]

Transposomes

It is estimated that less than 5% of the DNA in a human contains gene sequences that code for proteins. Noncoding DNA is assumed to consist of structural elements and repetitive sequences such as transposons. Transposons, also called *jumping genes*, are stretches of DNA that jump from site to site within one genome or "catch a ride" with a virus and move from species to species in so-called horizontal transmission.

Barbara McClintock won the 1983 Nobel Prize for her work on transposons in Indian corn; she postulated transposons as far back as the 1940s. It is now widely accepted that genes operate as parts of networks almost like those formed by neurons in the brain, and individual genes may either inhibit the action of another gene or cause it to be more likely to express itself (translate into proteins). Because of this additional complexity, it seems increasingly likely that a traditional reductionist approach will not be able to solve all of the issues involved in the expression of the 30,000 or so human genes. Answers may come from the science of complex systems instead, and, as a result, nonlinearity and complexity mathematics may become the most important contributors to the understanding of life and evolution. Along this line of thinking, Roy Britten at Caltech and many others believe that evolution cannot be explained by single-point mutations only.[30] Britten's position is that transposons are a much larger source of natural variation. To make his argument, he has focused on the Alu transposons that are unique to primates. Alus make up more than 5% of our DNA and could conceivably be responsible for the diversity of primates. Alus, for some unknown reason, seem to have started spreading widely around 30 to 50 million years ago. Wanda Reynolds discovered that Alus have the power to alter which sets of genes are expressed, by which hormone, and when they are triggered, thereby elevating or reducing gene expression in certain tissue so that the evolution of a species can gradually be changed (see also http://www.ncbi.nlm.nih.gov/pmc/articles/PMC38434/). This process might have been an important factor in creating diversity in primates. Despite the fact that the genetic distance separating us from pygmy chimps is only 1.6%, it is possible that not all our differences are the result of this (the chimp-human split occurred not much more than 10 million years ago, possibly even less than 5 million). The proximity of Alus might have regulated which sets of genes were triggered, enabling *Homo sapiens* to appear around 40,000 years ago. Transposons can be regarded as an internal environmental challenge to the evolution of life forms rather than the more familiar external environmental factors. Transposons, by posing a continuing challenge to increasingly complex genomes, have helped evolution along and might even be the reason why more complicated organisms such as humans could emerge at all. The reason for the dense packing of chromosomes in more complex living systems is probably to keep transposons better hidden within them. As a consequence, only 0.2% of all spontaneous human mutations are caused by transposons, although they make up a large portion of our genome. By contrast, fruit fly transposons constitute only 10–20% of the genome, yet they are responsible for as much as 85% of spontaneous fruit fly mutations. The chromosomes of the fruit fly hide the transposons much less than our chromosomes do.

In 1998, Schatz and coworkers at Yale University identified yet another major role that transposons might have played in evolution, namely, the development of a powerful immune system in vertebrates.[31] A powerful immune system is created by mixing gene fragments to create a vast array of antibody genes in millions of B cells.* Schatz's group has demonstrated that the mechanism for the creation of these antibody cocktails involves the RAG1 and RAG2 genes, which produce the proteins used to "cut and paste" the gene fragments into new combinations. RAGs have lost their ability to reinsert themselves at random in our genome, and as a result, they are not only harmless but actually beneficial. Transposons also might be dangerous, as their presence near or on a functional gene may damage or destroy it; they have been associated with diseases such as leukemia, hemophilia, and breast cancer.

Human Genome Project

The aim of the $3 billion Human Genome Project, started in the mid-1980s, was to decode the details of the software running the genetic code machinery depicted in Figure 2.12 by 2005.[32] Scientists working on the project located, mapped, and identified the ~30,000 individual genes carried on our twenty-three pairs of chromosomes by 2001. If the

* B cells develop from stem cells in the bone marrow. Each B cell is programmed to make one specific antibody. For example, one B cell will make an antibody to block a virus that causes the common cold, while another produces an antibody to attack a bacterium that causes pneumonia.

human genome is compared to a book, the twenty-three chromosomes may be considered the chapters, with each chapter containing many stories, or genes. Each story contains many paragraphs, called *exons*, interrupted by advertisements called *introns*. The words are *codons*, and each word is written in letters, or *bases*.[12] One of the project's approaches involved cloning fragments of DNA, inserting them into *E. coli* bacteria, and looking for shared sequences of DNA that could be identified chemically. The Human Genome Project stressed free access, including via the Internet, to all the information resulting from this multinational project (http://genomics.energy.gov). On the other hand, Celera, a competing private company, had already applied for patents on 6500 different gene sequences. In elucidating the genome, Celera used an approach in which the genetic material was shredded into tiny pieces and, after analysis, a supercomputer tried to put the pieces back together. In September 1999, Celera succeeded in mapping the genome of the fruit fly (*Drosophila melanogaster*), which contains "only" 180 million base pairs. The Human Genome Project announced the elucidation of the map of human chromosome 22, the second smallest, on December 2, 1999.[33] Although this chromosome holds only 1.6–1.8% of the human genetic code, it is believed to contain genes that play a role in at least thirty-five diseases, including congenital heart disease, mental retardation, leukemia, and schizophrenia. By June 26, 2000, Bill Clinton, president of the United States and Tony Blair, prime minister of the United Kingdom, simultaneously announced that the rough draft of the Human Genome was complete. We can now only hope that humankind has it in its genes to exploit the results for the benefit of humanity rather than for making a very few very rich individuals yet richer at the cost of their own country and especially to the detriment of third world nations.

Identification of the DNA sequence is only the beginning; the real payoff is still to come; molecular biologists worldwide that wrote down the 3.5-billion-bit genetic code are now trying to link specific genes and the proteins for which they code. It is becoming apparent that the expression of genes and the folding of proteins might better be solved by complexity theory than with a reductionist approach because of the extreme interconnectedness of the subcomponents.

Death and Aging

With the aging of the population in the industrialized countries, research funding for the study of death and aging is rapidly increasing. Some things about aging are known, but more needs to be learned, and plenty of controversy surrounds the topic. Here are some generally accepted facts. The simpler the life form, the easier it is to prolong its life or to regenerate parts of itself. Bacteria, for instance, continue to grow and produce progeny as long as they have food and a nontoxic environment. They do not die because of aging. Aging becomes increasingly important as a species acquires more sophisticated, specialized cells and organs. In these more complex life forms, there are two types of cells, namely germ cells and somatic cells. Germ cells are potentially immortal (like bacteria, they can reproduce the whole living entity forever); somatic cells are there for the greater efficiency of particular functions but cannot reproduce the whole organism. If a worm is cut in half, the tail, which contains germ cells, can reproduce a whole worm; the head, which only contains somatic cells, dies. The higher the ratio of somatic cells to germ cells, the more prone an organism is to aging. Mutating the genes that control how worms respond to hormones resembling insulin enable worms to live two to five times longer.[34,35]

Chromosomes are capped and protected by special DNA pieces called *telomeres*, in the same way that shoelaces are prevented from becoming frayed by aglets. These telomeres in mammals consist of thousands of TTAGGG repeats and act somewhat as an internal clock for the aging of cells: each time a cell divides, the telomeres on the chromosomes get shorter, and when they are short enough, they signal the cell to stop dividing. By adding telomerase, an enzyme that rebuilds telomeres, cells cultured in a dish become immortal (the patent rights for telomerase are owned by Geron in California). Cancer cells keep on dividing in part because they reactivate their telomerase! It is not yet clear how telomeres relate to death and aging in humans, since so many

other factors play a role. It is reasonable to expect, though, that some cells and tissue—especially those that divide quickly—become crippled by shortened telomeres. Short telomeres in old people might weaken the immune system and the capability of regenerating bone and skin, all of which contain rapidly dividing cells. The more times DNA is replicated, the more likely it is that errors are introduced. This link between age and mutation frequency is clear from parental-age-dependent genetic disorders such as Down's syndrome. Dolly, the cloned sheep, died in 2003, at age seven, very young for a sheep (see Figure 2.18). It has been speculated that Dolly started off life with a telomeric "clock" set already at her mother's age. This has put somewhat of a damper on cloning enthusiasm.

Many factors are claimed to influence average lifetime. Free oxygen radicals decrease the average life span (Linus Pauling's vitamin C is an antioxidant, and he claimed that it increased lifetime!).[36] Mutant forms of the enzyme catalase decrease the average life span; food restriction and less sexual activity (or more sex later in life—figure that one out!) will make you live longer. Cooler temperatures and an efficient glucose metabolism also prolong life (goodbye Florida). Although we can expect to live longer and healthier in the new millennium, Hayflick discovered in 1961 that normal human cells, when grown in a culture dish, divide a limited number of times (mean value is 61) and then die. This ultimate ceiling for life has been dubbed the Hayflick limit. Hayflick believes that our maximum life span is fixed by our genes at about 125 years.[37–40] Evolutionarily, there is no reason for human life to go on after the reproductive years are over (just look at the Rolling Stones).

Cloning

Introduction

The last decade of the twentieth century brought one of the most important milestones in molecular biology: the creation of the cloned sheep Dolly in 1996,[41] an event that made unprecedented headlines, both in scientific journals and in the popular press. The result was that suddenly the entire world was introduced to cloning. Cloning has become a household term, but most people are still unacquainted with the scientific processes involved in cloning and how cloning is actually executed. In this section, the cloning process and its implications from a molecular biology point of view is introduced. As shown below, there are three different types of cloning: 1) DNA cloning or recombinant DNA technology, 2) reproductive cloning, and 3) therapeutic cloning.

DNA Cloning, or Recombinant DNA Technology

First, we should mention that the terms cloning, gene cloning, recombinant DNA technology, and molecular cloning are used interchangeably to denote the same genetic engineering method. The term *cloning* denotes the process by which genetic material from one species is transferred to a self-replicating genetic element, such as a bacterial plasmid or a virus. A plasmid is an extrachromosomal circular piece of DNA distinct from the normal bacterial genome that contains genes and that is capable of producing multiples copies of itself when introduced into a host organism (i.e., bacteria, yeast, mammalian, or plant cells). Thus, plasmids replicate desired genes; in nature, these genes are from the plasmid's same organism (bacteria, yeast, etc.), but in recombinant DNA technology, scientists introduce foreign genes and use the plasmid to copy said genes. For example, in genetically modified foods, a gene encoding for a natural pesticide can be incorporated into a corn cell, giving rise to corn that now can produce its own pesticide.

Of all the terms used to denote cloning, *recombinant DNA technology* is perhaps the one that best defines what the procedure entails. As explained, the method involves taking a gene that encodes for a desired protein and introducing it within a bacterial plasmid, thereby recombining this foreign DNA with that of the plasmid. More specifically, the gene to be cloned first is isolated or cut out from chromosomal DNA by employing restriction enzymes, a type of enzymes that cleave DNA when they encounter a specific sequence of DNA bases that they recognize, leaving a distinct sequence of bases at both ends. The resulting DNA fragment

containing the gene is then isolated and purified from the mixture of chromosomal DNA, so it is ready to be inserted into the plasmid. Likewise, the plasmid is also cut open, or cleaved, with the same restriction enzyme, leaving the same sequence of bases at each end of the open plasmid as those present in the isolated gene. In that manner, when the gene fragment and the open plasmid are mixed together in the presence of another enzyme—ligase, whose function is to bind or "glue" together DNA sequences—they recombine like two pieces in a puzzle. After the plasmid is constructed, it is ready to be inserted into a host—most commonly a bacterium—where it will take advantage of the host's cell machinery to replicate in large amounts. The plasmid then expresses the gene that was inserted, thus producing the protein encoded by the desired gene. Figure 2.16 depicts a generic cloning procedure performed in bacteria.

Reproductive Cloning

The process of reproductive cloning is different from DNA cloning as described above. Its goal is to create an animal from the nuclear DNA of an existing one. The principle of this method involves the removal of the nucleus of an egg containing all the DNA of the cell and replacing it with that from another cell. In the laboratory, the newly created egg is then treated with chemicals or electrical stimulation so it can divide. Once the embryo reaches a certain cellular stage, it is inserted into the uterus of a female host, or "surrogate mother," where it grows and develops until birth (see Figure 2.17). It is important to note that cloned animals are not

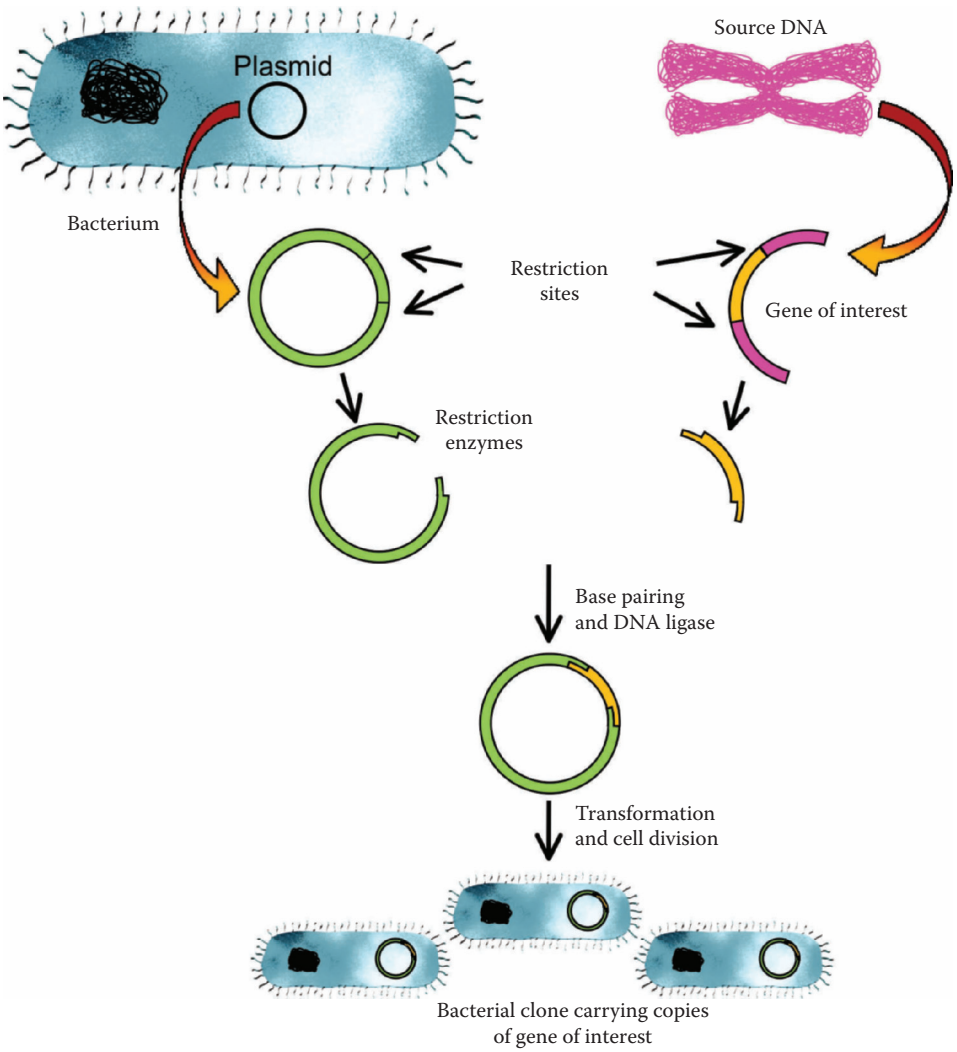


FIGURE 2.16 A generic cloning procedure performed in bacteria.

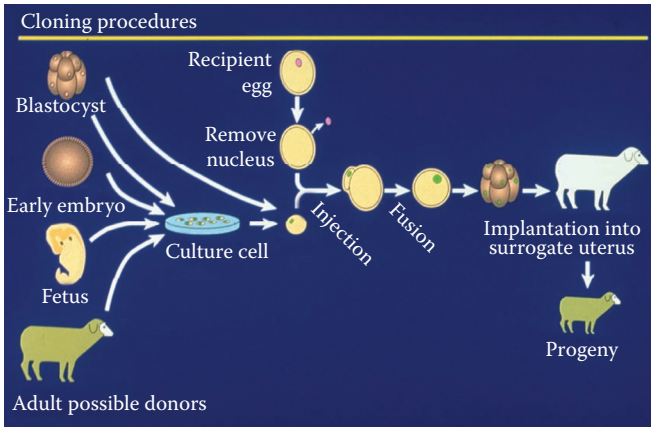


FIGURE 2.17 Reproductive cloning.

FIGURE 2.18 Dolly with her firstborn lamb, Bonnie. All of Dolly's offspring were bred the old-fashioned way. (Roslin Institute Image Library, http://www.roslin.ac.uk.)

exact replicas of their progenitors because the clones, in addition to the nuclear DNA from the progenitor, have mitochondrial DNA that is derived from the original egg. Mitochondrial DNA plays a role in the aging process, an important consideration when choosing the egg donor.

Reproductive cloning is not as effective as DNA cloning, where virtually any gene can be cloned into a suitable host, with the exception of a few incompatibilities. Reproductive cloning methods, although continuously improving, still are inefficient, with more than 90% of the embryos failing to produce viable offspring. To date, not enough data are available to allow one to predict the long-term effects of cloning, partly because a large number of the cloned animals produced are dying at a premature age. The first example of reproductive cloning, performed in 1952, resulted in a tadpole.[42]

Although other examples of cloned amphibians and other animals exist,[43] the most notable example of this type of cloning involves the creation of Dolly the sheep (named after the country singer Dolly Parton). Dolly (Figure 2.18) was cloned by a team of Scottish scientists led by Dr. Ian Wilmut,[41] currently at the Queen's Medical Research Institute at the University of Edinburgh. Dolly's cloning was exemplary in that she was the first large mammal to be cloned, thus bringing to reality the possibility of cloning other large animals, including humans. Unfortunately, Dolly's life was not easy; she suffered from crippling arthritis and lung cancer in addition to having to deal with paparazzi following her at all times—she was a Dolly Parton kind of celebrity! As a result, Dolly had to be euthanized when she was only six years old, a rather short life span compared to the eleven to twelve years of life of regular sheep.

Therapeutic Cloning, or Biomedical Cloning

The goal of this type of cloning is to produce an artificial organ or tissue that can be transplanted back into the individual who supplied the original DNA. The procedure, in its first stages, is almost identical to that of reproductive cloning, and it involves first removing the nucleus of an egg and then replacing the genetic material by inserting the nuclear DNA of the individual for whom a new organ or tissue is needed. The egg is allowed to develop until it has produced a large number of stem cells. After that, the stem cells are harvested from the embryo. This procedure results in the destruction of the embryo. The stem cells are allowed to reproduce in vitro and are chemically treated so as to develop in different types of stem cells that will give rise to the desired organ or tissue type. Once the artificial tissue or organ has been produced, it can be implanted into the patient (see Figure 2.19). The fact that the embryo is being destroyed during the cloning process brings up a number of ethical issues (see discussion in the section below). Nevertheless, the potential benefits of therapeutic cloning are tremendous, and researchers believe that these cells will be invaluable in the treatment of all sorts of diseases, among others, those of neurological origin

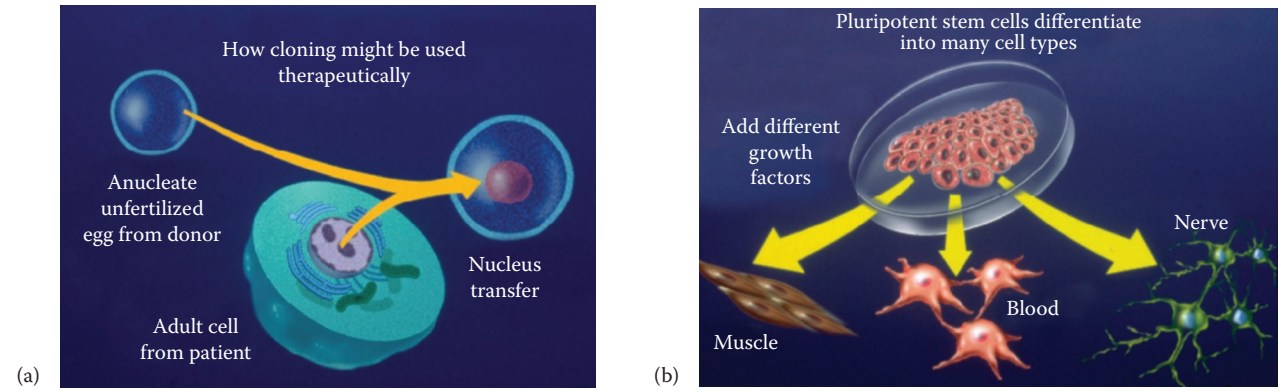


FIGURE 2.19 Therapeutic cloning.

(e.g., Parkinson's, Alzheimer's), cardiac diseases, and autoimmune diseases.

Stem Cells

The human body has more than 200 different types of cells, all derived from the same original stem cells. These stem cells are capable of reproducing themselves and transforming into different kinds, depending on what type of signal they receive while growing. This results in the production of a large number of diverse specialized cells responsible for organ and tissue formation in animals. Much research has been devoted to understanding the signaling events and pathways that lead to the transformation of a pool of identical unspecialized stem cells into this distinct very specialized cells originators of organs and tissues. Embryonic stem cells can develop into all different types of cells, whereas adult stem cells (present in fully developed adult individuals) are limited to the production of only certain types of specialized cells. In addition to embryonic and adult stem cells, rather recently, stem cells have been identified in placenta and in umbilical cord blood. The latter two types can develop into various types of blood cells. The characteristics and differences between embryonic and adult stem cells are shown in Table 2.4, below.

There is much to be learned about stem cells, but undoubtedly, the knowledge gained thus far already shows the vast number of applications and benefits that their manipulation and production will have in medicine, agriculture, and in the environment. A clear example of the potential of embryonic stem cells in medicine involves the preparation of active functioning nerve cells in mouse brains. In order to realize all the benefits of stem cells, protocols and methods will need to constantly be improved and simplified. In that regard, in 2006 researchers demonstrated the feasibility of producing embryonic stem cells from the morula (i.e., solid mass of cells that develops about six days after fertilization of an egg cell) of a mouse, thus showing that stem cells can be made without animal products in a culture. In November 2007, two separate teams of researchers were able to reprogram skin cells and turn them into stem cells.[44,45] The teams, led by James Thomson of the University of Wisconsin-Madison and by Shinya Yamanaka of Kyoto University in Japan, first identified four genes that confer the ability to reprogram certain types of cells. They then introduced these genes into skin cells from different sources and individuals by employing a retrovirus to essentially infect the skin cells and carry these genes into the nucleus of the cells. Once the new genes were part of the skin cells, they reprogrammed them, converting the skin cells into pluripotent stem cells. Because the original cells came from different types of tissue (i.e., facial skin, connective tissue, fetal skin, and baby foreskin) and from a variety of individuals (i.e., a 36-year-old woman, a 69-year-old man, a fetus, and a baby), the researchers are confident that the method will be able to be generalized and employable in most individuals. While the efficiency of the method was not the best—the Japanese team reported that 1 in 5,000 cells treated turned into a stem cell, and the US team, 1 in 10,000—both methods still result

Table 2.4 Stem Cell Overview

Characteristics of Stem Cells			
	Embryonic Stem Cells		
	In Vitro Fertilization	Nuclear Transfer	Adult Stem Cells: Adult Tissues
Properties	Produce all cell types Easy to manipulate and grow in the laboratory Can obtain embryos from in vitro fertilization clinics	Produce all cell types Easy to manipulate and grow in the laboratory Stem cells are genetically matched to patient	Successful in certain treatments Stem cells are genetically matched to patient
Limitations	Limited number of cells available for federally funded research Risk of creating tumors from transplanting undifferentiated cells	Not achieved with human cells yet Risk of creating tumors from transplanting undifferentiated cells	Results in limited numbers of cell types Not found in all tissues or organs Difficult to manipulate and grow in the laboratory
Ethical issues	Destruction of human embryos Require informed consent from embryo donor	Destruction of human embryos Require informed consent from embryo donor Possible misapplication in reproductive cloning	No ethical issues have been raised

Source: Adapted from the National Academies of Sciences, http://dels.nas.edu/dels/rpt_briefs/Understanding_Stem_Cells.pdf.

in a good number of cell lines that are viable. The implications of this discovery are vast, as this is the first real demonstration of conversion of cells from mature adult individuals into stem cells. The major downfall of these cell-reprogramming experiments is that they use a retrovirus to carry the reprogramming genes into the skin cells to be transformed. Retroviruses are known to cause tumors in human cells, and thus, the risk of inducing tumors cannot be trivialized. However, at this moment the possible positive outcomes of this new breakthrough technology outweigh the negative. Perhaps the most interesting application stemming from the work of these groups is the possibility of regeneration of organs and tissues of a patient from his/her own cells. Moreover, since the methods developed bypass the use of embryos or oocytes, the ethical dilemmas and controversy associated with their use will be avoided.

Finally, it is important to note that stem cells are also useful in studying human development, for testing new drugs and screening of toxins, and for testing gene therapy methods.

Reproductive Cloning and Stem Cell Research: Implications in Bioethics

Dolly's creation and life led to controversy, and as a consequence, reports regarding the ethics of cloning were drafted in the European Union (EU), the United States, and a number of other countries, including Canada and Australia. Specifically, in 1997, after Dolly's creation, President Clinton responded quickly to the prospect of human cloning, banning the use of federal funds for human cloning in the United States. Having considered the scientific and moral implications of said research, he also asked privately funded researchers to implement a temporary moratorium on human cloning research until the Bioethics Advisory Committee set up guidelines regarding this matter. Moreover, in 2001, then President George W. Bush created The President's Council on Bioethics (http://www.bioethics.gov/topics/cloning_index.html), whose mission is to advise the President on bioethical issues that may emerge because of advances in biomedical science and technology. Currently, there are no federal

laws in the United States that prohibit human reproductive cloning. To date, federal regulations on most human biotechnologies are not adequate, and more comprehensive regulatory policies are needed. Regarding stem cell research, in 2001, President Bush imposed a limitation on federal funding earmarked to embryonic stem cell research. This policy decision has been a source of controversy and debate, with many members of Bush's own Republican Party regretting this limitation and siding with the more progressive Democratic members of Congress. Also noteworthy are some of the most outspoken supporters of stem cell research, Republican public figures Nancy Reagan, former First Lady, and Arnold Schwarzenegger, the California governor. In that regard, then Governor Schwarzenegger led the country by proposing a bill, Proposition 71, in his state to fund stem cell research. Californians voted to pass the bill in 2004, which authorized up to $3 billion over 10 years for embryonic stem cell research in California. With this funding, the California Institute for Regenerative Medicine was created to foster stem cell research and appropriate the funds provided by the state of California. Other states have followed suit with the creation of stem cell research programs, albeit not as ambitious or comprehensive as the California one.

In 2005, the National Academies of Sciences of the United States published the Guidelines for Human Embryonic Stem Cell Research, which provide a common set of scientific and ethical guidelines aimed at aiding scientists from both the public and private sectors to establish new protocols in stem cell research. These guidelines state the need for the monitoring of all human embryonic stem cell research and outline specific methods for the derivation of new stem cell lines; for example, experimenting on human embryos by inserting stem cells into them is not allowed. Furthermore, the rapid advances in stem cell research prompted the National Academies to create a panel composed of experts in the field to monitor and review scientific developments; to change ethical, legal, and policy issues; and to prepare periodic reports to update the guidelines. More information on this can be found at http://www.nationalacademies.org/stemcells.

Within the EU, the European Convention on Human Rights and Biomedicine has prohibited human cloning in one of its protocols; however, only a few countries—namely Greece, Spain, and Portugal—have ratified this protocol. Furthermore, the ambiguity continues in the EU with regard to any kind of legal standing on reproductive human cloning despite the fact that the Charter of Fundamental Rights of the European Union has prohibited its practice. The United Nations decided to introduce debate on this issue among the country members of its General Assembly; as a result, in March 2005, the United Nations Declaration on Human Cloning was adopted.

For information on National Polices Governing New Technologies of Human Genetic Modification in a large number of different nations, see http://geneticsandsociety.org/article.php?id=304. For readers interested in the topic of bioethics, information can be found at http://ethics.sandiego.edu/Applied/Bioethics/index.asp. In addition, a good source of information and interesting reading can be found on the Web site of Professor Lee Silver of Princeton University, an expert in bioethics, http://www.leemsilver.net/challenging/index.htm. Dr. Silver's views are provocative as well as entertaining, and interested readers will enjoy his written work and audiovisual pieces.

Biological Cells and Their Machines

Prokaryotes and Eukaryotes

Earlier, we learned how cells might have emerged from a soup of chemicals, and their more detailed construction and operation will be instructive for the future of miniaturization science. Living cells and their organelles have dimensions in the micrometer and submicrometer range and constitute the first truly autonomous micromachines. There is much more membrane area (99%) surrounding the internal organelles than there is outer membrane surface (1%).

Microscopic sizes are the normal working size for the biological world, a biological cell of about 5–10 μm being one of the larger components. The human body contains approximately 100 trillion of these cells. Some of the simplest organisms, such as

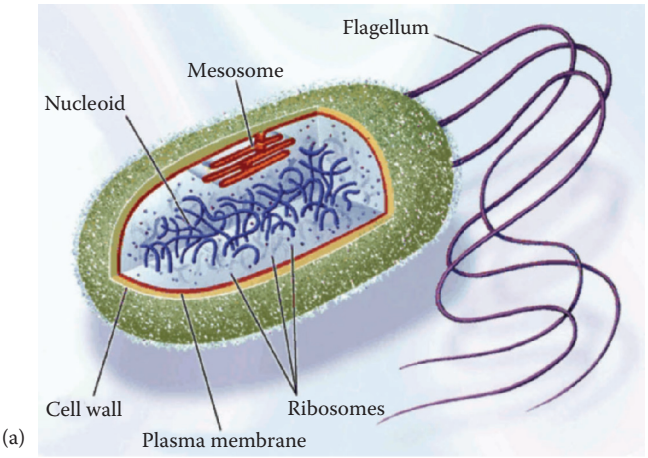
bacteria and algae, have no cell nucleus, and their genetic material is mixed in with the aqueous cytoplasm. These life forms are known as *prokaryotes*, a designation used to convey their relatively simple status as protocells, and they are the forerunners to multicellular organisms. Most multicellular organisms have their genetic material in a separate chamber within the cell, the *nucleus*, and are called *eukaryotes*, literally, "true nucleus." These two types of cells are compared in Figure 2.20.

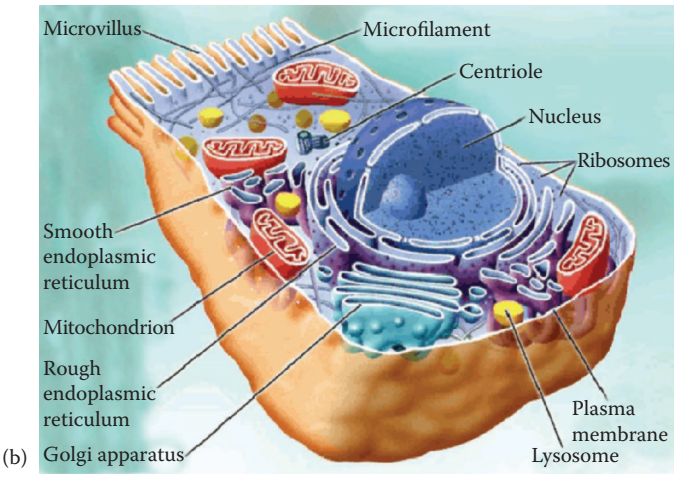
Prokaryotes (from bacteria to blue-green algae) took about 500 million years to evolve, but the evolution of the first eukaryotic cells took 2500 million years. Clearly the development of this more elaborate construction was a crucial step in the evolution toward more complex life forms; early eukaryotic cells began the evolutionary path that led to complex multicellular animals and plants. The more complex a multicellular organism, the more different types of cells it has; a blue whale, for example, has approximately 120 types of cells, whereas a mushroom has fewer than 10. Eukaryotic cells come in a very wide variety of shapes and sizes, but on average a cell has a volume of about 3.4×10^{-9} ml and weighs roughly 3.5×10^{-9} grams. Cell size, on the upper end (400 μm^3), is limited by the ability of diffusion to bring nutrients to the appropriate parts inside the cell and to dispose of waste products. The smallest size for a free-living organism (0.02 μm^3) most likely is set by the catalytic efficiency of enzyme and the protein synthetic chemistry machinery.

Organelles

Introduction

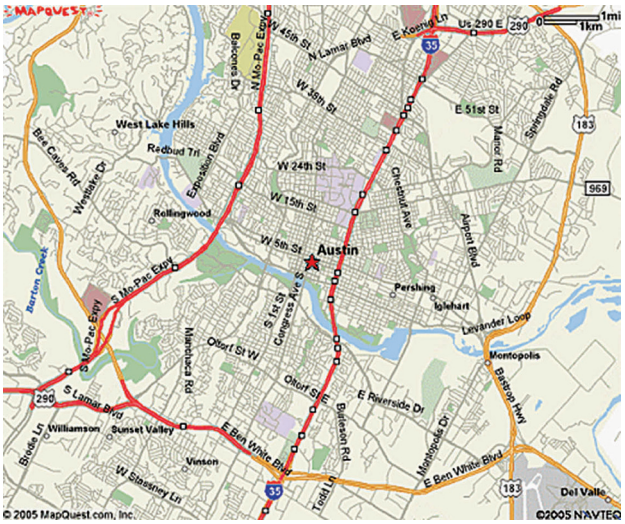
Organelles, visible in the eukaryotic cell in Figure 2.20b, are specialized structures. The functions of several of these cell organelles are described in the following sections. As depicted in Figure 2.21, a cell might be compared to a city, with the corresponding functions listed below the images. The high degree of spatial/temporal organization of molecules and organelles within cells is made possible by protein machines that transport components to various destinations within the cytoplasm. These different tiny motors are specialized for specific tasks, such as cell division, cell movement, crawling, maintenance of cell shape, movement of internal organelles, synthesis of ATP, RNA synthesis, DNA synthesis, etc.

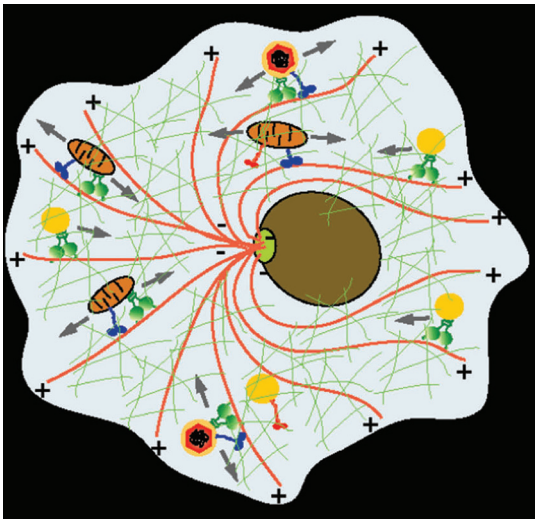




Prokaryotes	Eukaryotes
No nucleus—have a nuclear "area"	Nucleus present
No membrane-bound organelles	Have membrane-bound organelles
DNA present usually as 1 large chromosome	DNA in nucleus
Primitive cells usually quite small	"Modern" cells, can be quite large
Archebacteria, Eubacteria	Plants, Animals, Fungi, Protists

FIGURE 2.20 Comparison of a prokaryotic cell (a) and an eukaryotic cell (b).





Workers/machines	Proteins
Roads	Actin filaments
Highways	Microtubules
Trucks	Molecular motors
Power plants	Mitochondria
Factories	Ribosomes
Library	Nucleus with DNA
Post office	Golgi apparatus
Police	Chaperones
Gates, keys, passes	Ion channels
Electric fences	Membranes
Train tracks	Actin filament network
Motors, generators	ATP synthases
Vehicles	Hemoglobin
Train control center	Centrosome
Copy machines	Polymerases
Chain couplers	Ligases
Bulldozers, destroyers	Proteases, proteasomes
Internet nodes	Neuron synapses
Mail sorting machines	Protein sorting machines

FIGURE 2.21 The cell is like a city.

Mitochondria and Phylogenetic Trees

The nucleus contains the genetic material (library), and the mitochondria (power plants), through utilization of pyruvate and oxygen, produce ATP (adenosine triphosphate), a compound that provides energy for cellular activities. There is convincing evidence that these mitochondria used to be independent structures with a life of their own that became incorporated into eukaryotic cells over time—very much like bacteria that permanently colonize more complex cells.[46] Mitochondria, for example, contain DNA that uses a slightly different code for amino acids than does the same DNA in the nucleus of the cell. As in bacteria, the length of DNA is closed into a loop and the mitochondria reproduced by splitting in two. These are amazing findings, as the DNA code for amino acids is otherwise the same for species as different as humans and fish (see under "The Genetic Code"). While each of us necessarily has two parents, our mitochondria and mitochondrial DNA come from the ovum and hence from our mothers. Our mothers, in turn, got their mitochondrial DNA from their

mothers, and so on. Thus, while our nuclear DNA is a mish-mash of the DNA of our four grandparents, mitochondrial DNA (mtDNA) is an almost exact copy of the DNA of our maternal grandmother. The match is not exact because of mutations: mtDNA has a pointwise mutation substitution rate 10 times faster than nuclear DNA—so it is a better record of evolutionary change.

Evolutionary trees or phylogenies can be reconstructed by comparing mtDNA (Figure 2.22). In fact, the mutations in the mitochondrial DNA provide the molecular clock that allows us to determine how much time has elapsed since the *mitochondrial* Eve lived. University of California Berkeley's Allan Wilson, in 1987, compared mtDNA of many racial groups.[47] Knowing the average mutation rate, he figured out that we all derive from a single version of mtDNA that existed 200,000 years ago in the cells of an African woman—the common ancestor of all of humanity. This work is now seen as overly simplistic. Construction of an exact phylogenic tree is indeed still controversial: for example, it is not necessarily true that the molecular clock is constant; i.e., mutation rates may change, and environmental pressure may change (http://www.sciencemag.org/feature/data/tol). Moreover, the mathematical descriptions and algorithms that may lead to a historically correct phylogenetic tree remain to be developed (http://tolweb.org/tree).

Mitochondria and Power

Here we dwell for a moment as we consider the operation of the cells' power stations—the mitochondria—in more detail (Figure 2.23). We study how cells derive their power from the sun by relying on mitochondria. Energy is required for mechanical work, transportation, and chemical work. When considering building artificial cells, powering the micro- or nanounits is one of the first priorities. This might mean, in human engineering terms, capturing radiant energy (e.g., sunlight) with a solar cell

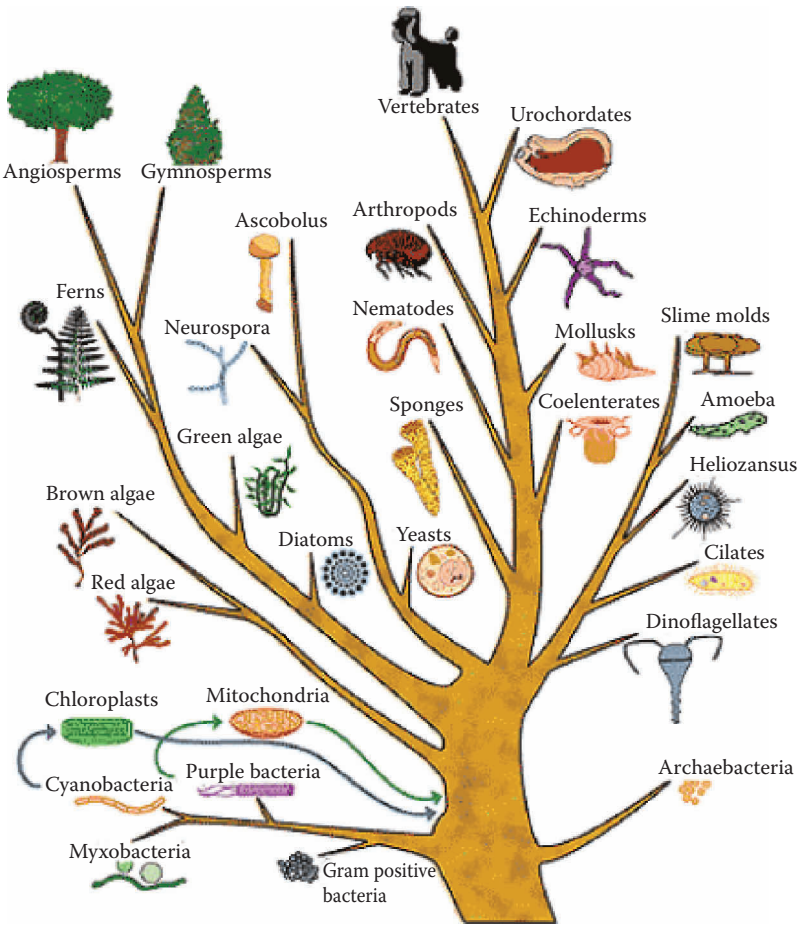


FIGURE 2.22 Evolutionary trees or phylogenies can be reconstructed by comparing mitochondrial DNA.

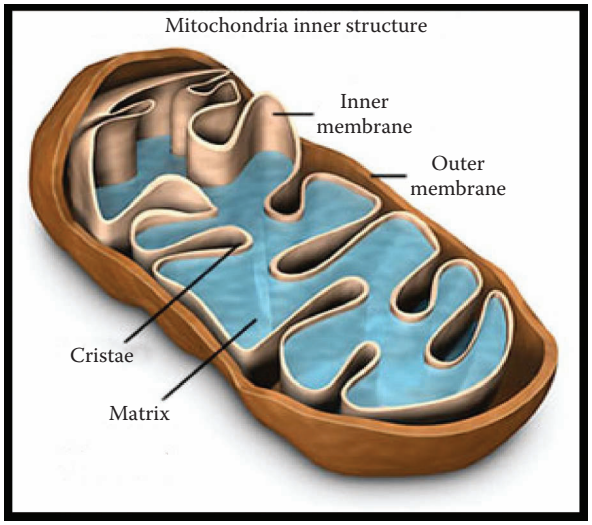


FIGURE 2.23 Mitochondria.

and converting it to a fuel that can be stored for later use (e.g., hydrogen through water electrolysis).

In Figure 2.24a we show the molecular structure of the energy currency of a biological cell, i.e., adenosine triphosphate or ATP, the immediate source of energy in all cells of all living organisms on Earth. The structure of an ATP molecule contains three molecules of phosphate linked to each other; at the pH of normal biological fluids, they have a negative charge. When energy is required, ATP is hydrolyzed to adenosine diphosphate (ADP) and inorganic phosphate (P_i). Back in the mitochondria, ADP is recycled back to ATP almost immediately, so that the energy is again available to do needed work (recharging of the batteries) (Figure 2.24b). The production of ATP is due to three different cell processes, namely photosynthesis, fermentation, and respiration. In photosynthesis, nature uses stacks of thylakoids in a chloroplast, a plant organelle as shown in Figure 2.25. We will detail the photosynthesis process, the conversion of carbon dioxide and water into glucose with the help of sunlight, shortly.

For now it suffices to say that through photosynthesis, plants, algae, and some bacteria store solar energy in the form of the high-energy chemical bonds of glucose. In that process carbon dioxide and water are consumed, and some leftover oxygen is released in the air. This energy consuming process is said to be anabolic; i.e., large molecules are synthesized from smaller ones in a scenario schematically illustrated in Figure 2.26a ($\Delta G > 0$). Most living organisms (including plants and animals) then reverse photosynthesis by taking glucose apart and releasing energy to power their own growth and development. In this second process, oxygen is consumed, and water and carbon dioxide, the original ingredients consumed by the plants, algae or bacteria, are released. This energy producing process is said to be catabolic, i.e., breaking down complex

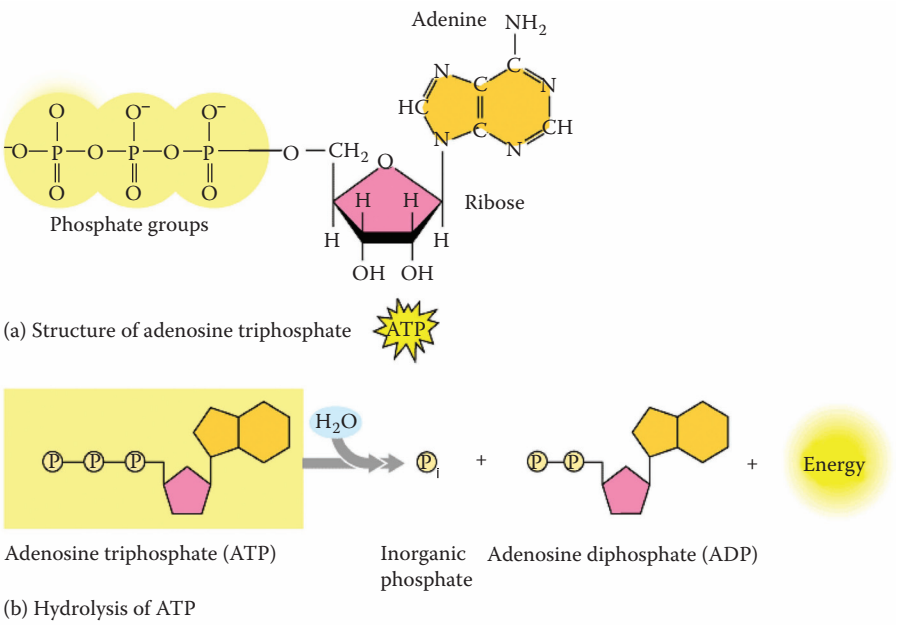


FIGURE 2.24 (a) Molecular structure of adenosine triphosphate (ATP). (b) Hydrolysis of ATP to ADP.

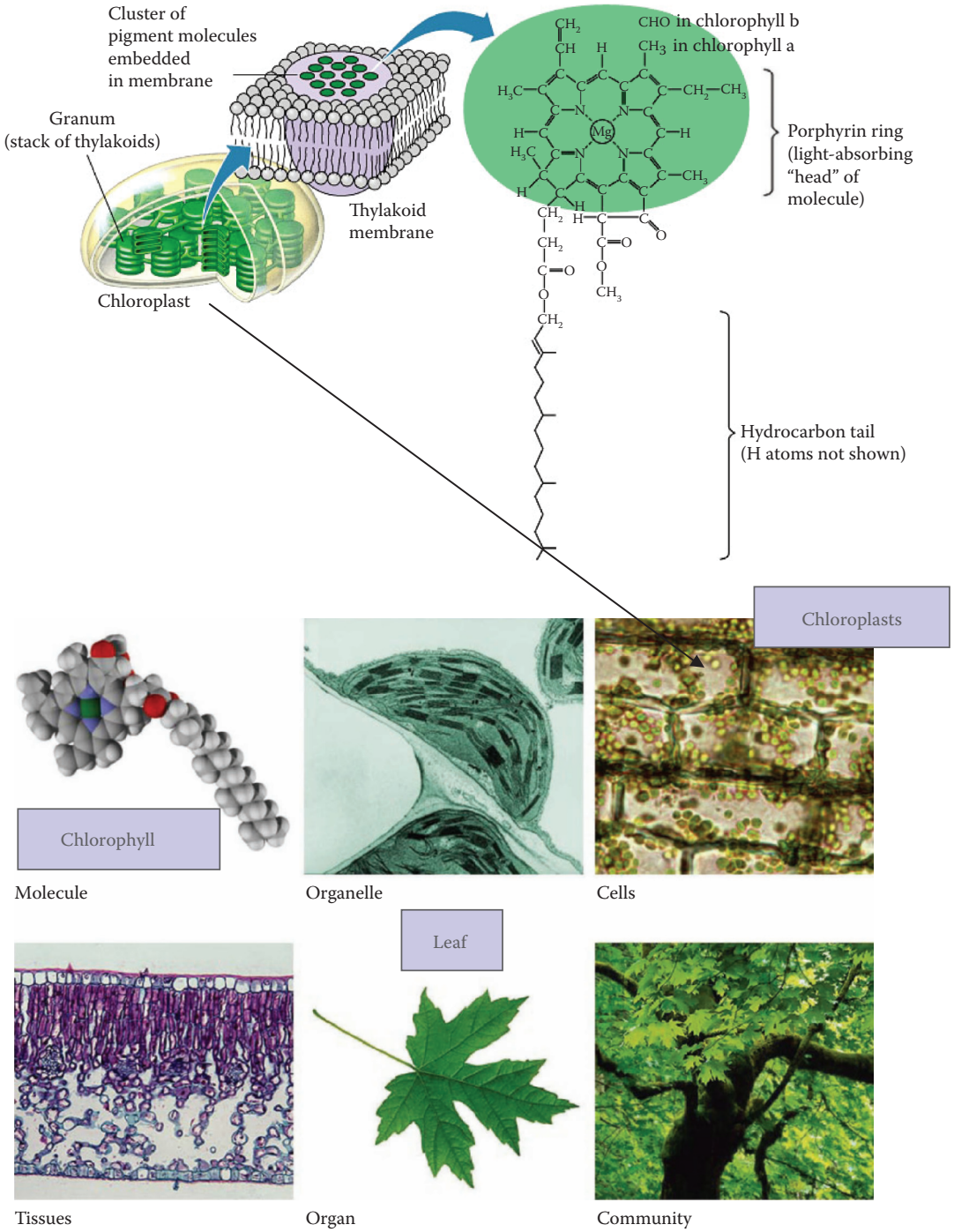
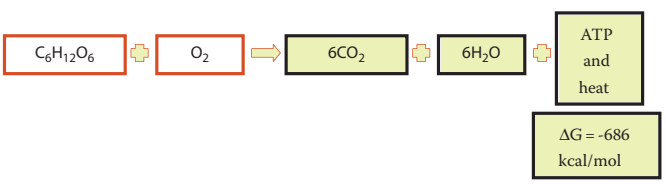


FIGURE 2.25 Photosynthesis in plants: from trees to pigment molecules in the thylakoids in chloroplasts.

molecules into smaller, simpler ones. This is illustrated in Figure 2.26b ($\Delta G < 0$).

Direct oxidation of glucose with oxygen would release 686 kcal/mol ($\Delta G = -686$ kcal/mol) (Reaction 2.1), but cells cannot use the energy stored in glucose directly.

In the cytoplasm of a cell, first, nine different enzymes must help convert the glucose (a chain of six carbons with twelve hydrogen atoms and six oxygen atoms) in nine successive chemical reactions



Reaction 2.1

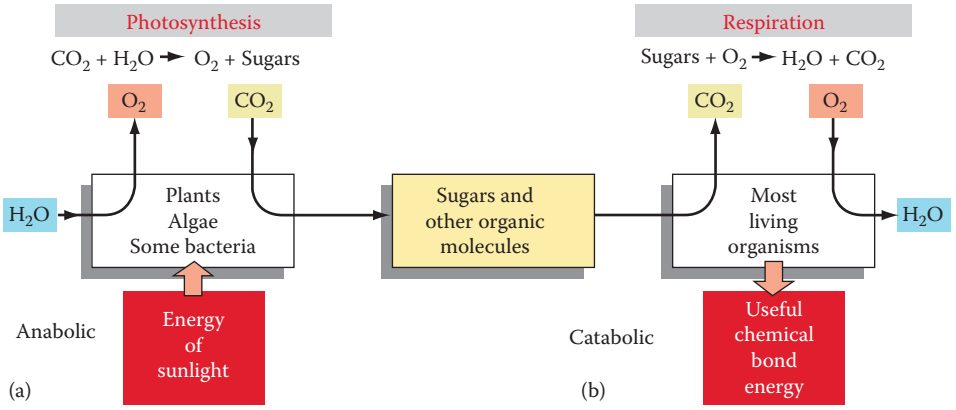


FIGURE 2.26 (a) Anabolic process: use of energy to conduct unfavorable/uphill reactions (from small to large molecules). (b) Catabolic processes: fuel breakdown to deliver energy (from large molecules to small).

to two molecules of pyruvate (a three-carbon chain with three hydrogens and three oxygens). This complicated process is called *glycolysis*. After glycolysis, it is the pyruvate that carries on as the energy source originally produced with the solar energy input, and transports it to the mitochondria. The mitochondria takes the energy garnered during photosynthesis (in the case of plants) or the digested food (in the case of animals) and stored as pyruvate and converts it to ATP. This conversion in the mitchondria occurs at a very high surface inner membrane in a complex set of reactions, including the famous Krebs cycle. The Krebs cycle is an aerobic process (i.e., it involves the use of oxygen) that is responsible for cellular respiration. Hydrolysis of the ATP result in the release of one of three P_i groups, along with 7 kcal/mol, and the formation of ADP (Figure 2.24b). ATP thus packs a wallop of concentrated solar power. These tiny solar batteries are released by the mitochondria into the cell to power all types of uphill reactions ($\Delta G > 0$). In releasing their power, ATP battery molecules shed instantly one phosphate (P_i) and are discharged to become ADP (ATP minus one P_i). In biochemistry, phosphorylation is the addition of a P_i group to a protein or a small molecule. Beyond doubt, protein phosphorylation is the most important regulatory event in eukaryotic cells. Many enzymes and receptors are turned on or off by phosphorylation and dephosphorylation. Phosphorylation is catalyzed by various specific protein kinases, whereas phosphatases dephosphorylate. The discharged ADP solar battery, in need of recharging now, diffuses until it finds a mitochondria power station, where it will regain its lost phosphate. Cells obviously recycle the ATP batteries they use for work.

To make ATP from pyruvate, as we show in Figure 2.27, the most prevalent and efficient pathway is cellular aerobic respiration. In cellular aerobic respiration, the conversion of 1 mol of glucose yields up to 36 mol of ATP; therefore, the energy efficiency is 36×7 kcal/mol divided by 686 kcal/mol = 252/686 = 36.7% (compared with about 3% efficiency in the burning of oil or gasoline). In the absence

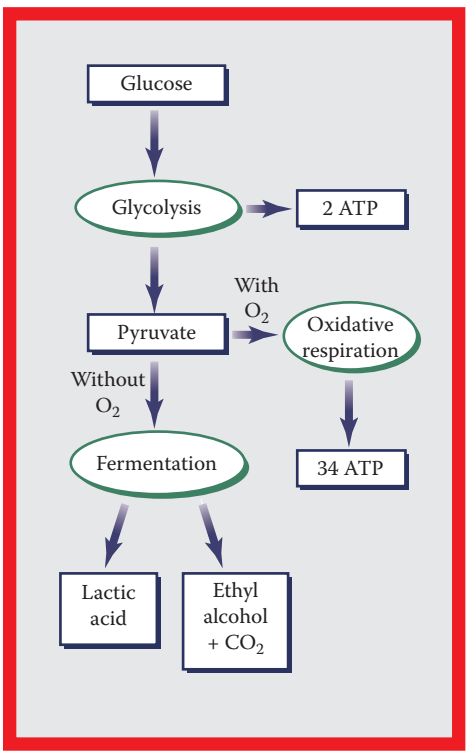


FIGURE 2.27 Various pathways to making of pyruvate, the food source for mitochondria.

of oxygen, certain cells are still able to make ATP, albeit in a less efficient matter. This can be accomplished through fermentation and anaerobic respiration. Fermentation is the breakdown of sugars in the absence of oxygen to release energy, carbon dioxide, and alcohol (or lactic acid). Yeast cells produce ethyl alcohol by fermentation.

Some prokaryotes are able to carry out anaerobic respiration, respiration in which an inorganic molecule other than oxygen (O_2) is the final electron acceptor. For example, certain bacteria known as sulfate reducers can transfer electrons to sulfate (SO_4^{2-}), reducing it to H_2S. Other bacteria, called nitrate reducers, can transfer electrons to nitrate (NO_3^-), reducing it to nitrite (NO_2^-). Other nitrate reducers can reduce nitrate even further, to nitrous oxide (NO) or nitrogen gas (N_2). The total energy yield per glucose oxidized is less than with aerobic respiration, with a theoretical maximum yield of 36 ATP molecules or fewer.

Cellular respiration, fermentation, and anaerobic respiration are catabolic, energy-yielding pathways. The totality of an organism's chemical reactions is called *metabolism*. Metabolism manages the material and energy resources of the cell. In Figure 2.28 the various processes that ATP energizes are summarized.

Since animals cannot make food directly from water and CO_2, chloroplasts and mitchondria work in tandem, as illustrated in Figure 2.29. Photosynthesis and respiration are complementary processes in the living world. Photosynthesis uses the energy of sunlight to produce sugars and other organic molecules. These molecules, in turn, serve as energy sources or food. Respiration is a process that uses O_2 and forms CO_2 from the same carbon atoms that had been taken up as CO_2 and converted into sugars by photosynthesis. In respiration, organisms obtain the energy that they need to survive. It is the cycling of molecules between chloroplasts and mitochondria that allows a flow of energy from the sun through all living things. Photosynthesis preceded respiration on the Earth by, probably, billions of years before enough O_2 was released to create an atmosphere rich in oxygen. Ultimately, the process of photosynthesis (Reaction 2.2) can be considered one of the most, if not *the* most, important chemical reaction on Earth.

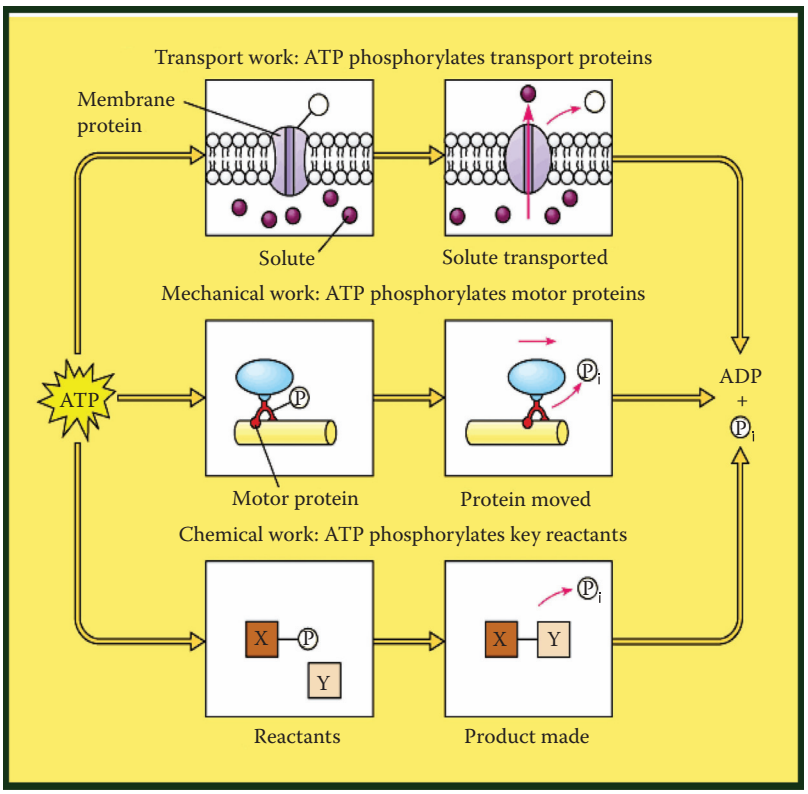


FIGURE 2.28 The various processes that ATP energizes: ATP at work.

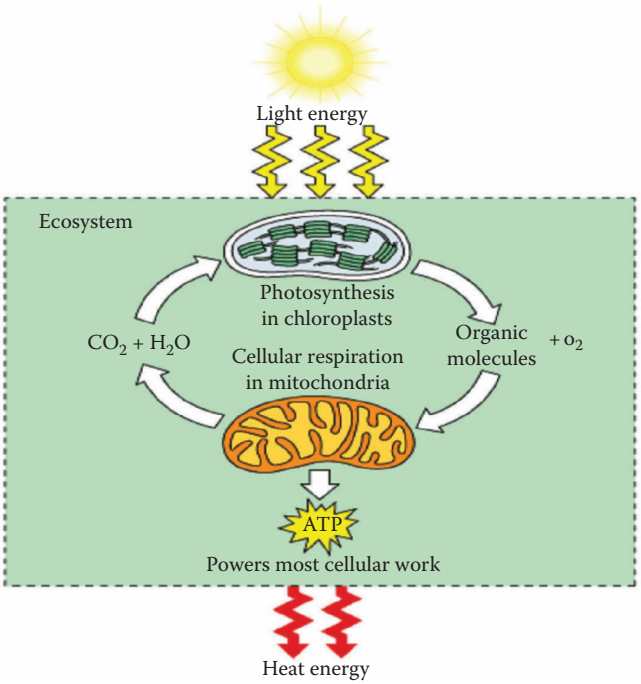


FIGURE 2.29 From light to photosynthesis to cellular respiration and heat energy.

$$6CO_2 + 12H_2O + \text{light energy} \Rightarrow C_6H_{12}O_6 + 6O_2 + 6H_2O \qquad \text{Reaction 2.2}$$

For building artificial cells, we would be better off building plant like-cells, as the source for fuel is in the atmosphere (water and/or CO_2) and harvesting energy from the sunlight is relatively simple. So let us take a more detailed look at how the chloroplasts in plant leaves do it (Figure 2.25).

Pigments in the leaves of plants absorb sunlight, principally through chlorophyll in the chloroplasts. Chloroplasts are found mainly in mesophyll cells forming the tissues in the interior of the leaf (Figure 2.25). Chlorophyll comes in two slightly different forms, called chl a and chl b. Most leaves contain at least two additional types of pigments, carotenes and xanthophylls, that are capable of absorbing light of different wavelengths. The more abundant chlorophylls usually mask the latter two types of pigments. There are about half a million chloroplasts per square millimeter of leaf surface. The color of a leaf comes from chlorophyll, the green pigment in the chloroplasts. Oxygen exits and carbon dioxide enters the leaf through microscopic pores, stomata, in the leaf. Veins deliver water from the roots, and carry off sugar from mesophyll cells to other plant areas (Figure 2.30).

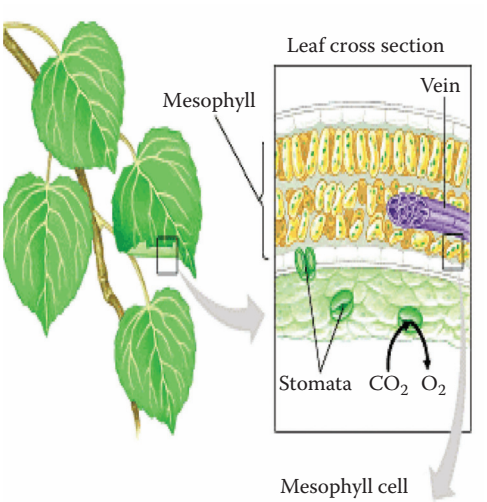


FIGURE 2.30 Stomata and veins in a green leaf.

Only chl a can participate directly in the light reactions. Other pigments in the thylakoid membrane (e.g., chl b, carotenoids) can absorb light and transfer the energy to chl a. These pigments thus broaden the spectrum of solar light (wavelengths) that can drive photosynthesis. Plants only use a narrow range of frequencies from sunlight between 400 and 700 nm. Light reactions on the thylakoid membrane convert solar energy and produce oxygen, and dark reactions fix CO_2 in the Calvin-Benson cycle to make sugar, as shown in Figure 2.31.

Molecular Motors

Other subcomponents of biological cells exhibit equally fascinating feats of nanoengineering. For illustration, consider molecular motors, composed of a discrete number of atoms that produce molecular or supramolecular motions. Molecular motors derive energy from ATP that is present in the cytoplasm and nucleoplasm of every cell and is the universal energy currency for all biological processes (see Figure 2.24). Living things can use ATP like a battery. Molecular motors are in a sense nanomachines that do mechanical work through the hydrolysis of ATP, the energy source of the cell. During hydrolysis of ATP, a shape change occurs within the motor protein (a mechanochemical process), and mechanical work is performed. This mechanical work is used to move the motor along a track in order to perform a load transport function or to apply forces to the filament for cell motility and cell

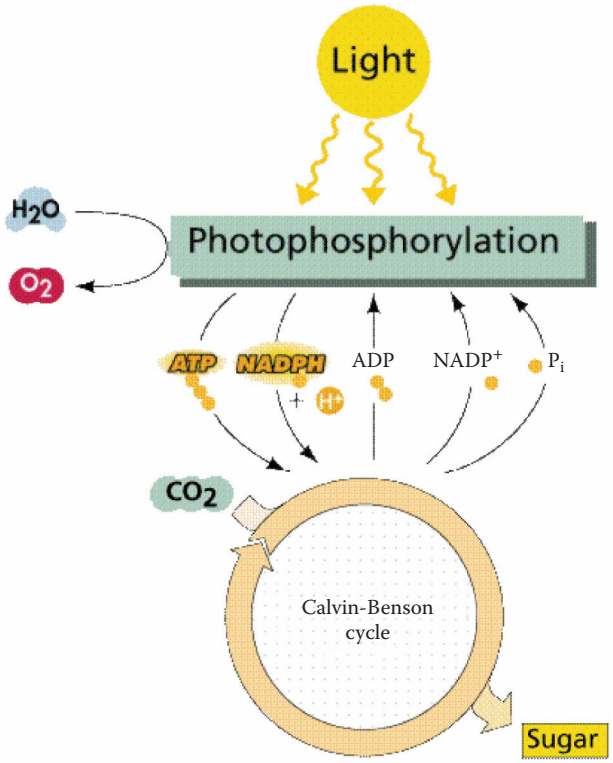


FIGURE 2.31 Photosynthesis with light reaction and dark reaction in the Calvin-Benson cycle.

division. Three types of motors can be distinguished, i.e., linear motors (such as myosin, dynein, and kinesin), rotary motors (F_0-F_1 ATP synthase) and polymerization motors (RNA, ssDNA); their tracks and functions are summarized in Table 2.5. Many of proteins can be considered nanoactuators, not only rotary or linear motor proteins. In some cases, even a simple binding of a protein and a specific target molecule—a chemical reaction—results in definite shape change—a mechanical effect. A conformational change might be put to good use in genetically engineered nanoactuators or biosensors (see calmodulin example, below).

Kinesin is a linear motor protein that steps along the microtubuli, the tubelike fibrils of the cell's skeleton, and is responsible for carrying organelles over a long distance within the cell (Figure 2.32). Most kinesins "walk" toward the plus ends of microtubules. A single kinesin molecule can drag along relatively large things like chromosomes and other organelles. Svoboda et al. demonstrated that kinesin can even work in vitro on a microscope slide.[48] This was demonstrated by gluing kinesin on microtubuli and observing it drag along a little ball of silica gel. Its strength is about 5 piconewtons (the force a laser pointer makes on a screen). Optical tweezers are well suited for studying molecular motors because of their low spring constants. Using optical traps, this same group determined that kinesin moves in individual steps that are 8 nm (80 Å) in length, a size consistent with its dimensions (Figure 2.33).[48] Microtubules can also move themselves on a glass slide surface containing bound kinesin (a single molecule bound to glass can move a microtubule). For the latest on kinesin research, visit the kinesin homepage at http://www.proweb.org/kinesin.

Noji et al. discovered that the γ-subunit of the hydrophilic F_1 portion of ATPase rotates in response to the synthesis/hydrolysis of ATP (Figure 2.34).[49] The moving part of an ATPase is a central protein shaft (or rotor, in electric-motor terms) that rotates in response to electrochemical reactions with each of the molecule's three proton channels (comparable to the electromagnets in the stator coil of an electric motor). The ubiquitous ATPase was the first rotary motor enzyme ever found, and the force generated by this motor protein (>100 pN) is among the greatest of any known molecular motor. The enzyme puts out a very large torque, comparable

TABLE 2.5 Molecular Motors and Their Tracks and Functions

Type	Track	Function
Myosin	F-actin, filament	Cell crawling, cell division, muscle contraction
Dynein	Microtubules	Eukaryotic cell (produces the axonemal beating of cilia and flagella and also transports cargo along microtubules toward the cell nucleus)
Kinesin	Microtubules	Organelle transport, chromosome segregation, moves cargo inside cells away from the nucleus along microtubules
Polymerase	RNA, single-stranded	Nuclear processes (transcribes RNA from a DNA template)

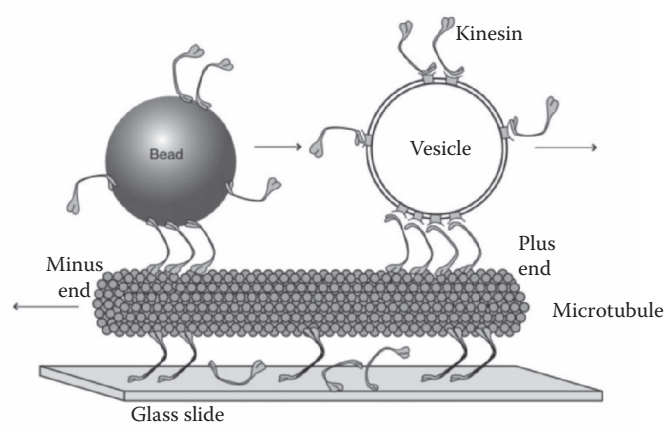


FIGURE 2.32 Kinesin motor. Schematic diagram showing movements mediated by kinesin, an ATP-driven motor, on microtubule tracks. Vesicles and beads containing kinesin molecules on their surface move toward the plus end of microtubules. Microtubules can also move themselves along a glass slide containing bound kinesin.[48]

to a person rotating a 150-m rod. The enzyme can spin such a long filament because it automatically ratchets down the rotation rate. A no-load rotational velocity of 1020 rpm was calculated, and the motor has a diameter of less than 12 nm. Montemagno et al.,[50] then at Cornell University, integrated these genetically engineered motor proteins with NEMS, for example, by attaching the protein nanomotors to arrays of nickel posts and attaching nickel rotor blades to the protein molecules.[50] The "handle" for attaching the ATPase motor to the nanofabricated metallic substrates is a synthetic peptide composed of histidine and other amino acids. The histidine peptide allows the molecular motors to adhere to nanofabricated patterns of gold, copper, or nickel—the three standard contact materials in integrated circuits.

The researchers attached fluorescent microspheres (~1 μm in diameter) to the ATPase molecule's rotor and observed the microsphere movement with a differential interferometer and with a charge-coupled device camera.

The authors envision that F_1-ATPase motors will be used to pump fluids and open and close valves in nano- and microfluidic devices and provide mechanical drives for a new class of nanomechanical devices.[50] This work was cited by *Discover* magazine as one of the most promising new technologies of 1999. Today, twelve years later, that promise remains unmet.

Bacteria swim in viscous liquid environments by rotating helical propellers known as flagella, shown in Figure 2.35 (see also Volume I, Chapter 6, "Mixing, Stirring, and Diffusion in Low Reynolds Number Fluids"). The bacterial flagellum is a nanomachine made of about twenty-five different proteins, each of them in multiple copies ranging from a few to tens of thousands. It is constructed by self-assembly of these large numbers of proteins, each into a different

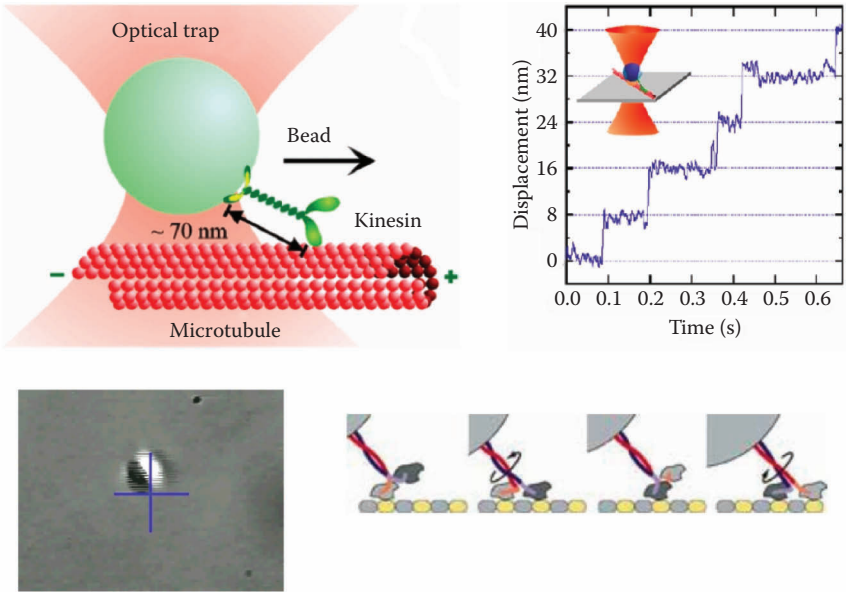


FIGURE 2.33 Laser tweezers were used to study kinesin motors with the kinesin molecule attached to a bead traveling over a microtubule. (From http://www.stanford.edu/group/blocklab.)

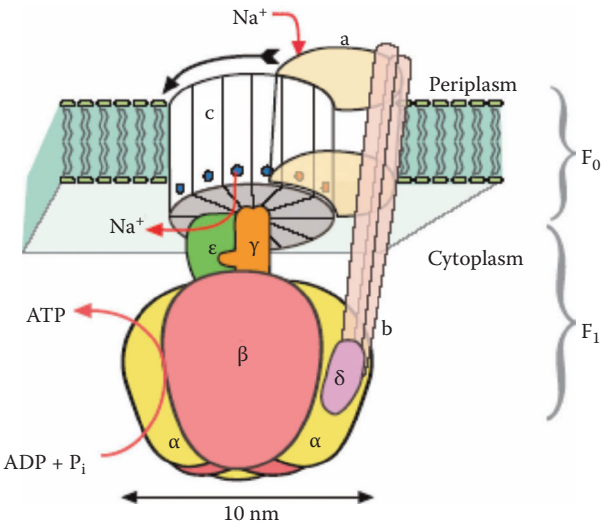


FIGURE 2.34 ATPase motors. A single molecule of F_0F_1-ATPase acts as a rotary motor—the smallest known. A central rotor of radius ~1 nm, formed by its γ-subunit, turns in a stator barrel of radius ~5 nm formed by three α- and three β-subunits.

part that exerts a different function, such as a rotary motor, bushing, drive shaft, rotation-switch regulator, universal joint, helical propeller, and rotary promoter for self-assembly. The rotary motor, with a diameter of only 30–40 nm, drives the rotation of the flagellum at around 300 Hz, at a power level of 10–16 W, with energy conversion efficiency close to 100%. These protein motors appear to be fueled by a flow of protons across a pH gradient and can push an average-sized cell at 30,000 nm/s or 15 body-lengths per second—and can also run in reverse![51] In comparison, the surface micromachined motors of Volume II, Chapter 7, are hopelessly impractical.

The structural designs and functional mechanisms revealed in the complex machinery of the bacterial flagellum could provide many novel technologies that may become the basis for future nanotechnology, from which we should be able to find many useful applications.

Mammalian muscle, nature's ubiquitous actuator for larger organisms, is another chemomechanical actuator. Muscle cells are long and cylindrical and made up of thick myosin filaments surrounded by thin actin filaments. A central myosin filament and surrounding actin filaments form a sarcomere unit, which has a length of 2.5 μm and a thickness of 10–20 nm. The thin and thick filaments are interdigitated, and upon muscle contraction, the thick filaments are pulled along the thin filament in a ratchet-like manner (Figure 2.36). The linear motion is effected by the myosin head, which is attached to the thick filament via a flexible hinge and binds to the actin filament (Figure 2.37). This linear protein motor again uses chemical energy released in the conversion of ATP to ADP and a phosphate ion (P_i). The hydrolysis of 1 mol of ATP releases about 10 kcal of energy (i.e., 6.96×10^{-20} J per molecule).

The maximum static muscle force generated per unit cross-sectional area (i.e., tension, stress) is about 0.350 MPa, a constant number for all vertebrate muscle fibers. In vertebrate muscle, the maximum force generated can be held only for short periods of time because of muscle fatigue. The maximum sustainable force usually is about 30% of the peak value. For this reason, the maximum static sustainable stress generated by muscle is about 0.100 MPa. The maximum power per unit mass (in W/kg) is an important figure of merit for robotic and prosthetic actuators. For human muscle, it typically measures

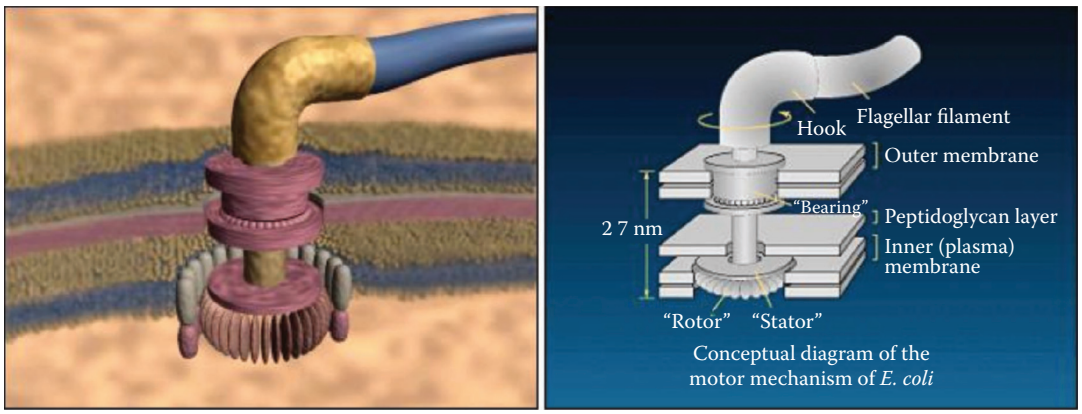


FIGURE 2.35 Bacteria swim in viscous liquid environments by rotating helical propellers called flagella.

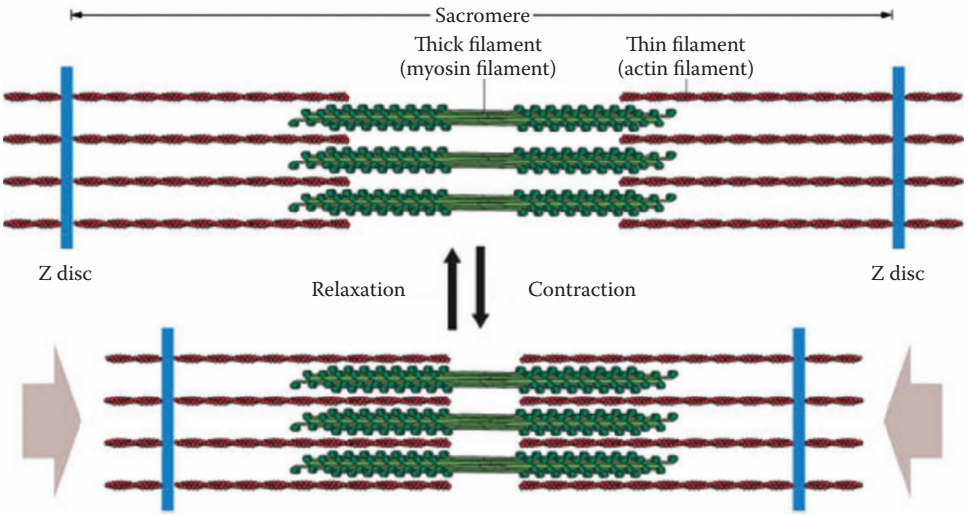


FIGURE 2.36 Muscle cells: the linear motion of muscles, which results in muscle contraction, is driven by chemical energy as ATP is hydrolyzed.

about 50 W kg^{-1} but can be as high as 200 W kg^{-1} for some muscles for a brief period of time. To illustrate the excellent cycle lifetime of a muscle, we need only look at cardiac muscle. The heart beats more than 3×10^9 times in the lifetime of an average person; an excellent lifetime compared with any artificial actuator. When trying to mimic a muscle with an artificial actuator, one of the most difficult properties to emulate is the extreme change in stiffness occurring between resting muscle and maximally activated muscle. Stiffness can increase as much as 5 times from rest to a 100% contraction.[52] On the basis of these observations, researchers preparing artificial muscles are now shifting from using stiff inorganic materials with a high Young's modulus, such as Si, to low-Young's modulus organic materials, such as hydrogels and redox polymers.

The beating motion of hair-like cilia on cells and of flagella depends on an interplay of a pair of proteins, dynein and tubulin (see Figure 2.38). Dyneins walk toward the minus ends (the microtubule part with less expand and shrink activity) of microtubules, causing sliding of microtubules, which leads to a beating motion. This protein pair is equivalent to the myosin and actin pair in muscle tissue (see Figure 2.37).

Unlike the cells of plants and fungi, animal cells do not have a rigid cell wall. This feature was lost in the distant past by the single-celled organisms that

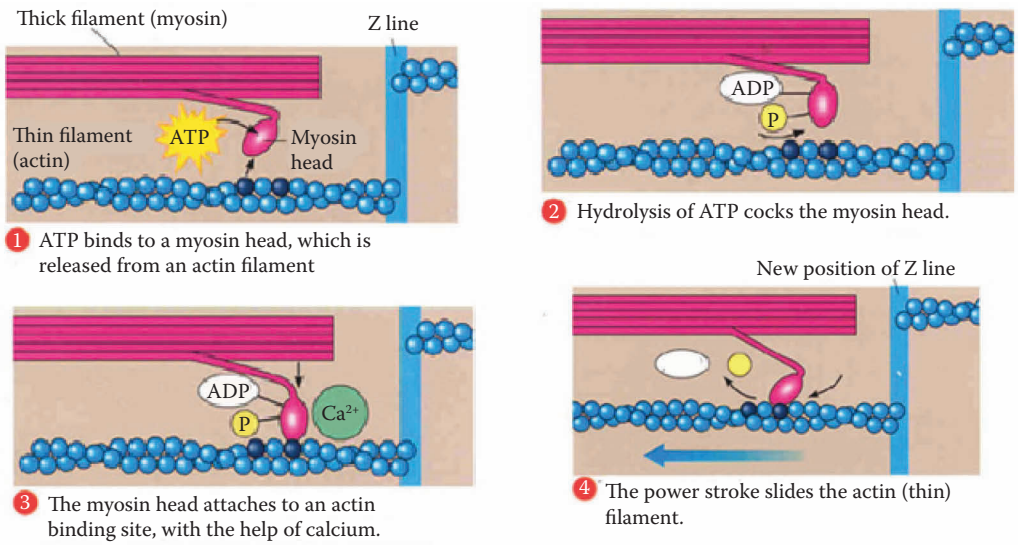


FIGURE 2.37 The linear motion is effected by the myosin head, which is attached to the thick filament via a flexible hinge, and binds to the actin filament.

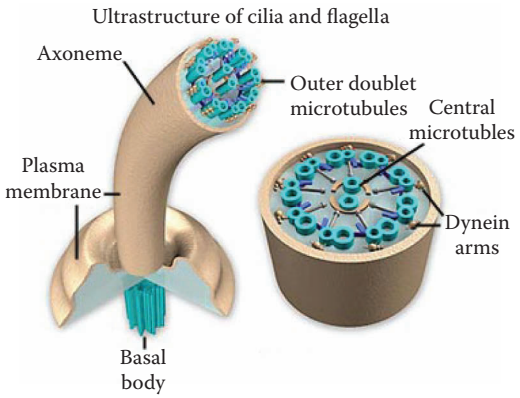


FIGURE 2.38 Dynein, found in cilia and flagella, causes a beating motion.

gave rise to the kingdom Animalia. Most eukaryotic cells, both animal and plant, range in size between 1 and 100 μm. The lack of a rigid cell wall allowed animals to develop a greater diversity of cell types, tissues, and organs. Specialized cells that formed nerves and muscles—tissues impossible for plants to evolve—gave these organisms mobility. The ability to move about by the use of specialized muscle tissues is a hallmark of the animal world, though a few animals, primarily sponges, do not possess differentiated tissues. Notably, protozoans locomote, but only via nonmuscular means, in effect, using cilia, flagella, and pseudopodia.

Different types of cells, such as neuronal cells and brain cells will be reviewed in more detail in the context of equipping micro- or nanomachines with a muscle and brains in Chapters 8 and 9, this volume, respectively.

Hydrophobic and Hydrophilic Forces Drive Self-Assembly in Nature

Introduction

Hydrophobic forces are most important with regard to self-assembly such as in protein folding, cell wall formation and formation of macromolecular assemblies (e.g., DNA hybridization). The concept of hydrophobic forces only was introduced in the 1950s and is hard to define, as the underlying physical principles are not easily understood.* Jencks's definition is one of the better ones: "hydrophobicity is an interaction between molecules that is stronger than the interaction of the separate molecules with water, and can not be accounted for by covalent, electrostatic, H-bonding or charge transfer processes."

As we saw in Volume I, Chapter 7, a hydrophobic molecule, also termed a nonpolar or lipophilic molecule, is a water-fearing or -repelling molecule. Since hydrophobic molecules are not electrically polarized, and because they are unable to form hydrogen bonds, water repels them in favor of bonding with other water molecules. Examples of hydrophobic molecules include the alkanes, oils, fats, and greasy substances in general. Hydrophilic or polar molecules is the term used for water-loving molecules. Examples include polar organic groups such as amino, carboxyl, and hydroxyl groups, such as those in amines, carboxylic acids, and alcohols. Amphiphilic or amphipathic molecules are molecules that have both a hydrophobic and a hydrophilic component, which makes them at the same time water-loving and water-repelling. They are compounds that exhibit both hydrophilic and hydrophobic properties simultaneously, for example, fatty acids and detergents. One type of an amphiphilic molecule has a water-soluble carboxylic group head and a water-insoluble hydrocarbon tail. Amphiphilic molecules of this sort are called surfactants. A surfactant is a soluble chemical amphiphilic compound that reduces the surface tension between two liquids. It is used in many detergents and soapy cleaning compounds because of its ability to surround grease particles and solubilize them. Another type of amphiphilic molecules, introduced below, is the phospholipids.

Hydrocarbon groups have greater polarizability (not to be confused with polarity) than water, so they prefer to interact with each other rather than with water. But this effect is rather small, so there must be a more dominant effect at work, especially since even if the force between nonpolar molecules is repulsive they are aggregating when put together in a polar medium. The complex that forms between nonpolar molecules has less unfavorable interaction with the solvent than the sum of the uncomplexed molecules (the lesser of two evils so to speak) (Figure 2.39). It is this latter effect that

* "The magnitude, range and origin of the hydrophobic interaction have been a mystery ever since the pioneering work by Kauzman and Tanford" (J. Israelachvili, 2005).[53]

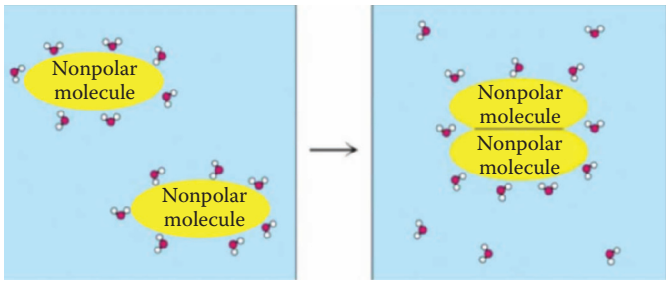


FIGURE 2.39 Aggregation reduces the hydrophobic surface area that requires ordering of water molecules for optimized solvation.

causes the hydrophobic interaction—which in itself is thus incorrectly named since the energetic driving force comes from the hydrophilic molecules.

The hydrophobic effect is illustrated in more detail in Figure 2.40 and in the sequence of events in Figure 2.41. Both figures clarify what happens when we dissolve lipid molecules in water. In Figure 2.40, we illustrate how hydrophobic compounds upset the structure of water and lead to the formation of a clathrate. A clathrate, clathrate compound, or cage compound is a chemical substance consisting of a lattice of one type of molecule trapping and containing a second type of molecule. The ordering of the shell of water molecules around the hydrocarbon solute in a clathrate compound (so as to minimize "dangling" H-bonds) causes a significant decrease in the entropy of the water ($\Delta S < 0$). Typically, the total change in entropy, ΔS, in dissolving small hydrocarbon molecules in water (at 298 K) is about –100 J K^{-1} mol^{-1}. This decrease in entropy (associated with the ordering of molecules) makes it unfavorable to mix "hydrophobic molecules" in water, or solvent entropy favors the removal of nonpolar solutes from solution. The overall enthalpy change of interaction of a nonpolar solute with water, on the other hand,

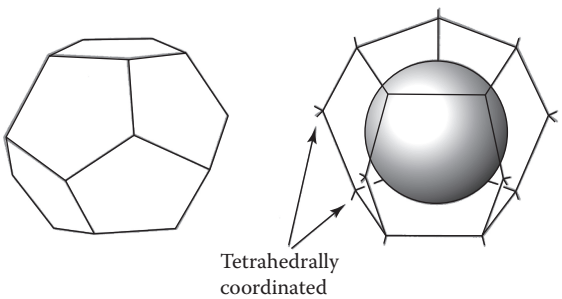


FIGURE 2.40 A water "cage" around another molecule or a clathrate.

is not particularly unfavorable ($\Delta H < 0$) because the nonpolar molecules induce cage-like ordering of the first shell of water molecules (clathrate), strengthening their H-bonding. Comparing the numbers, it is clear that the origin of the hydrophobic effect is mostly entropic.

The latter is also clear from the following example. When a hydrophobic molecule such as n-butane is added to water at 25°C, the Gibbs free energy changes as follows.

$$\Delta G = \Delta H - T\Delta S = -4.3 + 28.7 = +24.5 \text{ kJ/mol} \quad (2.7)$$

Generalizing, ΔH is related to bond formation/breaking, and ΔS is related to configurational freedom and water ordering. For a favorable reaction, we want the equilibrium constant to be $K > 1$, $\Delta G < 0$, $\Delta H < 0$, and $\Delta S > 0$, where $K = \exp(-\Delta G^{\circ}/RT)$ with $R = 8.3145$ $JK^{-1}mol^{-1}$. In the case of ΔH – negative, we have bond-making gain, and with ΔS – negative we lose entropy because of increased ordering. It is thus enthalpically favorable to place n-butane in water, but it is unfavorable from an entropic perspective (increasing order).

The sequence of events depicted in Figure 2.41 illustrates what happens when amphiphilic or amphipathic molecules are placed in water to form a micelle. A micelle is an aggregate of amphipathic molecules in water, with the nonpolar portions in the interior and the polar portions at the exterior surface, exposed to water. In Figure 2.41a, the water molecules around one amphiphilic molecule reduce the entropy of the water through ordering (forming of a clathrate). In Figure 2.41b, we show a dispersion of lipid molecules, where each molecule forces surrounding water molecules to become highly ordered. In Figure 2.41c, only lipids at the edge of the cluster force the ordering of water. Fewer water molecules are ordered, and the entropy is thus increased. Finally, in Figure 2.41d, all hydrophobic groups are sequestered from water; the ordered shell of water is minimized, the entropy is further increased, and a micelle is formed.

The two immiscible phases (hydrophilic and hydrophobic) change so that their corresponding interfacial area is minimized in the micelle construct. This effect constitutes a phase separation. Depending on their concentration and ionic strength in solution,

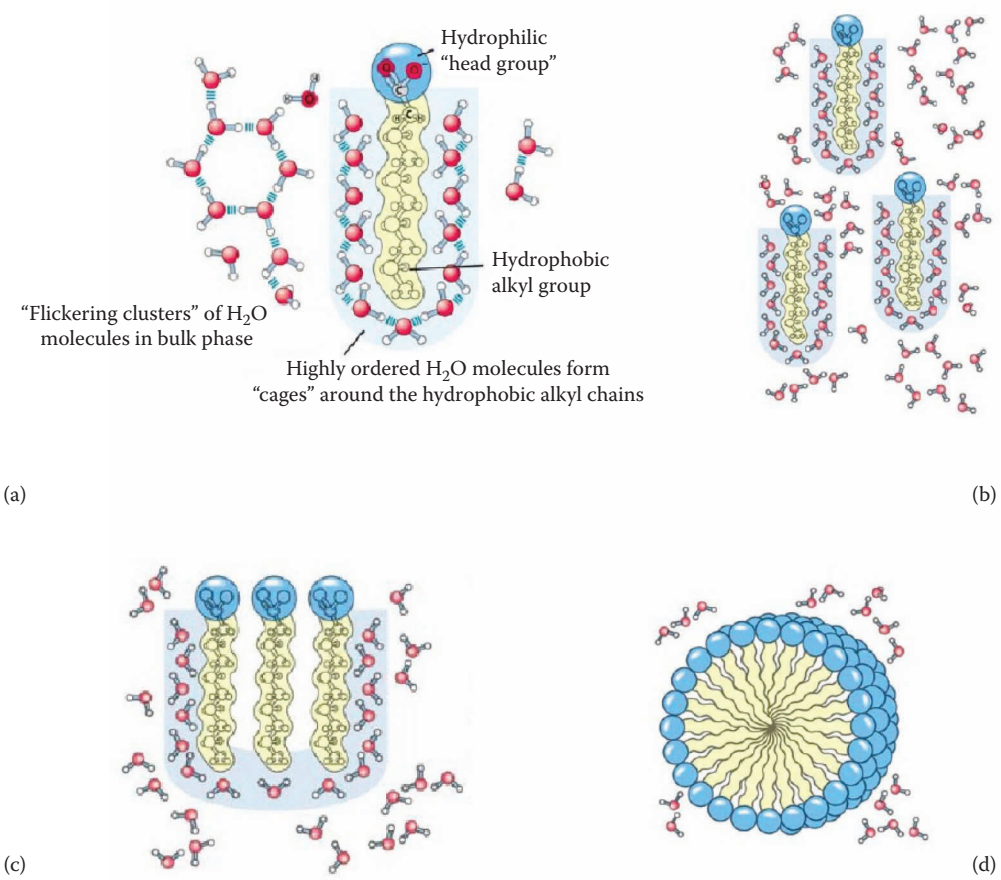


FIGURE 2.41 Steps involved in making a micelle. See text for explanation. (a) The water molecules around one amphiphilic molecule reduce the entropy of the water through ordering (forming of a clathrate). (b) Dispersion of lipid molecules, where each molecule forces surrounding water molecules to become highly ordered. (c) Only lipids at the edge of the cluster force the ordering of water. Fewer water molecules are ordered, and the entropy is thus increased. (d) All hydrophobic groups are sequestered from water; the ordered shell of water is minimized, the entropy is further increased, and a micelle is formed.

the temperature, and other factors, surfactants can assemble into spherical micelles, cylindrical micelles, or bilayers (membranes) or even saddle surfaces in bicontinuous structures (Figure 2.42). A bicontinuous surface, as shown in Figure 2.43, separates the oil phase and water phase. The oil and water are completely separated, but both are continuous across the system. Bilayer sheets are bent into "cubic phases," in which sheets are bent into a complex, ordered network of channels (see also http://www.nationalacademies.org/bpa/reports/bmm/bmm.html).

In a low-dielectric-constant solvent, micelles turn inside out, shielding their hydrophilic heads and forming a reverse micelle. Micelles form spontaneously—i.e., $\Delta G < 0$. When water is added to dried phospholipids, multilamellar vesicles (MLVs) form,

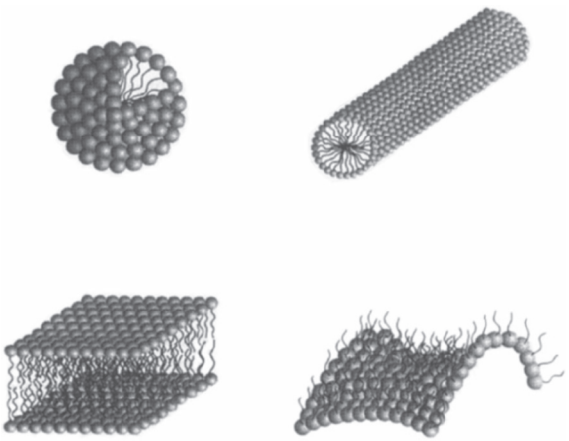


FIGURE 2.42 Surfactants can assemble into spherical micelles, cylindrical micelles, bilayers (membranes), or saddle surfaces in bicontinuous structures. (From I. W. Hamley. 2000. *Introduction to soft matter.* (2nd edition), J. Wiley, Chichester. With permission.)

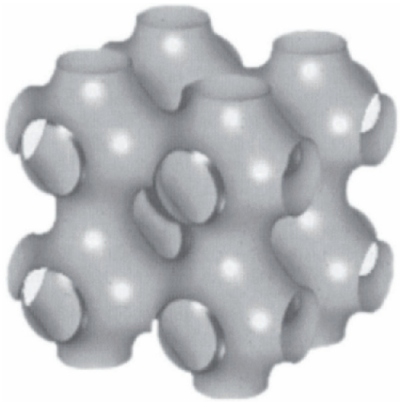


FIGURE 2.43 Surfactants can create a bicontinuous surface to separate an oil phase and a water phase. (From R. A. L. Jones. 2002. *Soft condensed matter.* Oxford University Press, Oxford. With permission.)

with water trapped between the successive polar head groups of the multiple bilayers. Sonicating these produces small unilamellar vesicles (SUVs).

From Micelles to Liposomes and Biological Cells

Introduction

It appears that the chemistry of life cannot take place in anything much simpler than a cell. A virus, a prototype life form, needs to invade a cell and assume control of the cell's chemical machinery by reprogramming it to produce more viruses. For life to have started, it seems essential that the important chemicals were somehow concentrated in order to increase their chance of meeting and reacting. This could have occurred when water evaporated in a small pond, or it could have involved reactions on a clay surface (meeting on a two-dimensional plane is more probable than meeting in three-dimensional space). Another possibility is that the chemicals became grouped in some kind of a vesicle. Thus, they could not diffuse away into the primordial broth. One of the many unanswered questions is, when did this cell-wall precursor take hold, before, during, or after the prebiotic chemical steps?

Micelles, Liposomes, and Biological Cells

Another type of an amphiphilic molecule (besides the detergents encountered above) consists of a phospholipid with an ionic phosphate head group and two hydrocarbon chains attached to it. A glycerol molecule links the two hydrocarbon chains together, and the phosphate group often has other hydrophilic molecules, such as choline and serine, attached to it. When a choline attaches, the molecule is called *phosphatidylcholine,* one of the most abundant phospholipids in all the cell membranes (Figure 2.44A). To form a micelle, water-soluble heads of amphiphilic molecules maintain contact with water, while the water insoluble ends are tucked inside.

A spherical micelle is the most stable configuration of amphipathic lipids that have a conical shape, such as fatty acids (Figure 2.44Ba). A bilayer is the most stable configuration of amphipathic lipids with a cylindrical shape, such as phospholipids (Figure 2.44Bb). Lipid bilayers are back-to-back arrangements of monolayers, and they typically form when there is more than one fatty acid chain connected to the polar head group. Double-chain fatty acids do not pack as well as single-chain fatty acids in a micelle configuration. Unilamellar vesicles, as shown in Figure 2.44Bc, are known as liposomes; multilamellar vesicles containing one or more bilayers arranged around the initial unilamellar vesicle. These types of vesicles are currently used as vehicles to deliver drugs in treatment of various diseases. For example, technologies have been developed to enclose chemotherapy and other drugs in unilamellar liposome bubbles.

The hydrophobic effect is essential for life. It forces the spontaneous clustering of lipids, forming biological membranes. Early in the development of life on Earth, chemicals organized themselves into cells, just like micelles. In humans, they are principally composed of phospholipids, formed from fatty acids, with some specialized proteins embedded in them, as shown in Figure 2.45. A wide variety of amphiphilic molecules can be accepted into a cell membrane. About half of a membrane surface is covered with proteins and sugars performing a wide variety of roles.

The bilayer that constitutes the cell surface—the plasma membrane—is permeable to some molecules, either passive or mediated by transmembrane proteins (pumps), impermeable to others (most water soluble molecules), and very flexible; call it a

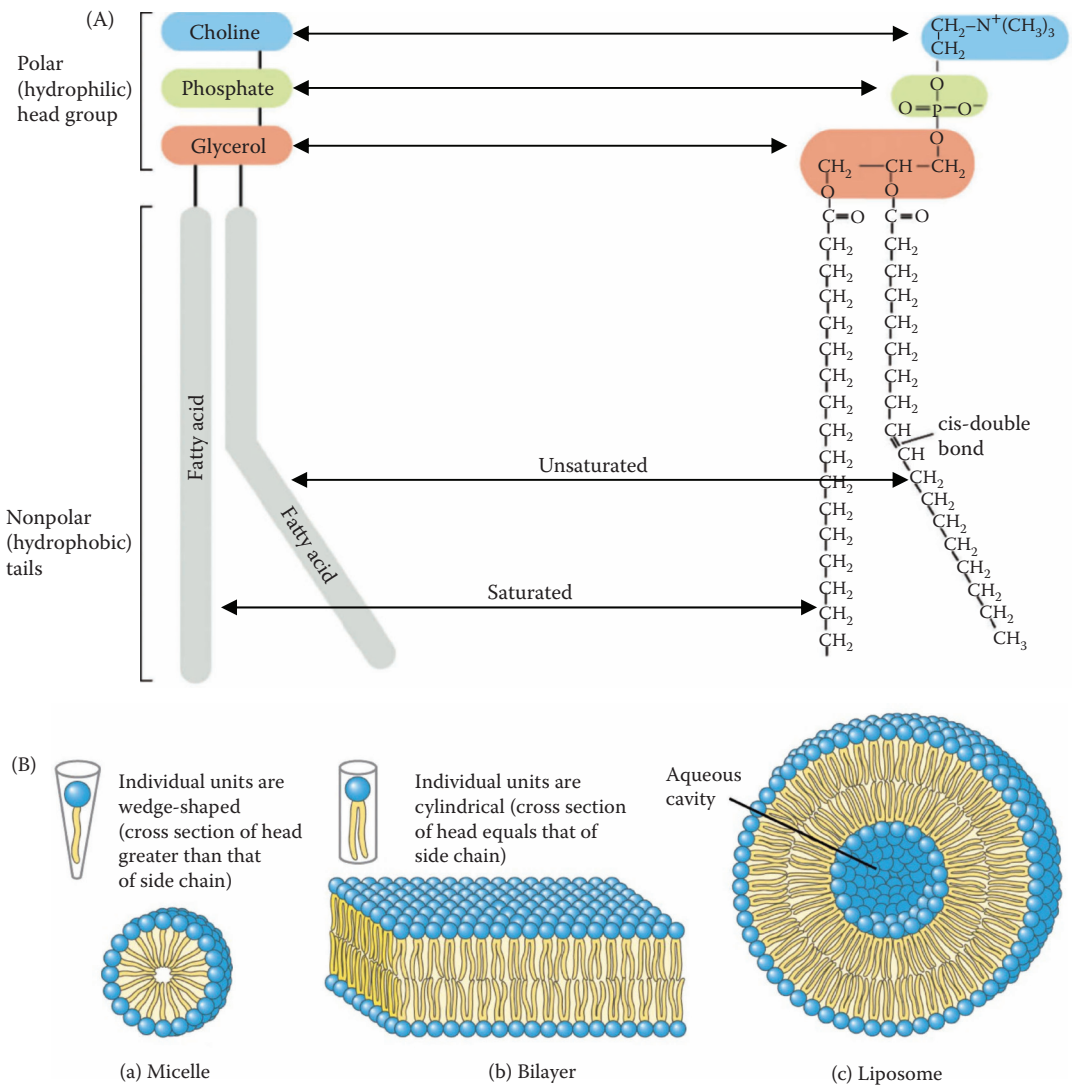


FIGURE 2.44 (A) Example of an amphiphilic molecule (phosphatidylcholine). (B) Amphiphilic molecules forming micelle (a), lipid membranes (b), and unilamellar liposome (c).

dynamic two-dimensional fluid in which lipids and proteins can move laterally. The membrane is also a very good electrical insulator, typically holding 70 mV over 45 Å (i.e., a field strength of 1.6 10^7 V/m), has no corners, and self-assembles.

As widely accepted as the idea of a plasma membrane with embedded proteins as ion channels and ion pumps has become, there are still valid objections to this model, as Gerald Pollack lucidly explains in *Cells, Gels and the Engines of Life.*[54] Pollack presents plenty of evidence that a cell is not enveloped by an ion-impermeant barrier. In this elegantly written book, he explains most cell functions by invoking the gel nature of the cytoplasm.

Natural Biopolymers

Even the most complex self-assembled structures of natural biopolymers, such as nucleic acids and proteins, are all driven by very simple balances of coil entropy and bond enthalpy; the coil entropy wants to expand, but the folded structure is held in place by enthalpic bonds such as hydrogen bonds, ionic bonds, and hydrophobic bonds. Being individually weak, these noncovalent interactions nevertheless can sum to an impressive total, often several hundreds of kilojoules, when many are present within a given macromolecule or between macromolecules. This explains the stability of the macromolecular structures. But being individually

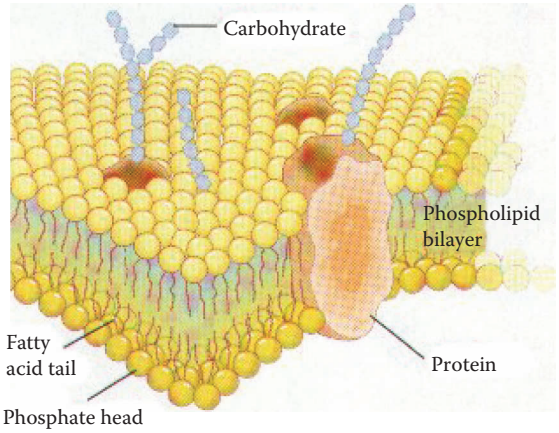


FIGURE 2.45 The cell wall of an animal cell is a phospholipid bilayer with embedded proteins regulating a series of functions, e.g., transport, recognition.

weak, individual noncovalent bonds in macromolecular structures can be broken and reformed rather easily. The latter in turn explains the flexibility in macromolecular structures. All biochemical processes have thus enthalpic and entropic contributions that combine to dictate spontaneity. DNA is a straightforward example of this, with a balance of mostly H-bonds and coil entropy (Figure 2.46). DNA structure arises from a favorable enthalpic term (enthalpy-dominated state) compensating for an unfavorable entropic term ($\Delta H < 0$ and $\Delta S < 0$). At a critical temperature $T_{critical}$, this reverses and entropy wins out. This *melting temperature* can be derived by solving $\Delta G = \Delta H - T\Delta S$, with ΔH from H-bonds balanced by ΔS from the amount of disorder, as follows.

$$T_{critical} = \frac{\Delta H}{\Delta S} \tag{2.8}$$

Because there are three hydrogen bonds between cytosine and guanine and only two between adenine and thymine, a higher C, G content makes for a higher melting temperature. The latter is illustrated in Figure 2.47.

In the 1980s, Nadrian Seeman (http://seemanlab4.chem.nyu.edu) from New York University took advantage of the intrinsic self-assembling properties of DNA molecules, as described here, to construct DNA molecular complexes. By synthesizing branched DNA and exploiting this behavior, Seeman and coworkers managed to assemble increasingly complex structures that do not occur in the natural environment (DNA-mediated assembly is described in Chapter 4, Figure 4.49). The use of stable branched DNA molecules permits one to make stick figures as depicted in Figure 4.49a, where we show a stable branched DNA molecule, and in Figure 4.49b, where we show such a molecule with sticky ends; four of these sticky-ended molecules are shown assembled into a quadrilateral. More exploits from the world of DNA nanotechnology, such as DNA origami, can be found in Chapter 3, this volume, on bottom-up manufacturing.

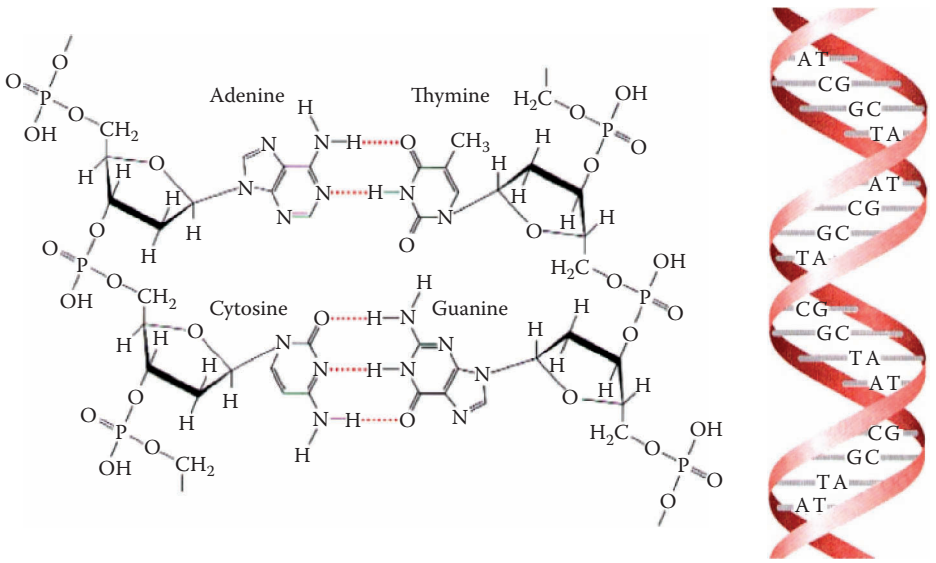


FIGURE 2.46 DNA: a question of entropy of coiling versus enthalpy of the hydrogen bonds (two between adenine and thymine and three between cytosine and guanine).

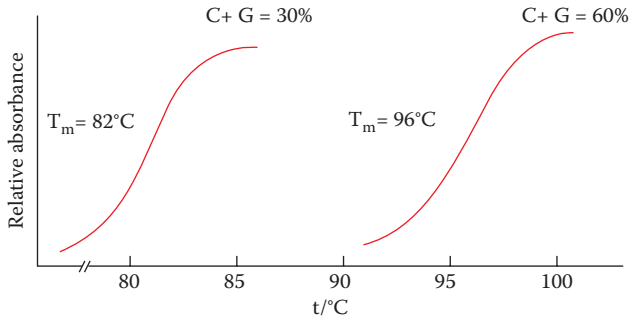


FIGURE 2.47 Melting temperature or critical temperature of DNA is reached at the moment that the entropy wins out. The higher the CG content, the higher the melting temperature.

Proteins are polypeptides prepared from 20 different amino acids, with a few different $\Delta H < 0$ bonds holding them together, to form a vast number of folded structures that govern biological function and biomaterials. As in micelle and DNA assembly, entropy and enthalpy balances are the basis of protein self-assembling behavior, such as the formation of a β-sheet, illustrated in the beautifully intricate "β-barrel" structure in green fluorescent protein (GFP) (see Volume I, Figure 7.108). In self-assembled, native proteins, the enthalpy wins over entropy (*H* dominated) and an increase in the temperature helps the entropy term till we observe denaturing or unfolding, as illustrated in Figure 2.48. Denaturing of a protein is the conversion of a biologically functional molecule into a nonfunctional form. There are many denatured states but only one native state, so entropically, the denatured protein state is by far the most favored condition for the protein to be in. Proteins can regenerate to their native state but only slowly. Denatured proteins have a greater tendency to aggregate. The transition from native to denatured happens at a critical temperature given by Equation 2.8. A typical coiling/ordering ΔS

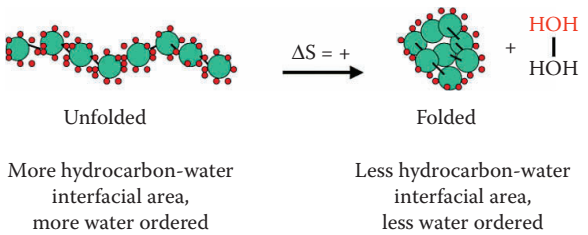


FIGURE 2.48 Protein folding and unfolding (denaturing).

penalty for a complex protein such as GFP may be of the order of 40,000 kJ, the number of H-bonds needed to balance this, at ~50 kJ each, is about 800 (ignoring any hydrophobic bond help). Most of the associative effects are enthalpic in origin, except for the hydrophobic effect, which is essentially entropic in origin.

In the exceptional case of a protein with an entropy dominated balance, such as with pyruvate carboxylase, a decrease in the temperature lowers the entropy term, and we observe denaturation at lower temperatures.

Biotechnology and Biomimetics

Introduction

Today, biotechnology is a major application target for miniaturization science, as engineered structures and nature's building tools have become similar in size, and financial rewards promise to be much larger for these life science ventures than they ever were for mechanical MEMS applications (see Chapter 10). After putting forward a definition for biotechnology and presenting examples of the use of miniaturization science in biotechnology, i.e., bio-MEMS, we make some concluding remarks on biomimetic approaches and finish this chapter by comparing some human and natural engineered structures. As examples of micromachining and nanomachining in biotechnology (bio-MEMS) we survey three examples here: genetically engineered proteins, miniaturized PCR and DNA chips. Other bio-MEMS and bio-NEMS examples, including electrochemical and optical immunosensors, microphysiometers, responsive drug delivery pills, glucose sensors, blood gas and blood electrolyte sensors, lab on a chips (LOCs), etc., can be found throughout the book. This section constitutes an attempt at sketching possible avenues for merging of biotechnology with nanochemistry and nanofabrication.

Biotechnology: Definition

Fermentation processes known from ancient times (the oldest known written recipe for beer brewing is written on a clay cuneiform tablet, part of an epic poem devoted to Ninkasi, the Sumerian goddess

of beer) evolved eventually into a larger endeavor called biotechnology, employing either cells or cell extracts (e.g., enzymes) to produce foods, drugs, insecticides, etc. Generally, modern biotechnology is an umbrella term that covers various techniques, applied mainly to cells or molecules, which take advantage of biological processes in very precise ways. The discoveries of three different enzymes in the late 1960s, especially, heralded its advent. The first of these discoveries was reverse transcriptase enzyme, which converts RNA into the corresponding DNA; the second, an enzyme that cuts DNA at predefined places; and the third, an enzyme that joins separate pieces of DNA. Using these gene technology tools, the pooled content of a cell can thus be used to manufacture proteins that can be used for human consumption (e.g., as medicine). Because the genetic code is universal, almost any cell in any organism can "read" a gene and translate it into a desired protein. Bacteria that have been genetically engineered to carry the human insulin gene, for example, can produce human insulin. Cells might also be reengineered and substituted in living things to correct for inherited deficiency or performance. "Gene therapy" has made great progress, especially in plant engineering.

Genetically Engineered Proteins

Binding chemistry of a target to a specific protein at the nanoscale translates into a mechanical movement/motion of the protein, known as a change in conformation. This conformational change of proteins upon binding with a target molecule has been exploited in the design of very selective and sensitive biosensors. This is accomplished by labeling the protein, say with a fluorescent probe, in such a way that the probe does not hinder the binding of the protein with its target, while the change in signal reported by the probe upon binding is maximized.

Two important examples of genetically engineered proteins: green fluorescent protein (GFP) and calmodulin. GFP, from the molecular system of the bioluminescent jellyfish *Aequorea victoria*, we saw in Volume I, Chapter 7, are making quite an impact on biotechnology today as fluorescent markers. Here we will explore genetic manipulation of the calmodulin protein and its uses.

Through judicious genetic manipulation of a protein such as calmodulin [a 148-amino acid (17-kD) molecule, shown in Figure 2.49], a rationally designed molecular biosensor can thus be developed. The key is to always select the right location for attachment of the probe in the protein structure. Calmodulin is a "regulatory" protein, involved in protein-protein interactions and in signaling pathways that involve stimulating or inhibiting the activities of many enzymes and proteins. Calmodulin is a notably tough molecule: it withstands a very low pHs and even survives boiling water. One of its binding partners is Ca^{2+} ion, which matches the versatility of the protein. Calcium ions are abundant in all biological systems, and its importance derives from its role as a universal messenger ion. The ion conveys signals received at the cell surface to the inside of the cell (the concentration of calcium inside the cell is 1,000- to 10,000-fold less than in the extracellular fluid). Calcium ions are involved in such processes as the regulation of muscle contraction; the secretion of hormones, digestive enzymes, and neurotransmitters; the transport of salt and water across the intestinal lining; and the control of glycogen metabolism in the liver. The binding of Ca^{2+} to multiple sites of a calmodulin molecule induces a major conformational change—a hinge-type motion—that converts it from an inactive to an active form, allowing for interactions with other proteins and peptides and thus regulating the function of

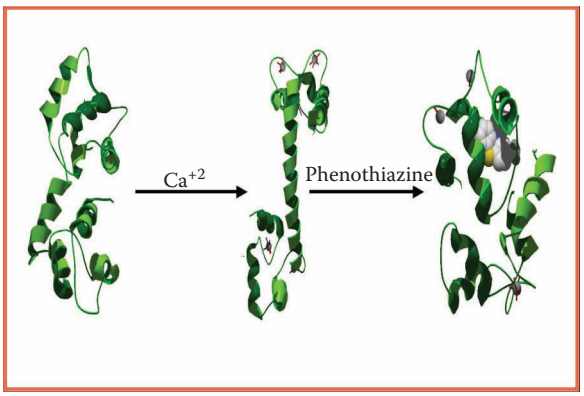


FIGURE 2.49 Model showing the exposure of drug binding site in calmodulin following binding of Ca^{2+}. Upon binding to calcium, calmodulin undergoes a change in conformation, which exposes two hydrophobic pockets located in the N- and C-domains. Certain hydrophobic peptides and the antipsychotic phenothiazine class of drugs interact with these exposed hydrophobic pockets.

the various enzymes and proteins. The calmodulin crystal structure has been well studied using x-ray crystallography and nuclear magnetic resonance (NMR) techniques, and it has been found to consist of two domains, the N- and C-domains.* Two high-affinity calcium-binding sites are located in the C-domain, and the other two calcium-binding sites (the low-affinity sites) are located in the N domain. Upon binding to calcium, calmodulin undergoes a change in conformation to an active form, which exposes two hydrophobic pockets located in the N- and C-domains. Certain hydrophobic peptides and the antipsychotic phenothiazine class of drugs interact with these exposed hydrophobic pockets, and in the presence of calcium ions, calmodulin also becomes a sensor for those peptides and drugs (Figure 2.49). Schauer-Vukasinovic et al.,[55] through site-directed mutagenesis of calmodulin, produced several mutants that varied from one another by a single cysteine residue at various positions in the protein skeleton. N-[2-(1-maleimidyl) ethyl]- 7-(diethylamino)coumarin-3-carboxamide (MDCC) (Invitrogen, CA, USA-prod #D-10253), a commercially available, environmentally sensitive, thiol-reactive fluorophore label was attached to the cysteine residue in each of the mutants, and the change in fluorescence signal upon binding the calcium ions was measured. Mutation sites were chosen on the basis of both the crystal and NMR structures of free calmodulin and calmodulin complexed with calcium. The MDCC-labeled calmodulin mutant showed a change in the fluorescence signal in the presence of calcium ions, the magnitude of which was dependent on the concentration of calcium.[55] By comparing the enhancement of the fluorescent response of the different mutants, a Ca^{2+} sensor with a detection limit as low as 2×10^{-9} M was built. The biggest change was noted for the mutant with a single cysteine at amino acid residue 109.[55] This sensing system is highly selective over other metals, including magnesium, europium, and barium. A calibration plot for calcium obtained with the MDCC-labeled calmodulin (CaM) mutant is shown in Figure 2.50.

* The N terminus of the protein denotes the first amino acid of its sequence, while the C terminus corresponds to the last one.

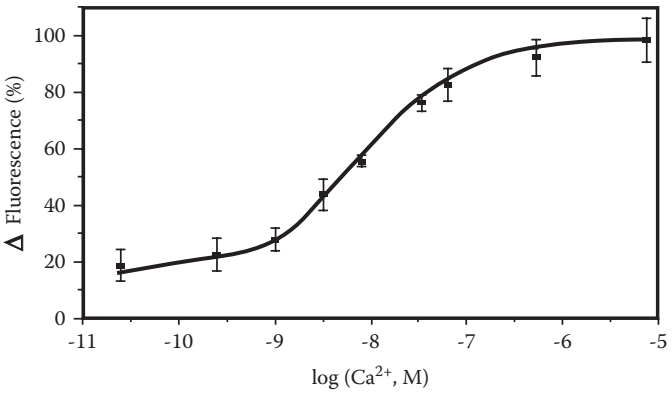


FIGURE 2.50 Calibration plot for calcium obtained with the MDCC [N-[2-(1-maleimidyl)ethyl]- 7-(diethylamino)coumarin-3-carboxamide]-labeled calmodulin (CaM) mutant. The *x*-axis denotes the concentrations of free calcium.[56]

In another study by the same team, when phenothiazine was added to calcium-bound fluorophore-labeled CaM109, a maximum of 100% quenching in the fluorescence was observed. This team then developed an compact disc based fluidic platform (see also Volume I, Chapter 6 on CD fluidics) for the sensing of phenothiazine-class of drugs based on this drug-dependent change in fluorescence signal.[56] This biosensing system also has been proposed as a model for the high-throughput screening of drugs that selectively interact with binding proteins.[57]

Polymerase Chain Reaction

The polymerase chain reaction (PCR) is to genes what Gutenberg's printing press was to the written word, but it took the 1995 O. J. Simpson trial to make PCR a household word. PCR perhaps is the most important tool of biotechnology today; it was conceived by Kary Mullis in 1983 (Figure 2.51). Mullis was granted a patent in 1989, and he received the 1993 Nobel Prize in Chemistry for his invention.[58] The invention of PCR, amplifying the amount of DNA in a sample, is as important for the biotechnology field as the transistor, amplifying electrons, is for electronics.

Mullis, a maverick, unbending to authority, never got the financial rewards or academic credit for his invention, a victim, to some degree, of commercial greed and academic jealousy. Mullis' technique enables the replication of any stretch of DNA as long as the arrangement of nucleotides at the two ends is known. Through PCR, a desired DNA sequence

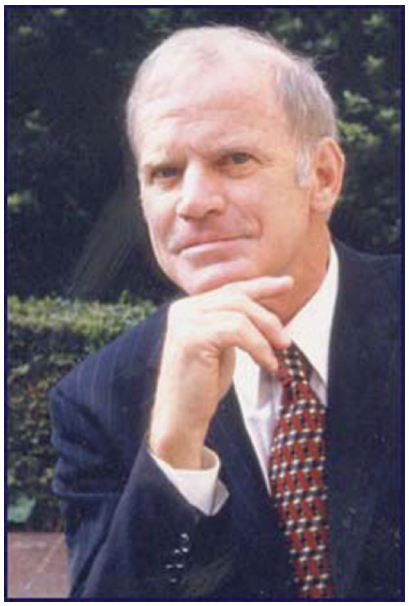


FIGURE 2.51 Kary Mullis, inventor of polymerase chain reaction.

of viral, bacterial, plant, or human origin can be amplified hundreds of millions of times in a matter of hours. By amplifying minute amounts of DNA to analyzable quantities and thus circumventing the need for lengthy cell-culture methods, the technique has had major impact on clinical medicine, genetic disease diagnostics, forensic science, and evolutionary biology. In addition, the ability to isolate genes from organisms by essentially copying them directly from their genome has allowed for the rapid cloning of a myriad of genes whose function was unknown. This has also benefited the fields of genomics and proteomics and thus had an impact on not only science but medicine and biotechnology as well.

Through miniaturization, PCR is becoming faster and applicable to ever smaller samples. The original PCR process is based on the heat-stable DNA polymerase enzyme from *Thermus aquaticus* (*Taq*) (see Figure 2.52), a thermophilic bacterium that can withstand high temperatures and that can synthesize a complementary strand to a given DNA strand in a mixture containing the four DNA bases and two primer DNA fragments (each about 20 bases long) flanking the target sequence.

The mixture is heated to separate (denature or melt) strands of double-helix DNA containing the target sequence and then cooled to allow 1) the primers to find and bind (anneal) to their complementary sequences on the separate strands, and 2) the *Taq* polymerase to extend the primers into new complementary strands. The *Taq* polymerase needs about 1 min to synthesize 1 kb pair, so the synthesis time depends on the length of the product. Repeated heating and cooling cycles multiply the target DNA exponentially, since each new double strand separates to become two templates for further synthesis. In Figure 2.53a, we illustrate how each PCR cycle doubles the amount of DNA.

A typical temperature profile (Figure 2.53b) for PCR is as follows:

1. Denature at 93°C for 15–30 s
2. Anneal primer at 55°C for 15–30 s
3. Extend primers at 72°C for 30–60 s

The primer extension step has to be increased by roughly 60 s/kb to generate products longer than a few hundred bases. The above are typical instrument times; in fact, the denaturing and annealing steps occur almost instantly, but temperature rates in commercial instruments (Figure 2.53c) usually are less than 1°C/s when metal blocks or water are used for thermal equilibration and samples are contained in plastic microcentrifuge tubes. Instantaneous temperature changes are not possible because of sample,

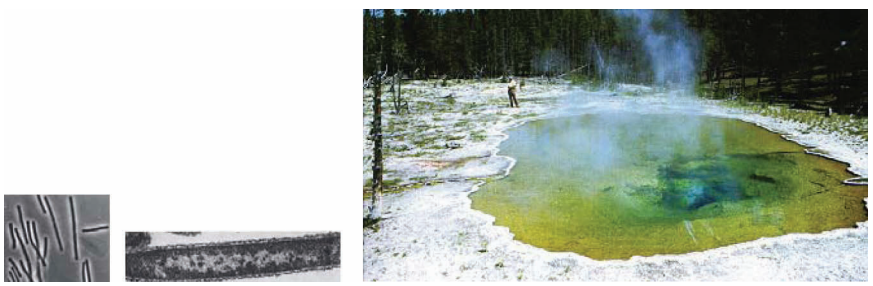
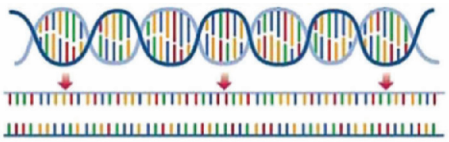
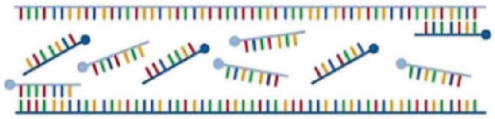


FIGURE 2.52 The bacterium *Thermus aquaticus* was first discovered in several springs in the Great Fountain area of the Lower Geyser Basin at Yellowstone National Park.

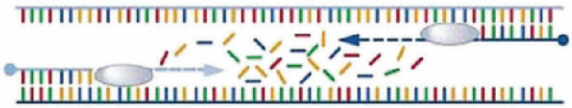
(a)



Step 1 – Denaturation (optimal temperature is 94°C). By heating the DNA, the double strand melts and opens to single-stranded DNA.

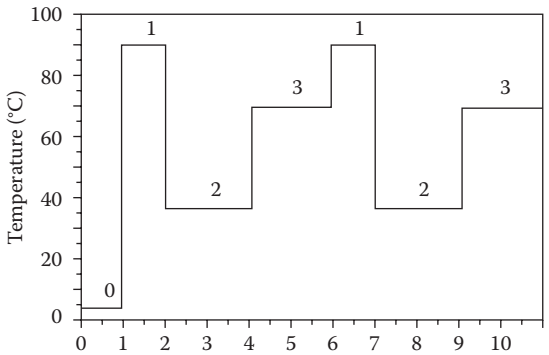


Step 2 – Annealing (optimal temperature is 60°C). The single-stranded primers bind to their complementary single-stranded bases on the denaturated DNA.



Step 3 – Extension: 72°C is the ideal temperature for the *Taq* polymerase to attach and start copying the template. The result is two new helixes in place of the first.

Step 4 – By applying this cycle several times, the quantity of DNA obtained is quickly enough to perform any analysis. Starting with one DNA molecule, after just 20 cycles there will be a million copies, and after 30 cycles there will be a billion copies.



(b)

(c)

FIGURE 2.53 Polymerase chain reaction (PCR). (a) How PCR cycle doubles DNA. (b) A typical PCR temperature profile (1, denature; 2, anneal primers; 3, extend primers; see text). (c) Commercial PCR instrument from Cepheid: the SmartCycler II System.

container, and cycler heat capacities, and extended amplification times of 2–6 h result. During the periods where sample temperature is making a transition from one temperature to another, extraneous, undesirable reactions occur that consume important reagents and create unwanted and interfering compounds.[59] Micromachining has improved PCR in several respects.

By micromachining thermally isolated, low-thermal-mass PCR chambers, one is able to mass-produce a faster, less expensive, more energy efficient, and more specific PCR instrument. To appreciate better the fast thermal cycling of a smaller heater, see Chapter 8 on miniaturization of thermal elements. Given the PCR chambers' small size, smaller amounts of sample and expensive reagents are required. Moreover, rapid transitions from one temperature to another ensure that the sample spends a minimum of time at undesirable intermediate temperatures, so that the amplified DNA has optimum fidelity and purity.[59] Northrup et al.[60] fashioned Si microchambers with integrated

thermally isolated heater elements and found that thermal cycling resulted in amplification without the need for active cooling. This is a direct result of the more effective cooling of these miniaturized Si structures (see Chapter 8). Northrup et al.[60] found that they could heat the PCR reactants as quickly as 35°C/s. Further work showed that integrated heaters on both sides of the reactant volume are required for optimal temperature uniformity. Later, it became evident that to make an economic miniaturized PCR instrument for diagnostic applications, the heater/cooler should be made part of a fixed reader instrument to enable a less expensive disposable unit. Along this line, Wilding et al.[61] performed a less integrated experiment with Si/glass microchambers of 4.4 to 8.9 μL inserted in a commercially available heater/cooler structure. The PCR total cycling time was reduced, but specificity was found to be inferior to that of commercial cyclers. This inferior performance may be linked to the fact that Si is poisonous to the *Taq* polymerase enzyme. The specificity problem can be avoided by derivatizing the Si surface or by machining PCR chambers from a more enzyme-compatible material. Another glaring problem with the two described Si-micromachined PCR approaches is that they do not provide an easy way to apply and remove the DNA sample. This is a major problem with miniaturized PCR and fluidic microsystems in general: the fabrication of connectors, receptacles, and packages that link the microstructure to the macroworld is still in its infancy (see Chapter 4). In principle, to avoid a hybrid connecting/packaging approach, this requires a manufacturing method with a very wide dynamic range, i.e., from microns to several millimeters or even centimeters. Also, for this reason using Si is not necessarily the best approach. The next three PCR examples illustrate nonsilicon alternatives.

Wittwer and Garling[62] and Wittwer et al.[63] optimized PCR temperature and cycle times, using a thermal cycler based on recirculating hot air and 10-μL samples in thin glass capillary tubes, with the sample temperature monitored by a miniature thermocouple. They found that product specificity could be improved significantly while decreasing the required amplification time by an order of magnitude. For a 536-base pair β-globin fragment from human DNA, optimal denaturation at 92–94°C took place in less than 1 s. Annealing for 1 s or less at 54–56°C gave the best product specificity and yield. Only elongation, at 75–79°C, required longer times. Denaturation yield actually decreased with times greater than 30 s, and nonspecific amplification was minimized with a rapid denaturation to annealing temperature transition. Wittwer's device has been commercialized by Idaho Technology; the RapidCycler from this company can perform 30 PCR cycles in less than 10 min.

The Si PCR work of Northrup et al.[60] eventually led to a commercial portable PCR instrument, i.e., the SmartCycler system (http://www.smartcycler.com) from Cepheid (see Figure 2.53c). This instrument incorporates a nonsilicon, rather large but thin PCR chamber with high thermal conductivity ceramic heater plates and integrated temperature sensors. At Cepheid it was understood early on that a major negative consequence of miniaturizing analysis equipment, in extremely small sample sizes, presents numerous chemicals and organisms that are present at levels below the typical detection limits. The topic of scaling of sample size was reviewed in Volume I, Chapter 6, on fluidics. There we saw that biological chemicals associated with clinical chemistry assays (between 10^{14} and 10^{20} copies/mL) and immunoassays (between 10^{7} and 10^{18} copies/mL) are readily assayed with very small sample volumes, in the range of picoliters and nanoliters. However, as evident from Figure 10.49, the minimum sample volume required for accurate DNA assays (from 10^{7} copies/mL to fewer than 100 copies/mL) is 100 μL, but a volume several milliliters of sample is more typical. As a consequence, an important Cepheid focus is the design of microfluidic diagnostic systems with sample preparation of larger input volumes. Along this line, the company fabricated a DNA capture device in which a continuous flow of biological sample results in a gradual accumulation of nucleic acid on the large silicon dioxide surface area of the chip.[59] It is not clear to this author why the latter device should be in Si, as there are many very large surface, porous materials available that are much less expensive.

DNA Chips

Introduction

With the rough draft of the Human Genome Project completed, the next most important question to solve is how genes relate to human health. Why do some people get certain diseases while others remain disease-free? A vast amount of expanding data supports the fact that it is rare for a single defect in one location on one gene to be responsible for a specific disease. The trend is toward the study of the biochemical and molecular functioning of a set of genes and how this complicated interaction leads to a diseased state. One way to tackle this study, we saw in Volume II, Chapter 8, is by using DNA microarrays. A DNA microarray, or DNA chip, consists of DNA molecules ranging from 20 to 1000 or more base pairs concentrated into specific areas on a solid support such as a glass chip.[64] The arrays are used as a capturing device of specific complementary gene products. Thousands of the same oligonucleotides can be attached in a specific location so that gene expression may be observed by counting the amount of oligos that bind. In this way, specific gene expression can be observed in particular cells, during specific stages of development, in diseased versus normal individuals, after administration of a drug, or in malignant cells, for example (see Volume II, Figure 8.23). Alternatively, one can incorporate many different oligonucleotides representing different genes on the same chip so that screening for the expression of many genes simultaneously is possible. The uses of this technology are broad and include disease diagnosis, individual response to drug therapy, and understanding the difference of gene expression between healthy and normal individuals.

Almost all current genetic testing either uses sequencing or hybridization. Sequencing involves hydrolyzing (digesting) a piece of DNA under controlled conditions, derivatizing the pieces as they come off the DNA strands, and analyzing them by chromatography or electrophoresis. In hybridization, pieces of genes bind to other pieces containing complementary sequences. Of the two techniques, sequencing is the more comprehensive and versatile, but hybridization is faster and, for small pieces of genes, more efficient. Rapid identification of genes or gene sequencing has become one of the main aims of medical diagnostics.

The field of DNA arrays involves microfluidics, MEMS machining, and self-assembly of DNA probes, which puts it at the forefront of mixing top-down and bottom-up manufacturing techniques. For the various microfabrication methodologies of DNA chips see Volume II, Chapter 8. In what follows we briefly review the major applications of microfabricated DNA chips.

Applications of DNA Arrays

In a typical cDNA array experiment, the genes of interest are first amplified using PCR (see above). After purification, the PCR products are printed on microscope glass slides by using a robot (see Volume II, Figure 8.22). For example, a DNA microarray, containing all of the genes in yeast (6400 of them) can be used to look at specific changes of gene expression in metabolism. Using in vitro reverse transcription, both the test and reference total RNA are labeled using two different fluorescent dyes. Laser excitation causes a specific emission that generates fingerprint-like spectra, which are measured using a scanning confocal laser microscope. The data are basically viewed as a change in the color of the two fluorescent readings at a specific point on the slide (a gene) when compared with a reference sample.

In an alternative DNA array approach, Fodor et al.[65] at Affymax (http://www.affymax.com) developed an elegant method combining on-chip combinatorial chemistry with established hybridization techniques and advanced fluorescence detection to make arrays of oligonucleotides (see Volume II, Figure 8.16). After synthesizing oligonucleotide arrays onto a glass slide, the glass slide is incubated with fluorescent-labeled DNA probes. After the DNA probes are hybridized to the oligonucleotide array, the complementary sequences are determined by fluorescent scanning. The oligonucleotide array format is used to detect mutations, locate target regions within genes, investigate gene expression, and identify gene function.[65] In this approach, the chemical synthesis and analysis of large numbers

of potentially useful compounds can be carried out on a single small chip (say, a centimeter square); therefore, it becomes orders of magnitude quicker and cheaper to identify potential ligands to build new drugs or analyze DNA sequences.[66] A key advantage of this approach is that it enables chips to be manufactured directly from sequence databases, thereby removing the uncertain and burdensome aspects of sample handling and tracking. Another advantage of this technology is that the use of synthetic reagents minimizes chip to chip variation by ensuring a high degree of precision in each coupling cycle. One disadvantage of this approach, however, is the need for photomasks, which are expensive and time-consuming to design and build. In Volume II, Chapter 8, we described how the more recent maskless array synthesizer might get around the latter obstacle (Figure 8.19). Also, the hybridization can occur only through DNA fragments finding each other through diffusion, which is slow.[67] The Nanogen DNA chip described in Example 8.2 (Volume II, Chapter 8) is much faster, as DNA fragments reach the chip hybridization sites through electrophoresis (Nanogen, a San Diego–based company, went out of business in early 2010). A more detailed list of pros and cons of the various DNA array approaches is presented in Volume II, Chapter 8. The integration of DNA arrays within a sample preparation platform (e.g., a microfluidic CD) was discussed in Volume I, Chapter 6, on fluidics.

Biomimetics: Some Concluding Remarks

Early examples of successful biomimetics include artificial materials for implants, tissue engineering, an artificial retina based on the architecture of the human eye, and neural network software based on the human brain. The devices mentioned are relatively large, however, and are fabricated using top-down methodologies. The neural network software example is only a very crude approximation of a human brain. Human-made devices crafted with bottom-up methods will open up a much more rewarding paradigm in human manufacturing; nanochemistry holds the promise of the versatility of design offered by nature itself, and molecular self-assembly and replication add to the tremendous appeal of this type of nanotechnology. Obviously, there is something very attractive about the small size of nature's preferred building blocks such as proteins. In biological systems, the energy efficiency is approximately proportional to the 2/3 power of the linear dimension. This is because metabolism is proportional to the second power of the linear dimension (surface of the organism, 1^2), and energy uptake (feeding) is proportional to the body volume (1^3), so the smaller organism is, the higher its efficiency (see the discussions of *scaling laws* in Chapter 7). Nature provides excellent examples in the design of efficient micro- and nanosystems as it optimizes scaling laws in the microdomain and even exploits the quantum-size effects of its components. These biosystems have been tested and selected by eons of evolution, and at a time when we find that micromachining tools are limiting the fabrication of ever smaller devices, exploration of some of nature's tools and building blocks promises to be very rewarding. As one of the many motivating examples of the promise of biomimetics, consider enzymes. There is a plethora of enzymes that outperform synthetic catalysts by several orders of magnitude. For instance, there is no artificial catalyst for producing ammonia from its elements at ambient temperature and pressure, as the nitrogenase of nodule-forming bacteria does (the industrial Haber-Bosch synthesis of ammonia from nitrogen and hydrogen at high temperatures and pressures pales in comparison).

Most miniaturization, especially in integrated circuits (ICs), over the next ten years will be based on top-down nanofabrication, and Moore's law will continue to run its course. In parallel, we will continue to adapt bottom-up nanochemistry principles and combine them with nanofabrication. There is a significant gap between the scale of individual molecular structures of nanochemistry and the sub-micrometer structures of nanofabrication. It is exactly in that gap, from about one nanometer to several hundred nanometers, where fundamental materials properties are defined. It is also on the nanometer scale that quantum effects become significant.

There is little doubt that biotechnology analysis tools will keep on improving at an increasingly

faster rate. Desktop DNA sequencers and 3D protein readers will be a reality in the not-so-distant future. Genetic engineering will have the most profound impact, though, on how humankind looks at manufacturing. Since it is already possible to synthesize a virus bottom-up, given the sequence of the bases in its genes, it seems quite likely that we will be able to manufacture synthetic viruses designed to enter a cell and carry out diagnostic and therapeutic tasks. Venter, we saw earlier, is using bottom-up synthesis methods to create a minimal gene setup required to sustain life in a test tube. Photovoltaic solar cells today have a high conversion efficiency of about 10–15% but are expensive to deploy and maintain. Crops grown for energy are also expensive, involve harvesting, and are only 1% efficient. It is conceivable that genetic engineering will enable the production of energy crops that convert sunlight into fuel at a 10% efficiency.[68] Through nanochemistry, the current digital information technology revolution might well be followed by a new analog manufacturing revolution. Today, computers let us shape our digital environment, but by giving computers the means to manipulate the analog world of atoms as easily as they manipulate bits, the same kind of personalization may be brought to our physical three-dimensional environment. In this context Gershenfeld from the Massachusetts Institute of Technology Media Laboratory envisions a personal fabricator akin to the personal computer.[69]

A human society based on nanomachining could perhaps be a much more balanced one, with a manufacturing approach based on how the species itself is made. Products will be based on a fundamental understanding of the assembly of their ultimate components, atoms, molecules, and proteins, and on how to induce self-assembly into useful objects. Materials will be degradable, flexible, and fully reusable. The smaller building blocks used in manufacture will enable products of more variety and intelligence. There will be less emphasis on the traditional engineering materials, such as steel, wood, stone, composites, and carbon, and proteins will become much more important.

The transition toward a nanosociety will require a major shift in workforce skill level as manipulation of data and applying knowledge of bioengineering will be part of a manufacturing worker's daily duties. In academia, less hyperspecialization and better grounding in all the sciences and engineering will be a must as the traditional "dry" engineering and sciences, such as electrical engineering and mechanical engineering, will merge with the "wet" sciences and engineering, such as biology and bioengineering.

Appendix 2A: Uses of Extremophiles

Source	Uses
Hyperthermophiles	
DNA polymerases	DNA amplification by PCR
Alkaline phosphatase	Diagnostics
Proteases and lipases	Dairy products
Lipases, pullulanases, and proteases	Detergents
Proteases	Baking, brewing, and amino acid production from keratin
Amylases, α-glucosidase, pullulanase, and xylose/glucose isomerases	Baking, brewing, and amino acid production from keratin
Alcohol dehydrogenase	Chemical synthesis
Xylanases	Paper bleaching
Lenthionin	Pharmaceutical
S-layer proteins and lipids	Molecular sieves
Oil degrading microorganisms	Surfactants for oil recovery
Sulfur oxidizing microorganisms	Bioleaching, coal and waste gas desulfurization
Hyperthermophilic consortia	Waste treatment and methane production

(*Continued*)

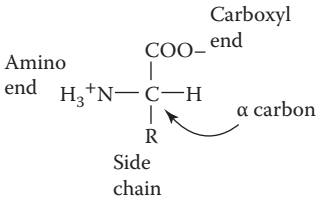
Appendix 2A: (*Continued*)

Source	Uses
Psychrophiles	
Alkaline phosphatase	Molecular biology
Proteases, lipases, cellulases, and amylases	Detergents
Lipases and proteases	Cheese manufacture and dairy production
Proteases	Contact-lens cleaning solutions, meat tenderizing
Polyunsaturated fatty acids	Food additives, dietary supplements
Various enzymes	Modifying flavors
ß-Galactosidase	Lactose hydrolysis in milk products
Ice nucleating proteins	Artificial snow, ice cream, other freezing applications in the food industry
Ice minus microorganisms	Frost protectants for sensitive plants
Various enzymes (e.g., dehydrogenases)	Biotransformations
Various enzymes (e.g., oxidases)	Bioremediation, environmental biosensors
Methanogens	Methane production
Halophiles	
Bacteriorhodopsin	Optical switches and photocurrent generators in bioelectronics
Polyhydroxyalkanoates	Medical plastics
Rheological polymers	Oil recovery
Eukaryotic homologs (e.g., *myc* oncogene product)	Cancer detection, screening antitumor drugs
Lipids	Liposomes for drug delivery and cosmetic packaging
Lipids	Heating oil
Compatible solutes	Protein and cell protectants in variety of industrial uses, e.g., freezing, heating
Various enzymes, e.g., nucleases, amylases, proteases	Various industrial uses, e.g., flavoring agents
γ-Linoleic acid, ß-carotene, and cell extracts (e.g., *Spirulina* and *Dunaliella*)	Health foods, dietary supplements, food coloring, and feedstock
Microorganisms	Fermenting fish sauces and modifying food textures and flavors
Microorganisms	Waste transformation and degradation, e.g., hypersaline waste brines contaminated with a wide range of organics
Membranes	Surfactants for pharmaceuticals
Alkaliphiles	
Proteases, cellulases, xylanases, lipases, and pullulanases	Detergents
Proteases	Gelatin removal on x-ray film
Elastases, keratinases	Hide dehairing
Cyclodextrins	Foodstuffs, chemicals, and pharmaceuticals
Xylanases and proteases	Pulp bleaching
Pectinases	Fine papers, waste treatment, and degumming
Alkaliphilic halophiles	Oil recovery
Various microorganisms	Antibiotics
Acidophiles	
Sulfur-oxidizing microorganisms	Recovery of metals and desulfurication of coal
Microorganisms	Organic acids and solvents

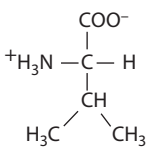
Appendix 2B: The Twenty Amino Acids Found in Biological Systems

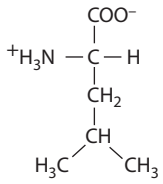
Alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine.

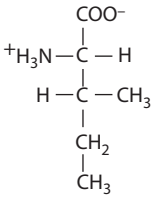
All amino acids have the same general formula:

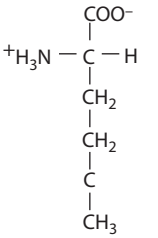


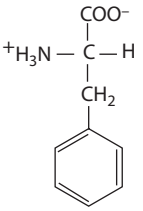
Amino acids with hydrophobic side groups


Valine (val)

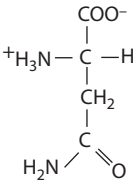

Leucine (leu)

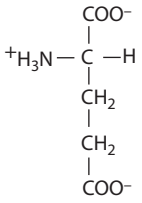

Isoleucine (Ile)

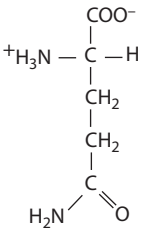

Methionine (met)

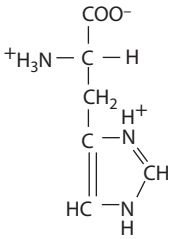

Phenylanine (phe)

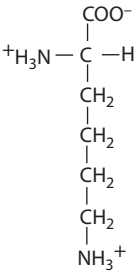
Amino acids with hydrophilic side groups

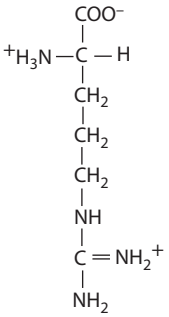

Asparagine (asn)

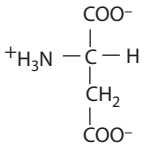

Glutamic acid (glu)


Glutamine (gln)

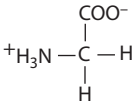

Histidine (his)

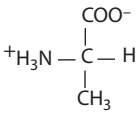

Lysine (lys)

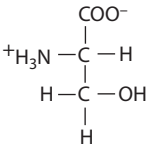

Arginine (arg)

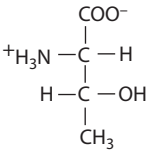

Aspartic acid (asp)

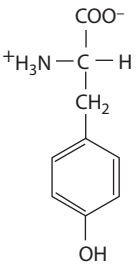
Amino acids that are in between

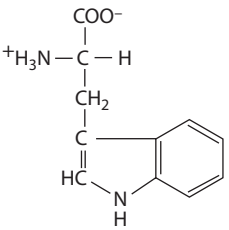

Glycine (gly)

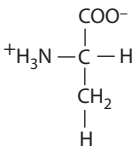

Alanine (ala)

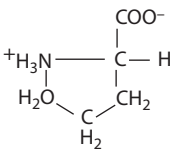

Serine (ser)


Threonine (thr)

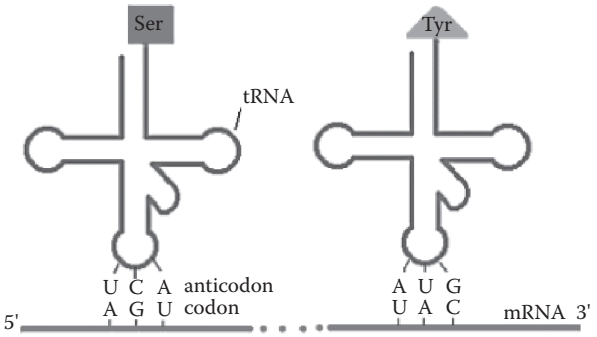

Tyrosine (tyr)


Tryptophan (trp)


Cysteine (cys)


Proline (pro)

Appendix 2C: The Universal Code



2nd base in codon

1st base in codon	U		A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

FIGURE A2C.1 The genetic code.

Questions

2.1: Immunoassays are based on the specific binding of antibodies with antigens.
- (a) Describe an ELISA.
- (b) Explain how nonspecific binding can produce unwanted signals in both immunoassays and in DNA assays. (*Nonspecific binding* refers to the binding of adsorption of molecules onto the surface without any preference or selectivity.)

2.2: With the aid of diagrams, describe:
- (a) A competitive immunoassay
- (b) A sandwich immunoassay

2.3: Why are living systems, at first glance, violating the second law of thermodynamics?

2.4: Briefly discuss the "who was on first" dilemma: DNA, RNA, or proteins?

2.5: What models are complexity and communication sciences contributing to the biogenesis question?

2.6: Draw a diagram to illustrate how proteins are manufactured.

2.7: List examples of emergent properties of complex systems.

2.8: Find some examples of man-made materials that feature self-repair.

2.9: Which of the following statements regarding GFP is True/False?
- _____ The excitation maxima of GFP are at 395 and 470 nm.
- _____ GFP is extracted from *A. victoria* jellyfish.
- _____ GFP has a highly unstable molecular structure.
- _____ GFP emits light at high intensity but with a very low quantum efficiency.
- _____ GFP gets activated by red or infrared light.

2.10: What are DNA microarrays? What are their applications?

2.11: Define what an antigen and an antibody are. What class of proteins are antibodies? Sketch an antibody, and illustrate where the antigen binds to the antibody. Explain what forces contribute to the antigen-antibody interaction.

2.12: Explain the error catastrophe in gene copying.

2.13: Explain the difference between DNA cloning, reproductive cloning, and biomedical or therapeutic cloning.

2.14: PCR is a very important analytical methodology to amplify genetic material. Briefly describe how it works and the merits/problems miniaturization of PCR entails.

2.15: Contrast human manufacturing techniques with nature's methods.

2.16: What does the term "binding constant" refer to in antibody chemistry? What is the value of a typical binding constant between an antigen and an antibody? What does the term "cross-reactivity" mean?

2.17: What is the difference between polyclonal and monoclonal antibodies? Briefly discuss the advantages and/or disadvantages of monoclonal vs. polyclonal antibodies.

2.18: The antigen-antibody interaction is used in a good number of bioanalytical techniques. One such example is affinity chromatography.

Describe the principle of affinity chromatography and give one example where you would need to use affinity chromatography.

2.19: Match the following terms with the appropriate definitions:

(1) Fluorophore	(a) Crystals composed of materials from periodic groups II-VI, III-V, or IV-VI and can be conjugated to biomolecules
(2) Stokes shift	(b) Method used to measure the ability of a molecule to move around in a cell over time
(3) FRAP	(c) Intense fluorescent probe used in many biological studies
(4) Quantum dots	(d) Difference in wavelength between fluorescence excitation maximum and the fluorescence emission maximum
(5) Green fluorescent protein	(e) Molecule that can be excited by light to emit fluorescence

Thanks to Mr. Youssef Farhat, UC Irvine.

2.20: What is a biosensor? Draw a schematic of a typical biosensor with its components and explain its operation with an example.

References

1. Vogel, S. 1998. *Cats' paws and catapults.* New York: W. W. Norton.
2. Gibson, D. G., G. A. Benders, C. Andrews-Pfannkoch, E. A. Denisova, H. Baden-Tillson, J. Zaveri, T. B. Stockwell, A. Brownley, D. W. Thomas, M. A. Algire, C. Merryman, L. Young, V. N. Noskov, J. I. Glass, J. C. Venter, C. A. Hutchison III, and H. O. Smith. 2008. Complete chemical synthesis, assembly, and cloning of a mycoplasma genitalium genome. *Science* 319:1215–20.
3. Morris, R. 1999. *Artificial worlds: Computers, complexity, and the riddle of life.* New York: Plenum.
4. Crick, F. 1981. *Life itself: Its origin and nature.* New York: Simon & Schuster.
5. Haynie, D. T. 2001. *Biological thermodynamics.* Cambridge, United Kingdom: Cambridge University Press.
6. Stewart, I. 1998. *Life's other secret: The new mathematics of the living world.* New York: Wiley.
7. Casti, J. L. 1989. *Paradigms lost.* New York: William Morrow.
8. Casti, J. L. 2000. *Paradigms regained.* New York: William Morrow.
9. Poundstone, W. 1999. *Carl Sagan: A life in the cosmos.* New York: Henry Holt.
10. Davies, P. 1999. *The fifth miracle: The search for the origin and meaning of life.* New York: Simon & Schuster.
11. de Duve, C. 1995. *Vital dust: The origin and evolution of life on earth.* New York: Basic Books.
12. Ridley, M. 2000. *Genome: The autobiography of a species in 23 chapters.* New York: Perennial.
13. Lee, D. H., K. Severin, Y. Yokobayashi, and M. R. Ghadriri. 1997. Emergence of symbiosis in peptide self-replication through a hypercyclic network. *Nature* 390: 591–94.
14. Ertem, G., and J. P. Ferris. 1996. Synthesis of RNA oligomers on heterogeneous templates. *Nature* 379:238–40.
15. Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci USA* 74:5088–90.
16. Capra, F. 1996. *The web of life.* New York: Doubleday.
17. Kauffman, S. 2000. *Investigations.* Oxford, United Kingdom: Oxford University Press.
18. Rebek, J. 1994. Synthetic self-replicating molecules. *Sci Am* 271:48–55.
19. Loewenstein, W. R. 1999. *The touchstone of life.* Oxford, United Kingdom: Oxford University Press.
20. Shapiro, R. 1991. *The human blueprint: The race to unlock the secrets of our genetic script.* New York: St. Martin's Press.
21. Watson, J. D. 1968. *The double helix.* New York: New American Library.
22. Crick, F. 1988. *What mad pursuit: A personal view of scientific discovery.* New York: Basic Books.
23. Maddox, J. 1988. *What remains to be discovered.* New York: Simon & Schuster.
24. Robinson, D. W. 1994. Medicine, molecules, and micro-engineering. *Nanobiology* 3:147–48.
25. Goldberg, A. L., S. J. Elledge, and J. W. Harper. 2001. The cellular chamber of doom. *Sci Am* 284:68–73.
26. Dawkins, R. 1989. *The selfish gene.* Oxford, United Kingdom: Oxford University Press.
27. Eigen, M., and P. Schuster. 1979. *The hypercycle: A principle of natural self-organization.* New York: Springer-Verlag.
28. Edwards, S. A. 2006. *The nanotech pioneers: Where are they taking us?* Weinheim, Berlin: Wiley-VCH Verlag GmbH & Co.
29. Dawkins, R. 1986. *The blind watchmaker.* New York: W. W. Norton.
30. Britten, R. J., and E. H. Davidson. 1969. Gene regulation for higher cells: a theory. *Science* 165:349–57.
31. Agrawal, A., Q. M. Eastman, and D. G. Schatz. 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* 394:744–51.
32. Cantor, C. R. 1990. Orchestrating the Humane Genome Project. *Science* 248:49–51.
33. Dunham, I., N. Shimizu, B. A. Roe, S. Chissoe, A. R. Hunt, J. E. Collins, et al. 1999. The DNA sequence of human chromosome 22. *Nature* 402:489–95.
34. Kenyon, C. 1996. Ponce d'Elegans: genetic quest for the fountain of youth. *Cell* 84:501–4.

35. Kenyon, C. 1996. Environmental factors and gene activities that influence lifespan. *C. elegans.* Ed. D. L. Riddle. Cold Spring Harbor Monograph Series no. 33. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
36. Passwater, R. A. 1985. *The antioxidants.* New Canaan, CT: Keats Publishing.
37. Hayflick, L., and P. S. Moorhead. 1961. The serial cultivation of human diploid cell strains. *Exp Cell Res* 25:585–621.
38. Hayflick, L. 1985. The cell biology of aging. *Clin Geriatr Med* 1:15–27.
39. Hayflick, L. 1994. *How and why we age.* New York: Ballantine Books.
40. Hayflick, L. 2000. New approaches to old age. *Nature* 403:365.
41. Wilmut, I., A. E. Schnieke, J. McWhir, A. J. Kind, and K. H. S. Campbell. 1997. Viable offspring derived from fetal and adult mammalian cells. *Nature* 385:810–13.
42. Briggs, R., and T. J. King. 1952. Transplantation of living nuclei from blastula cells into enucleated frogs' eggs. *Proc Natl Acad Sci USA* 38, 455–63.
43. Di Berardino, M. A., R. G. McKinnell, and D. P. Wolf. 2003. The golden anniversary of cloning: A celebratory essay. *Differentiation* 71:398–401.
44. Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131:861–72.
45. Yu, J., M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J. Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, I. I. Slukvin, and J. A. Thomson. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science* 5858:1917–20.
46. Margulis, L., and D. Sagan. 1986. *Microcosmos: Four billion years of evolution from our microbial ancestors.* New York: Simon & Schuster.
47. Cann, R. L., M. Stoneking, and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31–36.
48. Svoboda, K., C. F. Schmidt, B. J. Schnapp, and S. M. Block. 1993. Direct observation of kinesin stepping by optical trapping interferometry. *Nature* 365:721–27.
49. Noji, H., R. Yasuda, M. Yoshida, and K. Kinoshita. 1997. Direct observation of the rotation F1-ATPase. *Nature* 386:299–302.
50. Montemagno, C., G. Bachand, S. Stelick, and M. Bachand. 1999. Constructing biological motor powered nanomechanical devices. (Special conference issue of *Nanotechnology*: Sixth Foresight Conference on Molecular Nanotechnology.) *Nanotechnology* 10:225–31.
51. Kudo, S., Y. Magariyama, and S. Aizawa. 1990. Abrupt changes in flagellar rotation observed by laser dark-field microscopy. *Nature* 346:677–79.
52. Hunter, I. W., and S. Lafontaine. 1992. A comparison of muscle with artificial actuators. *Technical Digest: 1992 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC.
53. Meyer, E. E., K. J. Rosenberg, and J. Israelachvili. 2006. Recent progress in understanding hydrophobic interactions. *Proc Natl Acad Sci USA* 103:15739–46.
54. Pollack, G. H. 2001. *Cells, gels and the engines of life.* Seattle, WA: Ebner and Sons Publishers.
55. Schauer-Vukasinovic, V., L. Cullen, and S. Daunert. 1997. Rational design of a calcium sensing system based on induced conformational changes of calmodulin. *J Am Chem Soc* 119:11102–103.
56. Douglass, P. M., B. R. Wenner, R. D. Johnson, G. Barrett, J. C. Ball, P.-H. Hsu, J. Lee, L. G. Bachas, M. Madou, and S. Daunert. 2000. Biomimetic detection schemes on the lab CD. *Proceedings, NanoSpace 2000*. Houston, TX: The Institute for Advanced Interdisciplinary Research.
57. Wenner, B. R., P. M. Douglass, S. Shrestha, B. V. Sharma, S. Lai, M. Madou, and S. Duanert. 2001. Genetically designed biosensing systems for high-throughput screening of pharmaceuticals, clinical diagnostics, and environmental monitoring. *Progress in biomedical optics and imaging*. SPIE.
58. Mullis, K. 1998. *Dancing naked in the minefield.* New York: Vintage Books.
59. Petersen, K., W. McMillan, G. Kovacs, A. Northrup, L. Christel, and F. Pourahmadi. 1998. The promise of miniaturized clinical diagnostic systems. *IVD Technology* July [online].
60. Northrup, M. A., M. T. Ching, R. M. White, and R. T. Watson. 1993. DNA amplification with a microfabricated reaction chamber. *7th International Conference on Solid-State Sensors and Actuators (Transducers '93)*. Yokohama, Japan.
61. Wilding, P., M. A. Shoffner, and L. J. Kricka. 1994. PCR in a silicon microstructure. *Clin Chem* 40:1815–18.
62. Wittwer, C. T., and D. J. Garling. 1991. Rapid cycle DNA amplification: Time and temperature optimization. *BioTechniques* 10:76–83.
63. Wittwer, C. T., G. C. Fillmore, and D. J. Garling. 1990. Minimizing the time required for DNA amplifications by efficient heat transfer to small samples. *Anal Biochem* 186:328–31.
64. Schena, M., R. A. Heller, T. P. Theriault, K. Konrad, E. Lachenmeier, and R. W. Davis. 1998. Microarrays: Biotechnology's discovery platform for functional genomics. *TIB Tech* 16:301–06.
65. Fodor, S. P. A., J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251:767–73.
66. Eggers, M., M. Hogan, R. K. Reich, J. Lamture, D. Ehrlich, M. Hollis, B. Kosicki, T. Powdrill, K. Beattie, S. Smith, et al. 1994. A microchip for quantitative detection of molecules utilizing luminescent and radioisotope reporter groups. *Biotechniques* 17(3):516–25.
67. Friend, S. H. 1999. DNA microarrays and expression profiling in clinical practice. *BMJ* 319:1306.
68. Dyson, F. J. 1999. *The sun, the genome, and the internet—Tools of scientific revolutions.* Oxford, United Kingdom: Oxford University Press.
69. Gershenfeld, N. 1999. *When things start to think.* New York: Henry Holt and Company.
70. Gibson, D. G., J. I. Glass, C. Lartigue, V. N. Noskov, R.-Y. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z.-Q Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, C. A. Hutchison III, H. O. Smith, and J. C. Venter. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329(5987):52–56.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

3

Nanotechnology: Top-Down and Bottom-Up Manufacturing Approaches Compared

Outline

Introduction

Top-Down versus Bottom-Up Manufacturing

Important Concepts in Bottom-Up Manufacturing

Self-Assembly and Self-Organization

Biomimetic Nanomanufacturing

Supramolecular Chemistry

Molecular Assembler

Nanostructure Properties

Nanoparticle Manufacture

In Situ Fabrication

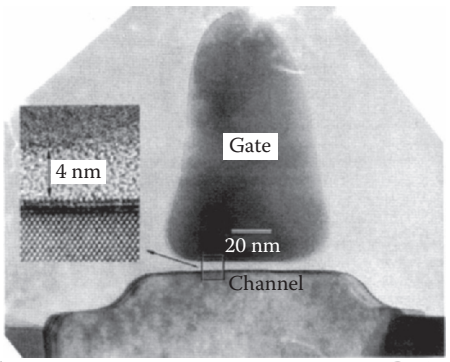
Gas Phase Synthesis

Carbon Nanostructures: Fullerenes

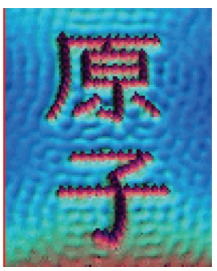
Carbon Nanotube Applications

Questions

References



Top-down nanotechnology. Electron micrograph of an SiO_2/Si interface in a Si N-channel MOS transistor fabricated using top-down methods.



(a)

(b)

(a) Kanji character for "atom," by Lutz and Eigler, IBM, Almaden. An example of manipulation/mechanosynthesis of iron atoms on a copper surface with a scanning tunneling microscope. (Courtesy of D. Eigler.) Mechanical manipulation of atoms is a bottom-up technique. (b) Cover of *American Scientist*: the carbon nanotube launch station. Carbon nanotubes are usually made with bottom-up techniques.

Our ability to arrange atoms lies at the foundation of technology.

K. Eric Drexel
1986

Introduction

Nario Taniguchi coined the word *nanotechnology* in 1974 to describe machining with tolerances of less than 1 μm (see Volume II, Figure 6.3). Humans are still best at building structures we can grasp with our hands and see with our eyes, and we have only relatively recently started to intentionally craft objects at the nanoscale with nanotechnology. Over the last two decades, nanotechnology has come to mean two quite different things. In the broadest sense it is any technology dealing with building structures between 1 and 100 nm in size, in a narrower sense, as promoted by K. Eric Drexel, it involves designing and building machines in which every atom and chemical bond is precisely specified.[1]

About the first definition, one commentator quipped that some companies are willing to say their product is based on nanotechnology if it is made of atoms![1] Clearly, some further refinement is in order for a working definition. Our criteria for nanotechnology are that the miniaturized structures have at least one dimension smaller than 100 nm, are crafted with a novel technique, and have been intentionally designed with a specific nanofeature in mind. Medieval church stained glass, with embedded gold nanoparticles, does not qualify, since the reason for the color dependence on the size of the particles was not yet appreciated in the Middle Ages (Figure 3.1). Finally, nanotechnology attempts to control or manipulate matter at the atomic scale. Our definition dovetails nicely with the criteria adopted by the National Nanotechnology Institute (http://www.nano.gov/):

1. Nanotechnology involves structures in the 1–100-nm range
2. Nanotechnology creates and uses structures that have novel size-based properties
3. Nanotechnology takes advantage of the ability to control or manipulate at the atomic scale.

The manufacture of devices with dimensions between 1 and 100 nm is based on either top-down manufacturing methods (starting from bigger building blocks, say a whole Si wafer, and chiseling them down into smaller and smaller pieces by cutting, etching, and slicing) or bottom-up manufacturing methods (in which small particles such as atoms, molecules, and atom clusters are added for the construction of bigger functional units). The top-down approach to nanotechnology we call *nanofabrication* or *nanomachining*—an extension of the micromachining approaches [microelectromechanical system (MEMS)] introduced in previous chapters. The bottom-up approach we call *nanochemistry*. Nanochemistry is also called *molecular nanotechnology*; it is the arranging and building of more complex organic and inorganic structures at molecular level by chemistry, molecular biology, or mechanical positional assembly. An example of this second approach is the self-assembly of a monolayer from individual molecules on a gold surface (see Volume II, Figure 8.14). Bottom-up methods are nature's way of growing things. In biomimetics, one studies how nature, through eons of time, developed manufacturing methods, materials, structures, and intelligence. These biomimetic studies, expanded upon in Chapter 2, this volume, are now inspiring design, materials, and engineering of human-made miniature objects.

Both nanomachining and nanochemistry are considered nanotechnology. We start this chapter off by contrasting these two opposite methods to obtain nanostructures. Along the way we highlight the Drexel-Smalley debate about the "correct path" toward nanotechnology. We then list all the

FIGURE 3.1 Stained glass: *Labors of the Months* (Norwich, England, ca. 1480). The ruby color is due to embedded gold nanoparticles.

properties that make nanodevices special, i.e., physical and chemical behavior that depend on particle size and allow one to engineer their properties. In one extreme, we have atoms and molecules, generally less than a nanometer in size, studied in quantum chemistry, and in the other extreme we have bulk condensed matter physics that deals with solids with an infinite array of bound atoms. Nanoscience deals with the in-between mesoworld where neither quantum chemistry nor the classic laws of physics apply. The listing of nanoparticle properties is followed by a section on preparation of nanostructures in general: from high-energy ball milling to photolithography to gas phase and liquid phase processes. Finally, as more specific examples, we detail properties and highlight bottom-up manufacturing approaches of carbon nanotubes, graphene sheets, and nanowires.

Top-Down versus Bottom-Up Manufacturing

General Characteristics of Top-Down and Bottom-Up Manufacturing

Top-down nanofabrication is principally based on lithography and traditional mechanical machining, etching, and grinding, whereas bottom-up nanochemistry is based on chemical synthesis, self-assembly, and positional assembly. The distinction is clarified in Figure 3.2. These complementary developments constitute different aspects of nanotechnology; both aim to create and use structures, devices, and systems in the size range of about 0.1–100 nm (covering the atomic, molecular, and macromolecular length scales).

There are some important distinctions to be made between these two approaches to making very small things. Whereas top-down methods essentially impose a structure or pattern on a substrate, in self-assembly and self-organization-based bottom-up methods the assembly or organization of atomic and molecular constituents into larger structures is inherent through processes in the manipulated system. Self-assembly and self-organization are driven by thermodynamics and hold the intriguing potential of delegating more and more mechanical or logical operations to materials instead of to manufacturing machines. Assembly of atoms or molecules into larger aggregates in chemical synthesis is driven by chemical reactions with the Gibbs free energy changes dominated by the enthalpy changes associated with breaking and creating bonds ($\Delta G = \Delta H - T\Delta S$). In self-assembly the driving forces are

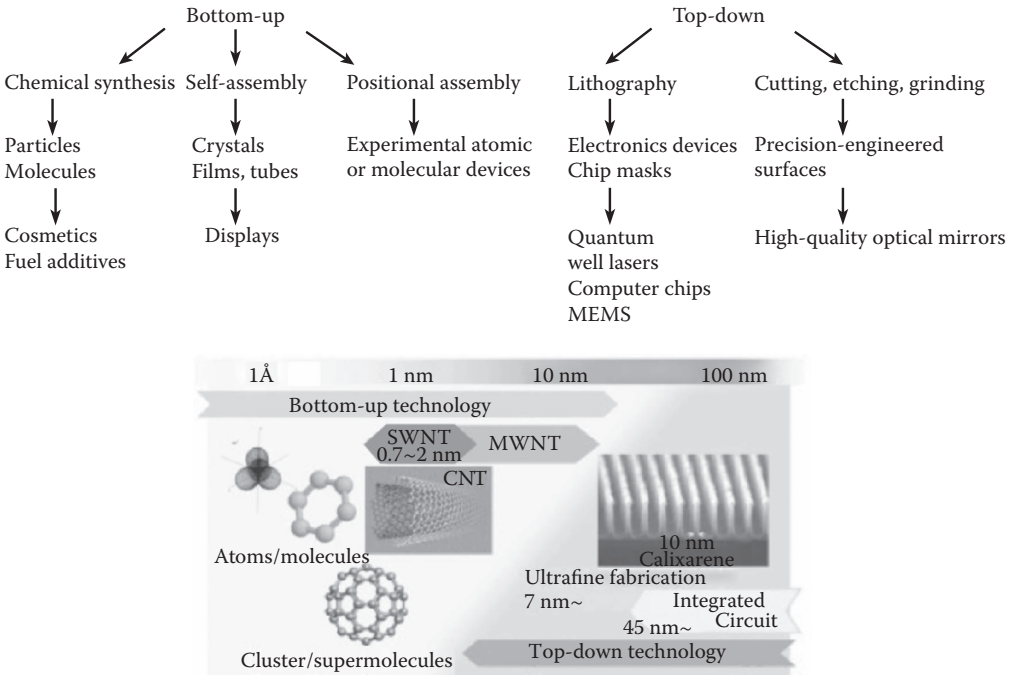


FIGURE 3.2 Contrasting bottom-up nanofabrication with top-down nanochemistry.[2] (SWNT: single wall nanotube; MWNT: multiwall nanotube; CNT: carbon nanotube.)

weak interactions (hydrogen bonds, van der Waals forces, hydrophobic/hydrophilic interactions, etc.; see Volume I, Chapter 7) and are usually driven by entropy changes only. Both chemical synthesis and self-assembly-based bottom-up processes are based on growth phenomena, and any growth scenario is governed by the competition between kinetics and thermodynamics. Positional assembly is the least mature bottom-up technique and uses mechanical means, e.g., an atomic force microscope (AFM), to assemble nanostructures by forcing atoms or molecules in close enough proximity and with the right orientation to form or break bonds (see Volume II, Figures 2.43 to 2.45).

A top-down manufacturing process (for example, the writing of an e-beam pattern) is deterministic, but a bottom-up process (for example, the chemical synthesis of a quantum dot from a colloidal solution) is stochastic. In the latter case, instead of dealing with only one possible "reality" of how the process might evolve under time, in a stochastic or random process there is some indeterminacy in its future evolution, described by probability distributions. This means that even if the initial condition is known, there are many possibilities to which the process might go, but some paths are more probable and others less (statistics take over). Current top-down manufacturing techniques are in a sense crude and wasteful compared with the bottom-up manufacturing techniques we survey in this chapter. Top-down, making a smooth edge requires grinding away and discarding trillions of atoms (i.e., removing excess material). The hope for bottom-up manufacturing is to build with atomic precision without wasting material.

Applying modified integrated circuit (IC) methods, MEMS led to major advances in manufacturing strategies, but the method still presents a limited choice of materials and works with building blocks that are large and crude compared with the nanochemistry arsenal. As humankind learns to build with the same construction set nature uses, we are bound to challenge nature in nanoengineering. Because it uses relatively large building blocks, human manufacturing is rapid. The expectation is that nature, because it uses much smaller building blocks, for example, atoms with a diameter of 0.3 nm, must be very slow. To offset the time it takes to work with small, basic building blocks, nature, in growing an organism, relies on additive processes featuring massive parallelism and self-assembly. Massive parallelism and self-assembly are two areas where human technology still is very backward compared with nature's methods. In Table 3.1 we compare some of the key characteristics of bottom-up manufacturing (nanochemistry) with MEMS technology.

In Figure 3.3 we illustrate the convergence over time of top-down and bottom-up methodologies. The lines representing top-down manufacturing in Figure 3.3 are a simplified representation of a 1974 diagram by Norio Taniguchi (see Volume II, Figure 6.3). The higher of the two lines, marking

Table 3.1 MEMS and Nanochemistry Compared

MEMS Technology	Nanochemistry
Top-down	Bottom-up
Miniaturization with micrometer and submicrometer tolerance	Miniaturization with atomic accuracy
Builds mostly on solid-state physics and the band model from quantum mechanics	Builds almost exclusively on quantum mechanics
Evolved from the integrated circuit fabrication technology	Evolved from molecular engineering, colloid chemistry, supramolecular chemistry, and scanning proximal tool assembly
Many MEMS products are commercially successful	Only a few nanomaterials are on the market
Deterministic processes	Stochastic processes
Limited set of building materials	Huge number of possible building materials
Relatively large building blocks make for a fast process	Very small building blocks necessitate massive parallelism and self-assembly

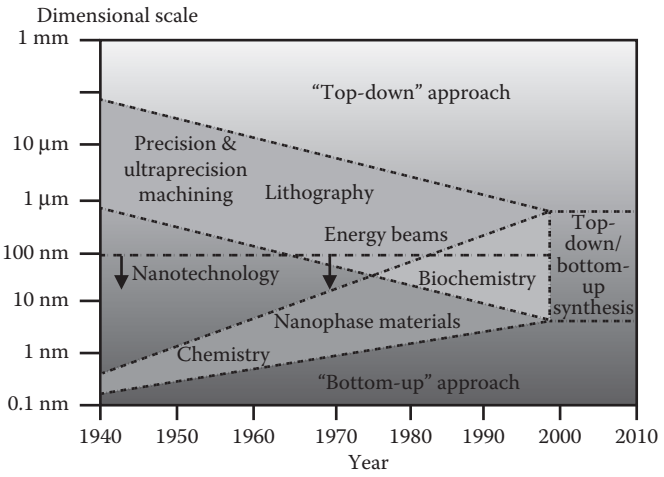


FIGURE 3.3 Convergence of bottom-up nanochemistry and top-down nanofabrication. The nanotechnology line is placed at 100 nm[2].

progress of absolute tolerance in top-down machining, corresponds to precision machining. The lower line represents absolute tolerance in ultraprecision machining, lithography, and energy beam techniques. Some caution is in order about the use of the term *absolute tolerance* in nanotechnology: a more rigorous definition of nanotechnology should differentiate between absolute tolerances in lateral and vertical nanostructures. Deposition of nanometer-thick films is relatively easy with all types of "traditional" chemical vapor deposition (CVD) and physical vapor deposition (PVD) techniques, but patterning of a thin film laterally, with critical dimensions (CDs) that extend from about 100 nm down to atomic orders of magnitude, is still quite a challenge when employing top-down methods. Bottom-up methods hold the potential for nanometric control in all three dimensions. This explains why we tend to make quantum wells with lithography-based top-down methods, since only the thickness direction requires tight control, but quantum wires and quantum dots are mostly made with bottom-up methods, since for those quantum structures we also need control in the lateral directions. Vertically layered structures, i.e., nanoscaled thin films, are obviously much easier to achieve than laterally patterned features. When people discuss breakthroughs in top-down nanotechnology, they mostly refer to better lateral patterning techniques. In the same context, in Volume II, Chapter 6, we explain how the excellent absolute tolerances for precision and ultraprecision mechanical machining only hold true for surface finishes, i.e., in the thickness direction, while in the lateral directions absolute tolerances are again worse. Mechanical machining lateral tolerances are worse than the lateral tolerances achievable with lithography, and those in turn are worse than what we can achieve with nanochemistry techniques: using electron beam machines we can cut shapes with a precision of 2.5 nm, about 1,000 times better than what is achievable in a standard machine shop (2.5 μm is standard, and 1 μm may be achieved in specialty shops), and precision in nanochemistry is about 1 Å, 25 times better than what e-beam offers today. The two lines in Figure 3.3 representing the bottom-up manufacturing progress in achievable absolute tolerances show how bottom-up processes evolved to enable, ever-larger structures through progress in nanochemistry. It is seen from this figure that from the mid-1970s on, feature sizes that could be achieved by nanofabrication (top-down) and nanochemistry (bottom-up) started to overlap. Over the next ten years, the size overlap of objects fabricated with either approach will continue, and molecular engineers and supramolecular chemists, jointly with IC, MEMS, and nanoelectromechanical system (NEMS) specialists, now making devices with 45-nm feature sizes and below, will find new ways to combine top-down and bottom-up manufacturing methods.

In Volume II, Chapter 6, we commented that the reason for the slowdown in electronic device miniaturization hinges on securing more mechanical stability in the positioning stages, but even more so on manufacturing cost and the fact that transistors are becoming too leaky below the 32-nm node, where quantum effects start playing a role. To further extend manufacturing accuracy in electronics one has to look beyond current manufacturing methodologies and consider methods such as molecular engineering and nanofabrication of quantum devices. Micromachining of special lithography tools will play a pivotal role in those new machining approaches. For example, the scanning tunneling microscope (STM), based on micromachining ultrasharp tips on Si cantilevers, creates machining opportunities with accuracies on the atomic level, well beyond the predictions of the

Taniguchi and Moore curves in Volume II, Figure 6.3. For mechanical NEMS structures, operation in the micro- and nanodomains imposes barriers other than those described for electronic devices. New design philosophies, producing results that are more error-insensitive with simpler mechanisms, smaller numbers of parts, and, wherever possible, flexible members rather than rigid ones must be adopted. Most of this book deals with top-down machining; in this chapter we deal in some more detail with bottom-up manufacturing.

Important Concepts in Bottom-Up Manufacturing

Introduction

Before reviewing nanostructure properties and listing methods for bottom-up nanotechnology manufacturing, some important contributing concepts and their definitions are in order. What follows are clarifications and definitions for concepts that are often utilized in bottom-up manufacturing but are not always well understood, such as self-assembly, self-organization, biomimetics, supramolecular chemistry, self-replication, and molecular assemblers.

Self-Assembly and Self-Organization

Self-Assembly

Self-assembly and self-organization are important concepts in both nanotechnology and biology. According to Maasden and Kaiser, about 10% of papers dealing with nanotechnology published between 1990 and 2005 address the concept of self-assembly, and among those, 3% deal with self-organization.[3] The distinction between self-assembly and self-organization is not always readily apparent, and this causes considerable confusion, especially when self-replication (see below) is thrown into the discussion.

Thermodynamics represents the simplest way to differentiate between self-assembly and self-organization. Self-assembly concerns equilibrium situations and is purely driven by the tendency of systems to minimize their free energy in accordance with the second law of thermodynamics. Self-organization, on the other hand, occurs in systems that are open and are driven by a constant input of energy and that are far from thermodynamic equilibrium. As we pointed out earlier, in self-assembly free energy minimization is often driven by entropy changes, ΔS (Equation 2.7). For natural examples of self-assembly, consider phospholipids with hydrophobic and hydrophilic ends placed in aqueous solution spontaneously forming stable structures (Figures 2.41–2.43, this volume) or the intricate three-dimensional (3D) patterns in protein folding (Figure 2.11, this volume). In technology, self-assembly is taken advantage of, for example, in sensors such as DNA probes and immunosensors, but is not used much in human manufacturing yet. We use self-assembly in manufacturing, for example, when growing crystals, in building self-assembled monolayers (see Volume II, Figure 8.14), and in colloid chemistry. In colloidal chemistry one can assemble inorganic solids, such as CdSe crystals, with astonishing regularity in size and properties, comparable to what can be accomplished with nanolithography, and this method has become an important route for building quantum dots (see Figure 3.23).

There is an upswing in self-assembly work that is bringing a renewed excitement to the nanotechnology field. In Chapter 4, this volume, on packaging, for example, we detail the differences between deterministic and stochastic parallel microassembly and clarify the need and tremendous opportunities for self-assembly in packaging of nanocomponents. Recent examples of work in self-assembly based packaging, clarified in Chapter 4, include DNA-mediated self-assembly of metallic and semiconductor components and micro-origami assembly of optical components and of small polymer sheets (see also the example that follows).

Example of Self-Assembly: Scaffolded DNA Origami

Weight

The reproducible generation of atomically precise patterns and shapes is at the heart of nanotechnology. Although research has shown improvement at creating nanosized shapes, it is still a tremendous challenge to create desired nanostructures that always do have the same size and shape with

FIGURE 3.4 Origami art. (From Origami.com, http://www.origami.com.)

atomic accuracy. Here we review a recently developed method that holds that potential. It involves a technique for folding long single strands of DNA (ssDNA) into arbitrary two-dimensional shapes using synthesized DNA staple strands in so-called "scaffolded DNA origami." Scaffolded DNA origami builds on Seeman's DNA nanotechnology, discussed in Chapters 2 and 4, this volume (e.g., Figure 4.49) and pushes it quite a few steps further.

Origami is the art of folding paper; it originated in China as "Zhe Zhi" and was adapted by Japanese monks in the sixth century AD (see an example of origami in Figure 3.4). Ancient origami is not self-assembly, but the scaffolded DNA origami reviewed here is. DNA by its nature does exhibit self-assembly (see base pairing in Chapter 2, Figure 2.6). The same characteristics that make DNA so effective genetically also make it a good choice for programmed self-assembly in nanotechnology:

- DNA can be manipulated to create exact copies that are extremely accurate (atomic accuracy).
- DNA is predictable and programmable.
- DNA also has the ability to store enormous amounts of information.

The most important part of controlling the way a DNA strand folds is knowing its genetic nucleotide structure. The scaffold strand that has been used most often is circular genomic DNA from the virus M13mp18 (natural M13mp18 is 7249 bases long). Synthesized DNA staple strands then bind to complementary base pairs on the scaffold M13mp18 viral genome as illustrated in Figure 3.5. Both the DNA strand and the staple strand must be known to avoid unwanted pairing and to ensure that the stapling takes place in desired locations only, as illustrated in Figure 3.6.

The design for a desired shape is made by raster filling the shape with the ~7-kilobase single-stranded scaffold molecule and by choosing more than 200 short oligonucleotide staple strands to hold the scaffold in place. Once synthesized, a large excess of staples (100×) is mixed with scaffold and annealed in a polymerase chain reaction (PCR) machine (<2 h) (for details on how PCR works, see Chapter 2, this volume, Figure 2.53). Synthesized DNA staples self-assemble, with the viral strands resulting in folded area shapes (squares, disks, five-pointed-star, etc.) approximately 100 nm in diameter, with a spatial resolution of 6 nm. A nucleotide may serve as a 6-nm pixel in these designs, so any type of word or image can be programmed. Individual DNA structures can also be made to form larger assemblies of two-dimensional (2D) periodic lattices. Sample DNA origami are deposited on mica and imaged with an AFM in tapping mode. Experimentally verified by the creation of a dozen shapes and patterns, the method is easy, has a high yield, and lends itself well to automated design and manufacture. So far, computer-aided design (CAD) tools for scaffolded DNA origami are simple, require hand-design of the folding path only, and are restricted to two-dimensional designs. If the method gains wide acceptance, better CAD tools will be required.[4]

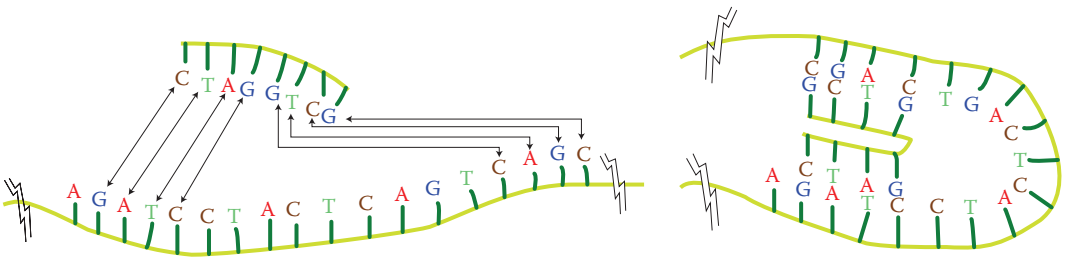


FIGURE 3.5 DNA staple concept. The sequence of DNA provides a unique address for each location. Staple strands bind locations together according to the design.

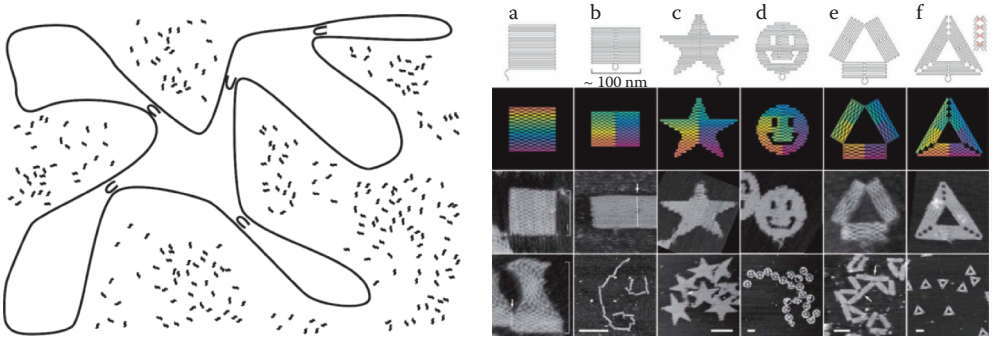


FIGURE 3.6 The design for a desired shape is made by raster filling the shape with the 7-kilobase single-stranded scaffold molecule and by choosing more than two hundred short oligonucleotide staple strands to hold the scaffold in place.[4]

Self-Organization

Self-organization is a term for those types of pattern forming systems that are open and thus driven by a constant input of energy and that are far from thermodynamic equilibrium. The Rayleigh-Bénard instability, or, more simply, Bénard heat convection, is regarded as a classic case of self-organization. This type of heat convection is named after French chemist Henri Bénard, who first recorded the effect in 1900. Because the British physicist Lord Rayleigh further investigated Bénard's observation, the conditions that produce it came to be called the *Rayleigh-Bénard instability*. Anyone can observe the formation of Bénard cells by heating a thin layer of oil in a pan, as illustrated in Figure 3.7. With the oil uniformly heated from below, a constant heat flux is established, moving from the bottom to the top of the oil layer in the pan. The oil itself remains at arest, and when the temperature difference between top and bottom of the oil layer reaches a certain critical value, thermal conduction is replaced by heat convection, in which the coherent motion of large number of molecules transfers the heat. At this point a very strikingly ordered pattern of hexagonal cells (a "honeycomb") appears, in which hot oil rises through the center of the cells, while cooler liquid descends to the bottom along the cell walls. The same physics underlies a weather phenomenon that leaves hexagonal patterns behind in sand or snow. The sun heats up the desert sand or a snowfield, and warm air from the surface rises, while cold air descends from the upper layers of the atmosphere. The conflict is resolved again by the emergence of orderly patterns of fluid flow under the action of an imposed temperature gradient, resulting in hexagonal circulation vortices that leave their imprints on sand dunes in the desert and on arctic snowfields.

During the 1960s, Ilya Prigogine (1917–2003), winner of the Nobel Prize in Chemistry, 1977, developed a new, nonlinear, irreversible thermodynamics, that mathematically describes the self-organization phenomenon described here.[5] It is

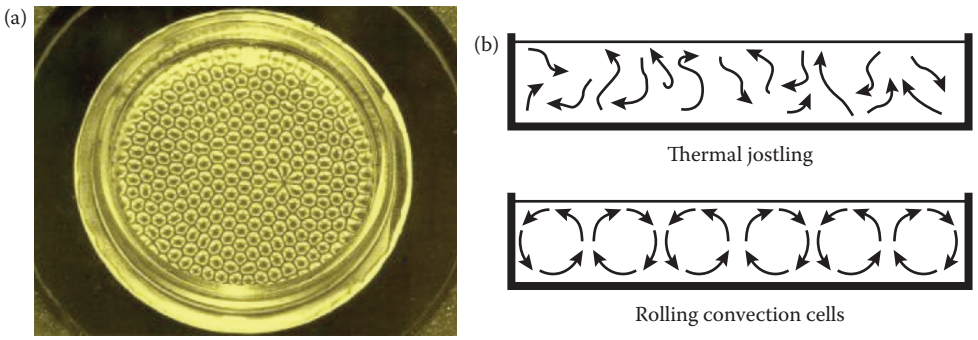


FIGURE 3.7 (a) Patterns of hexagonal Bénard cells viewed from above. The most interesting stable patterns appear right at the edge of chaos. This is happening far from equilibrium, thanks to the energy flowing through an open system and being dissipated. (b) The hexagonal cells—the simple global order—is more efficient at dissipating heat than the simple local motions.

quite unlikely that anyone would have deduced these phenomena from the Navier-Stokes equations (see Volume I, Chapter 6). On the basis of Prigogine's work, Bénard cells are now understood to arise from the properties of the liquid state at a bifurcation point in which a self-organizing structure emerges. The huge number of degrees of freedom of individual molecules in the fluid at this critical point suddenly organize into a coherent and simple flow pattern that dissipates heat more effectively than simple thermal conduction. In this new emergent behavior, convection (transport of heat by mass movement) in self-organized spatial structures becomes the dominant process. This change is initiated by an instability (noise) in the system, and the result obeys the second law of thermodynamics, since this mechanism is a more efficient way of disposing of the system's heat. More disorder must generally prevail (increase in ΔS), but one can still have a local increase in order, such as in the hexagonal Bénard cells, as long as overall they contribute to the wider disordering of the surroundings, i.e., the dissipation of energy as entropy or waste heat.

What is happening at a molecular level in the frying pan is that when the pan is only gently warmed, each heated molecule moves by itself, and the molecules all chaotically barge about and only eventually emerge at the surface, where they shed their excess heat. But at a certain critical temperature, a phase transition occurs. The jostling molecules are becoming entrained into a coherent flow, the rolling convection current that moves them smoothly from the heated bottom to the cooler surface. The hexagonal cells—the simple global order—is more efficient at dissipating heat than the simple local motions (see Figure 3.7b). But when one heats a Bénard cell just a little more, its orderly hexagonal cells soon break up and instead turbulence results and disorder appears again. The order turns out to exist only for a very finely controlled balance between heat input and heat dissipation.

In self-organization what differs from what happens in self-assembly is that internal variations, called *fluctuations* or *noise*, trigger self-organization. The increase in organization corresponds to a decrease of entropy equivalent to an increase in redundancy, information, or constraint. In other words, after the self-organization process, less ambiguity about which state the system is in results. Prigogine called systems that continuously export entropy in order to maintain their organization *dissipative structures*, i.e., systems that exist far from thermodynamic equilibrium, dissipate the heat generated to sustain them, and have the capacity to change to higher levels of orderliness or self-organization. In classic thermodynamics the dissipation of energy in heat transfer, friction, and the like was always regarded as waste. Dissipative structures introduced a radical change in this view by showing that in open systems, dissipation becomes a source of order. According to Prigogine, dissipative systems contain subsystems that continuously fluctuate. At times, a single such fluctuation or a combination of them may become so magnified by possible feedback that it shatters the preexisting organization in what is called an emergent new property. At such moments, called *bifurcation points*, it is impossible to determine in advance whether the system will disintegrate into "chaos" or leap to a new, more differentiated, higher level of "order." While dissipative structures receive their energy from outside, the instabilities and jumps to new forms of organization are the result of internal fluctuations amplified by positive feedback loops. Thus, amplifying "runaway" feedback, which was traditionally seen as destructive, now appears as a source of new order and complexity in this model.

Self-assembly and self-organization are fundamentally different from a thermodynamics point of view: while self-assembly stems from minimization of the free energy, self-organization requires an open system and a constant energy input.* Whereas self-assembly is quite well understood and is clearly important at small length scales in biology (protein folding, for example), self-organization is not as well understood, and the length scale at which it comes into play in biology is still under debate. For the specific case of cell membrane assembly, it is believed that both processes eventually lead to quite similar patterns, but that their evolution occurs on very different length and time scales. Self-assembly produces periodic protein patterns on a spatial scale below 0.1 μm in a few seconds followed by extremely slow coarsening,

* Unfortunately, important players in this field, such as Whitesides, seem to maintain their own definitions of self-assembly, and J.M. Lehn uses both terms more or less interchangeably to characterize the synthesis of supramolecular structures.

whereas self-organization results in a pattern wavelength comparable to the typical cell size of 100 μm within a few minutes, suggesting different biological functions for the two processes.[6] One thing that everyone agrees on is that living organisms cannot arise wholly from self-assembly, because in the absence of a continuous supply of energy they die. In summary, crudely said: viruses self-assemble, but elephants (perhaps) self-organize (http://www.softmachines.org/wordpress/?p=178). Another parallel to living systems here is that such dynamical structures are not devised to exhibit this behavior: they develop spontaneously from random initial conditions (we emphasize: not from special initial conditions).

The behavior of interest, such as that of the Bénard cells, is often found in the transition between order and chaos—at the edge of chaos—and is classified as a kind of organized complexity. The same type of behavior is involved in adaptive path minimization by ants, wasp and termite nest building, army ant raiding, fish schooling and bird flocking, coordinated cooperation in slime molds, synchronized firefly flashing, evolution by natural selection, game theory, the evolution of cooperation, etc. When a flock of birds is taking off from a lake (see Figure 3.8), birds compete for position, try to get free of the maelstrom of their fellow birds and go up into the air. They want to be part of the flock, yet at the same time try to avoid collisions with each other. Computer models show that each bird's attempt to keep a minimum and maximum distance from others causes flight paths to couple into feedback loops of attraction and repulsion. Positive and negative feedback balance so that the individual birds appear transformed into a single organism. The late Prigogine wrote about the reception his irreversible thermodynamics received at first:

> It is difficult today to give an account of the hostility that such an approach was to meet. For example, I remember that towards the end of 1946, at the Brussels IUPAP meeting, after a presentation of the thermodynamics of irreversible processes, a specialist of great repute said to me, in substance: "I am surprised that you give more attention to irreversible phenomena, which are essentially transitory, than to the final result of their evolution, equilibrium."
>
> **Ilya Prigogine**
> *1980*

FIGURE 3.8 Flock formation as an example of emergent, spontaneous organization. (From http://www.cs.vu.nl/~schut/bird12.jpg.)

Biomimetic Nanomanufacturing

Introduction

Biomimetics is the application of designs, materials, and systems found in nature to engineering and technology. As we saw in Chapter 2, this volume, where we cover biomimetics in much more detail, engineers and scientists tend to increasingly turn toward nature for inspiration on designs, materials, and systems, spawning a number of technology innovations. There is an obvious reason behind this trend: the rigorous competition of natural selection means that waste and inefficiency are not tolerated in natural systems, unlike many of the technologies developed by humans. For this same reason, many researchers now give natural, biomimetic manufacturing methods another look. Within this context, the guiding principles involved in the construction of living organisms are as follows.[7,8]

- The use of composites (rather than monolithic materials in human engineering)
- Several successive levels of organization (hierarchy), often in the form of two prevalent natural design motives: fractals (trees, lungs, the brain, the cardiovascular system, etc.) or layered structures (tree rings, onions, etc.) and combinations thereof
- Very wide range of manufacturing sizes: from atoms to several meters, i.e., 10^{12} orders of magnitude

- The use of carbon instead of Si
- Ambient temperature versus the high temperatures used in human engineering
- Low energy (ATP) versus high energy
- Mobility of components (Brownian motion) versus rigidity with no or few fluctuations
- Operation in aqueous solution rather air (wet vs. dry engineering)
- Compartmentalization in cells
- Plasticity: soft, flexible materials that change conformation in response to the environment versus rigid components
- Self-assembly (see above), self-organization (see above), and self-replication all through weak bonds
- Order out of noise versus noise as a nuisance (see above)
- The use of templates (e.g., the reading of genes by messenger RNA to build proteins in the cells' ribosomes, as shown in Figure 2.12)
- Variable number of components versus fixed number of components
- Adhesive surfaces (van der Waals) versus separated surfaces
- Robustness through stochastics versus through redundancy

These principles provide inspiration for the implementation of human nanotechnology. As noted in Chapter 2, this volume, biomimetics should not be taken too far beyond inspiration from nature, because the ultimate goal in nature is spreading genetic information to the next generation, not to generate the most efficient manufacturing technique. We add a few comments now for several of the attributes of a natural growing process listed above.

Nanocomposites

Nanocomposite materials are widespread in biological systems, such as bone, cell walls, and soft body tissue. The nacre (mother-of-pearl) of abalone shell, an oriented coating composed of alternating layers of aragonite ($CaCO_3$) and a biopolymer, is an oft-studied example (Figure 3.9). The laminated structure provides strength, hardness, and toughness. Nacre is twice as hard and 1000 times as tough as its constituent phases. Humankind has

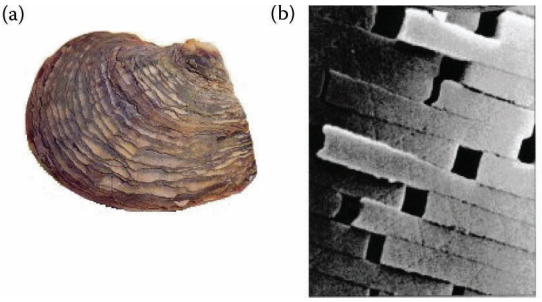


FIGURE 3.9 Abalone shell nanocomposite structure: (a) abalone shell and (b) scanning electron microscope image of the aragonite composite structure.

accumulated plenty of experience with composite materials, but attempts to make "biomimetic" nanocomposite assemblies are quite recent. Along this line, Sellinger et al. reported on an efficient self-assembly process for preparing nanolaminated organic-inorganic coatings that mimic nacre.[9] New biomimetic materials such as scaffolds for regrowth of tendons tend to be composites too, although not on the nanoscale; in the case of tendon scaffolds, composites are made from carbon fibers and polylactic acid.[7]

Hierarchical Organization and Range of Manufacturing Sizes

Proteins, bone, and tissue fibers have a hierarchical organization (see Figure 2.11 for the various protein structural levels) much more complex than that of materials of the nonliving world (e.g., crystals). This organization bestows many benefits on living organisms, such as a high strength-to-density ratio and the ability to serve several functions simultaneously. The ability of humans to mimic this feat is still limited and will improve only if we become more adept at manipulating a wider variety of materials at the nanoscale, use smaller building blocks, and build in three dimensions with more dexterity. Traditional lithography is mostly limited to 2D structures and is impractical for printing feature sizes below 0.1 μm. While 3D lithography techniques, as described in Volume II, Chapter 9, are limited in feature size, they do offer a possible route to 3D microstructures.

Two ubiquitous design patterns of hierarchical structures in nature are fractals (trees, brains, river

(a) (b)

FIGURE 3.10 Examples of fractals. (a) Romanesco cauliflower. (b) The fractal coastline of eastern Greenland, with its many fjords.

deltas, fjords, snowflakes, blood circulation system, etc.; see Figure 3.10) and layers (onions, tree rings, etc.) and combinations thereof. Consider a tree where the trunk forks into branches and the branches into smaller twigs. Twigs have leaves, which themselves repeat the dendrite pattern in their veins. The tree also has a layered build-up radially in its stem, branches, and twigs. An object is called a *fractal* if it has structures at a large number of levels. If they are identical, the fractal is said to be *self-similar* (see Chapter 7, this volume, on scaling, for more details on the similarity concept). Self-similarity can also be called *scale invariance* because we always find the same pattern upon zooming into at different length-scales.

In Chapter 9, this volume, we show that the underlying reason for fractal hierarchical structures is simply that such a shape maximizes surface area while minimizing the energy involved in reaching each point on that surface. A fractal network can both absorb and excrete at a maximal rate, whether it is water and electricity, oxygen, carbon dioxide, bloodstream nutrients, or even neural control information. We calculate in Chapter 9 that for a fractal the resistance is minimized and the surface-to-volume (S/V) scale is invariant. We also demonstrate that in a zero-order approximation, a layered structure attains the same goals as a fractal, with both resistance and S/V scaling, in the same way a fractal does when miniaturized.

Using these ubiquitous hierarchical design patterns (fractals and layers) and combinations thereof, nature covers sizes of biological life forms from the atom to several meters (e.g., a 10-m-tall tree), i.e., it covers at least twelve orders of magnitude in size. Human manufacturing techniques, on the other hand, typically are limited to much smaller ranges of size capability (e.g., lithography from 90 nm to 1 cm, i.e., hardly three orders of magnitude). In Figure 3.11, we sketch the challenge involved in making an artificial fractal, say a carbon tree. As indicated in this drawing, as the size ranges we work with approach the atomic and molecular size, stochastic manufacturing techniques (1 Å–100 nm), perhaps from a solution in a beaker, must take over from the larger scale deterministic manufacturing means, such as lithography (100 nm–1 cm) and computer numerical controlled (CNC) machining (1 cm–10 m). In other words, we are forced to switch to different manufacturing options as we move across different length-scales.

The Use of Carbon Instead of Si

Although silicon, human's foremost technological material, is abundant in silicate minerals, carbon is the key to life. In Chapter 5, this volume we explore how in nature, carbon occupies a very special place as the essential building block of life, and we notice how also in human enterprise carbon in its various forms has acquired an enviable position. To name a few, diamonds are made of carbon, and graphite, coke, diamond-like carbon, and glassy carbon are all forms of carbon. So are the more recently discovered buckyballs, carbon fibers, and carbon

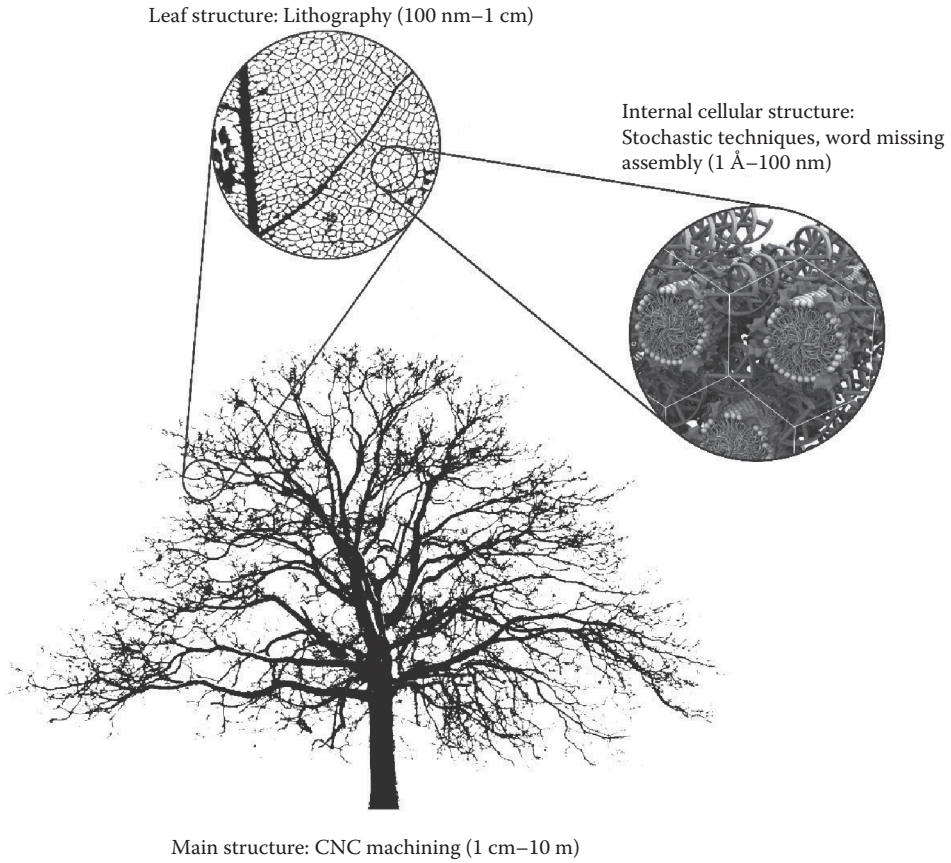


FIGURE 3.11 Making an artificial carbon tree would require switching to various manufacturing strategies, depending on the size. (Drawing by George O'Quinn, UC Berkeley.)

nanotubes (CNTs). It can rightfully be argued that carbon is now also becoming a more and more important material in human technology. Some key advantages of carbon compared with Si are summarized in Table 5.9. As we see from this table, carbon polymerizes better than Si (making organic chemistry a field all on its own); it comes in all types of forms (amorphous, graphite, nanotubes, etc.); it has a wide electrochemical stability window; and it is biocompatible, low cost, chemically inert, and easy to functionalize.

In Chapter 5, this volume, we further introduce the reader to a vision of a carbon-based MEMS and carbon-based NEMS world.

Ambient Temperature, ATP, Soft Materials, Compartimentalization (Cells), and Water Management

Of the many motivating examples of the promise of biomimetics we encountered in Chapter 2, this volume, consider enzymes. There are a plethora of enzymes that outperform synthetic catalysts by several orders of magnitude. For instance, there is no human-made catalyst for producing ammonia from its elements at ambient temperature and pressure, as the nitrogenase of nodule-forming bacteria does (the industrial Haber-Bosch synthesis of ammonia from nitrogen and hydrogen at high temperatures and pressures pales in comparison).

Molecular motors, also covered in Chapter 2, this volume, are in a sense nanomachines that do mechanical work through the hydrolysis of ATP, the energy source of the cell (Figure 2.24). During hydrolysis of ATP, a shape change occurs within the motor protein (a mechanochemical process), and mechanical work is performed. The force generated by a motor protein is as large as 100 pN, a very large torque, comparable to a person rotating a 150-m rod. Such a rotary motor, with a diameter of only 30–40 nm, drives the rotation of the flagellum at around 300 Hz at a power level of 10–16 W, with energy conversion efficiency close to 100%. These

protein motors appear to be fuelled by a flow of protons across a pH gradient and can push an average-sized cell at 30,000 nm/s, or 15 body-lengths per second—and can also run in reverse! The structural designs and functional mechanisms revealed in the complex machinery of the bacterial flagellum could provide many novel technologies that may become the basis for future nanotechnology, from which we should be able to find many useful applications.

In comparison, the surface micromachined motors of Volume II, Chapter 7 (see Figure 7.64) are hopelessly impractical. The biological motors are more error-insensitive, with simpler mechanisms and smaller numbers of parts, and wherever possible, flexible members rather than rigid ones are adopted.

In biology, it is important to note that the compartmentalization in cells constitutes an efficient water management system. Since neither proteins nor DNA functions in a dry environment, water needs to be actively controlled by the cell wall. It is no coincidence that more and more of the nanotechnology is focusing on nanochemistry (wet) instead of nanofabrication (dry). In human technology, when working with small droplets the evaporation rate is very high because of the high surface-to-volume ratio, requiring a water management system. Similarly, when working with natural polymers, results can only be obtained when working in an aqueous environment.

Self-Replication and the Use of Templates

We defined self-assembly and self-organization earlier in this chapter: self-replication is yet different. Self-replication is a process in which an object makes a copy of itself. Biological cells, we saw in Chapter 2, this volume, will, given a suitable environment, reproduce by cell division. During cell division, DNA is replicated and is transmitted to offspring during reproduction. Biological viruses can reproduce, but only by hijacking the reproductive machinery of cells. Computer viruses reproduce using the hardware and software already present on computers. Memes reproduce using the human mind and culture as their reproductive machinery. In self-replication, we saw in Chapter 2, this volume, molecules such as DNA encode the genetic information for cells and the resulting organisms, thus allowing the information to be passed on from cell to cell and from generation to generation.[10] PCR and research results with RNA and certain peptides, described in Chapter 2, this volume, are manifestations of this self-replication capability of biological molecules.

Example of Self-Replication in Human Engineering: Self-Replicating Robots

Self-replicating machines have been the subject of theoretical discussion since the early days of computing and robotics. Only three human made mechanical devices that can replicate themselves have been reported so far. One uses Lego parts assembled in a two-dimensional pattern by moving them along tracks, and another uses an arrangement of wooden tiles that tumble into a new arrangement when given a push. The third and most recent self-replicating machine is the most fascinating one. Hod Lipson, from Cornell Laboratory, made the first robots that self-replicate.[11] His robots are made up of a set of modular cubes called *molecubes;* each cube contains identical machinery and a computer program for replication. These machines are simple compared with biological replication, but they demonstrate that mechanical replication is possible and is not limited to biology. Electromagnets on the face of the cubes allow them to selectively attach to and detach from one another, and a complete robot consists of several cubes linked together. Each cube is divided in half along a long diagonal, which allows a robot composed of many cubes to bend, reconfigure, and manipulate other cubes (Figure 3.12).

A tower of cubes can thus bend itself over at a right angle to pick up another cube (Figure 3.13). To start

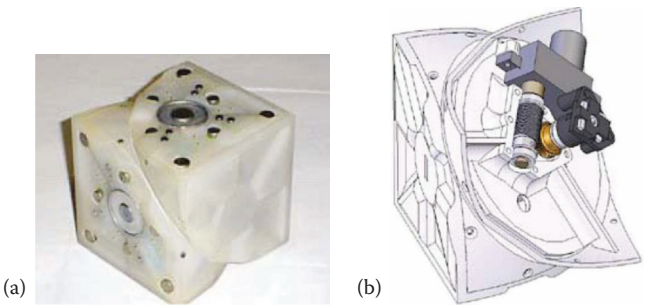


FIGURE 3.12 (a) Each module of the self-replicating robot is a cube about 4 in. on a side, able to swivel along a diagonal. (b) A cutaway drawing shows the motor mechanism.[11]

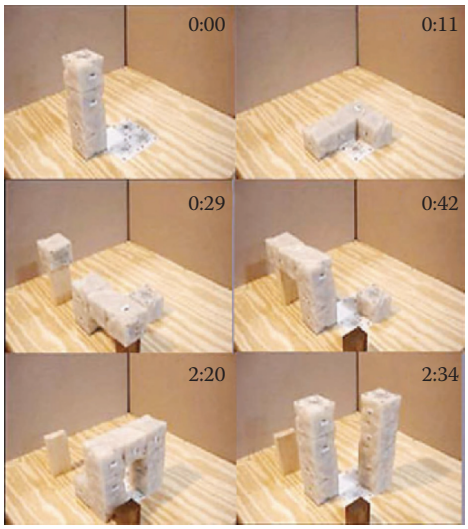


FIGURE 3.13 Frames from a video show the replication process. A robot consisting of a stack of four cubes begins by bending over and depositing one of its cubes on the table. The remaining three cubes pick up additional cubes from "feeding stations" and transfer them to the new robot, which assists in the process by standing itself up. Video is available online at http://www.news.cornell.edu/stories/may05/selfrep.ws.html.

the replication process, a stack of cubes bends over and sets its top cube on the table. Then it bends to one side or another to pick up a new cube and deposit it on top of the first cube on the table. Repeating this process, one robot made up of a stack of cubes creates another just like itself. One robot cannot reach across another robot of the same height. The robot being built assists in completing its own construction. These replicating robots are very dependent on their environment; they draw power through contacts on the surface of the table and cannot replicate unless the experimenters "feed" them by supplying additional modules.

Supramolecular Chemistry

Chemistry beyond the molecule gave rise to supramolecular chemistry—a new field of chemistry pioneered by researchers such as Nobel laureate Jean-Marie Lehn from the University of Strasbourg.[12,13] Self-assembly leads to large supramolecular architectures: no chemist needs to specify the individual atom or molecule positions in these systems.[14] For example, after disintegration of ribosomes (see Figure 2.12 for a ribosome in action) into their protein and RNA building blocks, the molecules will reorganize into functional units in a test tube given the right conditions.

An early suggestion of supramolecular chemistry came from Nobel laureate Hermann Emil Fischer who, in 1890, suggested that enzyme-substrate interactions take the form of a "lock-and-key,"* implying the concepts of molecular recognition and host-guest chemistry. The same lock-and-key assembly is involved in neurotransmitters assembling with molecular recognition sites on proteins that elicit chemical responses, for example, acetylcholine recognition at nerve-muscle junctions, resulting in muscle contractions.[15] The fundamental concepts underpinning supramolecular science are processes in which atoms, molecules, aggregates of molecules, and components arrange themselves into ordered, functioning entities without human intervention.[16] In the early twentieth century noncovalent bonds were understood in gradually more detail. The weak interactions typically involved are as follows (see Volume I, Chapter 7, for more details):

- Hydrogen bonds (described by Latimer and Rodebush in 1920)
- Electrostatic attractions
- Van der Waals forces (the existence of intermolecular forces was first postulated by Johannes Diderik van der Waals in 1873)
- Hydrophobic interactions
- Weak covalent bonds

Through supramolecular chemistry with weak ionic and hydrogen bonds and van der Waals interactions (0.1–5 kcal/mol), structures with dimensions of 1–100 nm are possible. For chemical synthesis, the largest of these dimensions often is too large, and for microfabrication, it often is too small. The types of interactions listed above allowed scientists to finally interpret protein and DNA behavior. Chemists started to study synthetic structures based on no covalent interactions, such as micelles and microemulsions. Eventually, these concepts, derived from

* The lock-and-key concept of biochemical interactions has, over the years, been modified in favor of a "hand-in-a-glove" concept, because some reorganization (conformational change) is required to make two biomolecules complementary to each other.

biological systems, were applied to synthetic chemical systems. The first big breakthrough along this line came in the 1960s with the synthesis of the crown ethers by Charles J. Pedersen.[17] Donald J. Cram, Jean-Marie Lehn, and Fritz Vögtle became active in the synthesis of shape- and ion-selective receptors, and throughout the 1980s research in the area gathered a rapid pace, with concepts such as mechanically interlocked molecular architectures emerging. The emerging science of nanotechnology also had a strong influence on the subject, with building blocks such as fullerenes, nanoparticles, and dendrimers becoming involved in synthetic systems. By the 1990s, supramolecular chemistry had become an established avenue toward nanoelectronic and nanomechanical systems, with researchers such as Fraser Stoddart developing molecular machinery and highly complex self-assembled structures[18] and researchers such as Itamar Willner developing sensors and methods of electronic and biological interfacing.[19] Stoddart pioneered the development of the use of molecular recognition and self-assembly processes in template-directed protocols for the syntheses of two-state mechanically interlocked compounds (bistable catenanes and rotaxanes) that have been employed as molecular switches and as motor-molecules in the fabrication of nanoelectronic devices and NEMS (see Figure 3.14 for a rotaxane switch).[20] Rotaxanes can act as molecular switches because they can oscillate back and forth between two stable states.

Yet more recently, electrochemical and photochemical motifs have become integrated into supramolecular systems in order to increase functionality, research into synthetic self-replicating system began, and work on molecular information processing devices began. For example, the Stoddart and Zink groups have developed donor-chromophore-acceptor-based molecular triads that can transduce light into electrical energy by mimicking the photosynthesis pathway.[21] A nanoscale power supply in the form of a light-harvesting molecular triad has been developed (Figure 3.15). Donor-chromophore-acceptor-based molecular triads that can convert light to electrical energy are some of the most effective molecules that can serve the purpose of energy transduction. Yet more impressively, it has been demonstrated that the photocurrent generated by the triad can be utilized to drive a supramolecular machine in the form of a pseudorotaxane.[22] The molecular triad is based on tetrathiafulvalene-porphyrin-fullerene (TTF–P–C60), which generates electrical current by harnessing light energy when self-assembled onto gold electrodes. The triad is composed of three unique electroactive components: 1) an electron donating TTF unit, 2) a chromophoric porphyrin unit, and 3) an electron-accepting C60 unit. A disulfide-based anchoring group was tagged to the TTF end of the molecule in order to allow self-assembly onto gold surfaces. In a closed electrical circuit, the triad-functionalized working electrode generates a switchable photocurrent of ~1.5 μA/cm^2, when irradiated with a 413 nm krypton-ion laser. The electrical energy generated by the triad is then exploited to drive a supramolecular machine in the form of a [2]pseudorotaxane composed of a π-electron-deficient tetracationic cyclobis(paraquat-

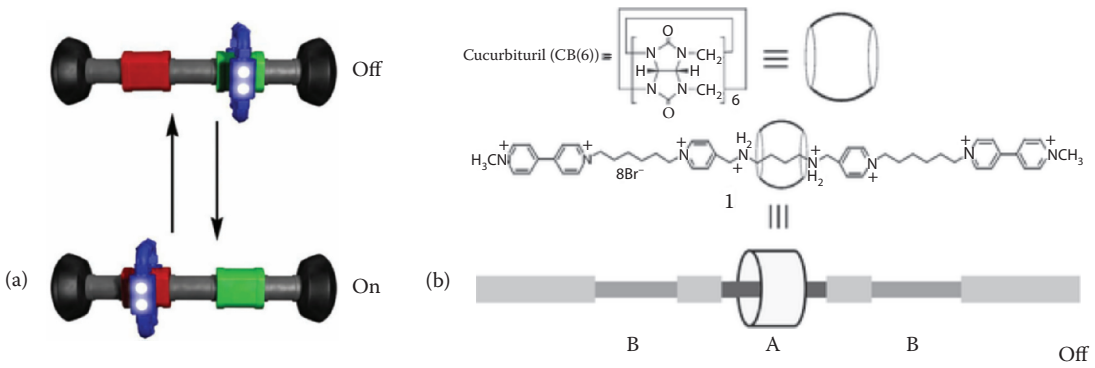


FIGURE 3.14 Rotaxene switch. (a) Model; (b) molecular structure and another model. A **cucurbituril** is a macrocyclic molecule consisting of several glycoluril [=$C_4H_2N_4O_2$=] repeat units, each joined to the next one by two methylene [-CH_2-] bridges to form a closed band. Cucurbiturils are commonly written as **cucurbit[*n*]uril**, where *n* is the number of repeat units. A common abbreviation is **CB[*n*]**. Adapted from Stoddart and Nepogodiev.[20]

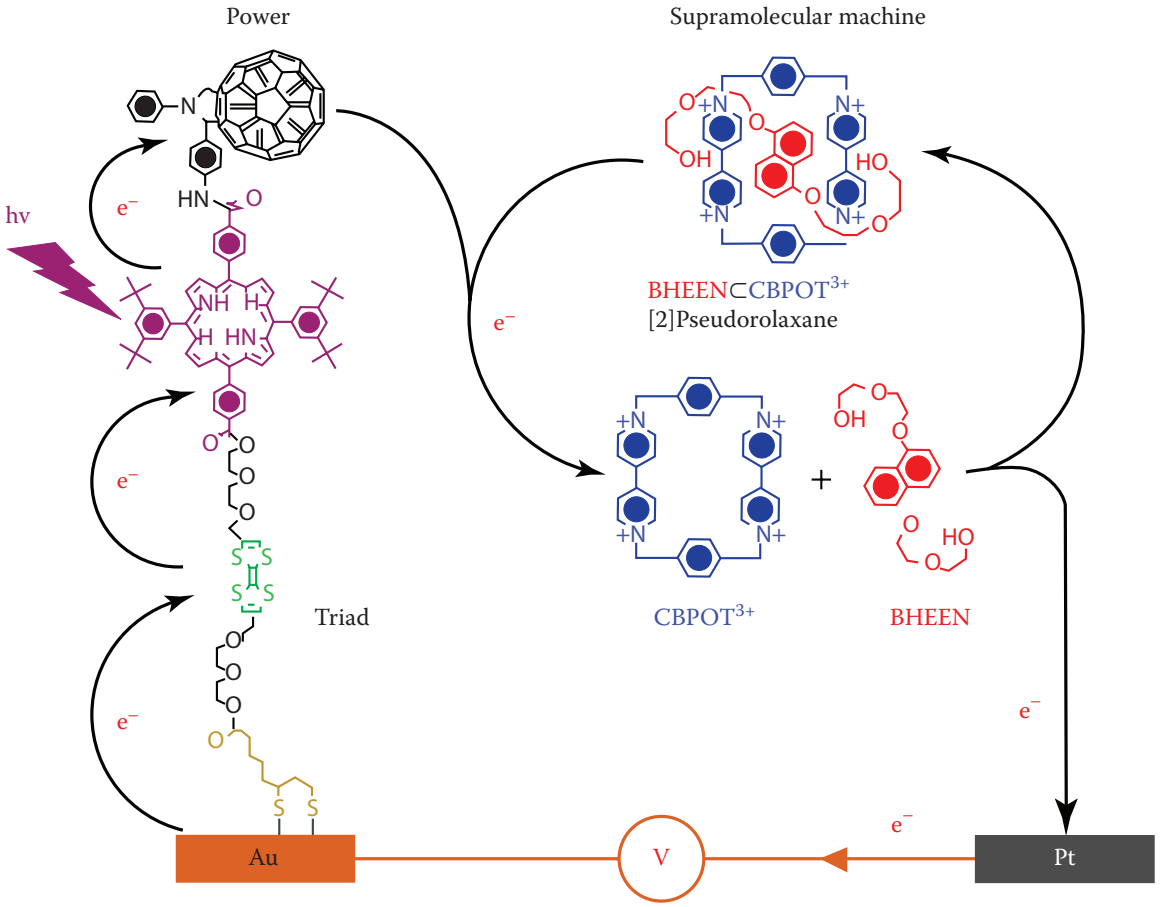


FIGURE 3.15 The photocurrent generated by the triad can be utilized to drive a supramolecular machine in the form of a pseudorotaxane. The [2]pseudorotaxane is composed of a π-electron-deficient tetracationic cyclobis (paraquat-*p*-phenylene) (CBPQT4+) cyclophane and a π-electron-rich 1,5-bis[(2-hydroxy-ethoxy)-ethoxy] naphthalene (BHEEN) thread. Adapted from Saha et al. 2005. Powering a supramolecular machine with a photoactive molecular triad. *Small* 1:87–90.[22]

p-phenylene) (CBPQT4+) cyclophane and a π-electron-rich 1,5-bis[(2-hydroxy-ethoxy)-ethoxy] naphthalene (BHEEN) thread (Figure 3.15). The dethreading of CBPQT4+ cyclophane from the BHEEN thread was monitored by measuring the increase in the fluorescence intensity of the BHEEN unit. A gradual increase in the fluorescence intensity of the BHEEN upon illumination confirmed that the dethreading process is driven by the photocurrent generated by the triad.

Dendrimers form a special set of molecules in supramolecular chemistry. They are globular polymer-based macromolecules of monodisperse size in which all bonds emerge treelike from a central core in a regular branching pattern and may be as small as 3 nm across (the size of a typical protein). These fractal-like macromolecules, introduced by Tomalia et al.[23] and Newkome et al.[24] in 1985 have attracted a lot of attention because of their novel structure and intriguing properties.[25,26] Dendrimer synthesis allows for nearly complete control over the critical molecular design parameters, including size, shape, reaction site distribution, surface/interior chemistry (e.g., number and size of internal voids), flexibility, and topology. The process involves a stepwise, repetitive sequence that guarantees completed reactions for each generation of reaction products, resulting in monodisperse final products. There are three major strategies involved in effective dendrimer synthesis: the starburst divergent strategy, the convergent growth strategy, and the self-assembly strategy. In the divergent approach, growth starts at the core and proceeds radially outward toward the periphery, and in the convergent approach, growth starts at what will become the periphery of the dendrimer and proceeds inward. Although the majority of dendrimers prepared to date have been constructed of covalent bonds[25,27] using convergent or divergent

strategies, several noncovalent dendrimers[25,28] have been prepared by a variety of self-assembly processes involving, for example, hydrogen bonding[29] or supramolecular coordination chemistry.[30]

Balzani and coworkers[31,32] described the use of metal branching coordination centers, such as ruthenium and osmium, and multidentate ligands for the construction of metallodendrimers centers more than a decade ago. Since this work, a large number of self-assembled dendritic structures have been prepared by using metal ligand complexes (Figure 3.16a), hydrogen bonding moieties (Figure 3.16b), or ionic interactions (Figure 3.16c) for the assembly.

Figure 3.16a shows a small bimetallic dendrimer obtained in a flexible synthesis strategy that is used to prepare many larger dendrimers with a variety of arrangements of the internal and peripheral metal centers.[32] These polynuclear complexes may be redox-active and/or luminescent, can be used as electrochemical and/or photochemical molecular devices, and may prove useful in areas such as energy harvesting and conversion, signaling, and diagnostics. Among the numerous self-assembling dendritic systems that have been reported, those involving assembly into columnar or spherical structures are particularly intriguing. Percec and coworkers[33,34] used van der Waals interactions and H-bonding to self-assemble functionalized polyether dendrons into cylindrical columnar or spherical assemblies that mimic the shape and architecture of certain types of viruses. Similarly, Aida and coworkers[35] have used weak metal-metal interactions to effect the hierarchical self-assembly of small dendrimers into luminescent superhelical fibers. Given their sizes, optical properties, and electronic properties, such assemblies could well find display, storage, or other applications in micro- and nanoelectronics.

As summarized in Figure 3.17, dendrimers can be designed to aggregate to form cylinders or spheres depending upon the nature of the fundamental building unit. Pie-shaped dendrimers assemble into cylinder segments (donuts) that then further aggregate into cylinders (Figure 3.17a). Cone-type dendrimers organize into spheres (two to twelve per sphere depending on generation), which then pack into the unusual Pm3-n (A-15)-type cubic crystal structure having spheres at the corners and body center of the cube plus two on each face (Figure 3.17b).

The layered architecture of dendrimers, their globular shape, highly controlled monodisperse size,

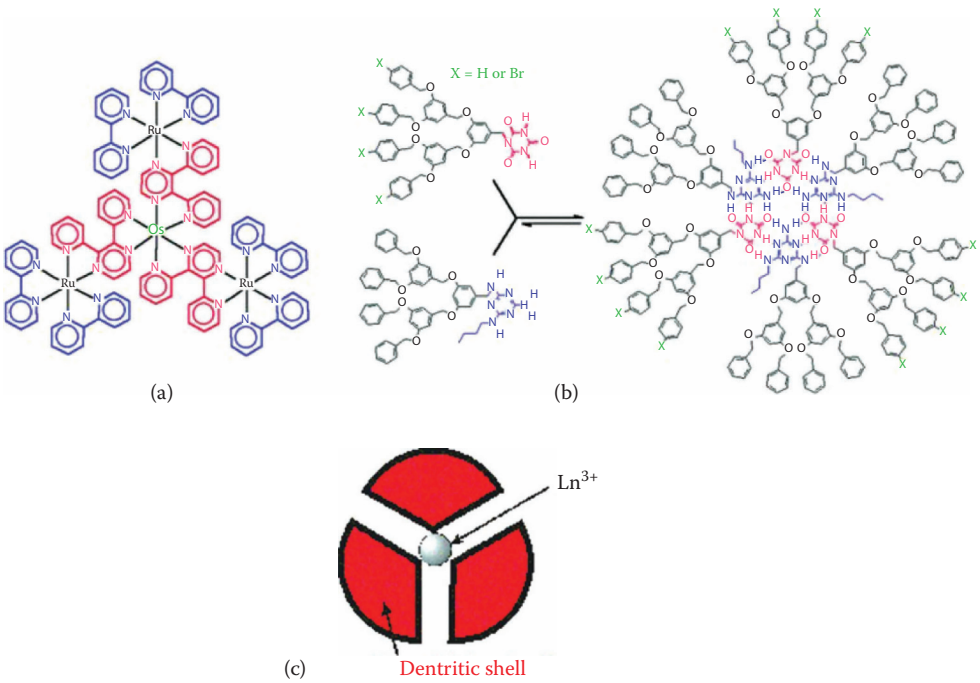


FIGURE 3.16 (a) Self-assembled metallodendrimer. (b) Self-assembly of dendrimer by hydrogen bonding. (c) Schematic representation of self-assembled ionic dendrimer. (From Denti, S., et al. 1992. *J Am Chem Soc* 114:2944–50; Balzani, V., et al. 1996. *Chem Rev* 96:759–833; Fréchet, J.M.J. 2002. *Proc Natl Acad Sci USA* 99:4782–87. [31,32,36])

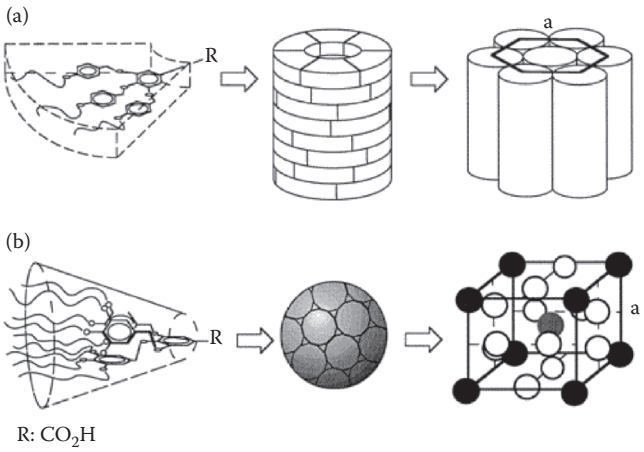


FIGURE 3.17 Assembling cylinders (a) and spheres (b) from dendritic molecules.[37]

radially controlled chemical makeup, multivalent periphery, variable inner volume, and controlled intramolecular dynamics endow dendrimers with unique features such as the ability to morph into a variety of supramolecular arrangements in response to external stimuli. On the basis of these properties, dendrimers may provide the key to reliable and economical fabrication of functional nanoscale materials with unique electronic, optical, optoelectronic, magnetic, chemical, or biological properties. Some of these applications are already being researched. Specially designed dendrimers with a hydrophobic interior and a hydrophilic[38] periphery, so-called "unimolecular micelles," are capable of molecular inclusion and solubilize hydrophobic molecules, such as pyrene in aqueous solution, and are at least as good at it than traditional micelles; moreover, they do not display a critical micelle concentration.[39,40] Classic micelles are thermodynamic aggregates of amphiphilic molecules and therefore dynamic assemblies of small molecules, while "unimolecular micelles" are static and retain their cohesion regardless of concentration. Also, dendrimer monolayers are finding a number of applications from surface-confined chemical sensor arrays[41,42] or affinity biosensors[43,44] to resists for nanolithography.[38,45]

Molecular Assembler

A molecular assembler was defined by K. Eric Drexler as a "device able to guide chemical reactions by positioning reactive molecules with atomic precision."[46] Some biological molecules, such as ribosomes, constructing a wide variety of proteins from their amino acid constituents, fit this definition (see Figure 2.12). However, the term *molecular assembler*, as defined by Drexel, refers to artificial mechanical machines, envisioned to work in air rather than in an aqueous solution. Molecular assemblers are broader than self-replicating machines (see above and Figures 3.12 and 3.13), as in theory they can both self-replicate and make other products. Drexel logically pointed out that if self-replication of molecular assemblers is not restrained, it might lead to competition with naturally occurring organisms. This is referred to as the "gray goo" scenario.[47] One method to construct molecular assemblers in solutions is to mimic natural evolutionary processes. As a result, production of complex molecular assemblers might evolve from simpler systems.

One of the most outspoken critics of the mechanical "molecular assemblers" was Richard Smalley (1943–2005). Smalley argued that such assemblers were not physically possible and introduced two major objections to them, termed the "fat fingers problem" and the "sticky fingers problem."[48] He believed that these two problems would exclude the possibility of molecular assemblers that worked by precision picking and placing of individual atoms.

In the fat fingers problem, Smalley argues that in an ordinary chemical reaction, five to fifteen atoms near the reaction site engage in an intricate three-dimensional waltz in a very cramped region of space measuring no more than a nanometer on each side, so there is not enough room in the nanometer-size reaction region to accommodate all the fingers of the mechanical manipulators necessary to exert complete control of the chemistry. Drexel's counterargument is that two reactants can be brought together with controlled trajectories if one reactant is bound to a substrate and the second reactant is positioned and moved by a single finger. He correctly points out that this has already been done experimentally by using the tip of an STM.[49] For example, Ho and Lee[50] physically bound a CO molecule to an iron atom on a silver substrate using an STM (see Volume II, Figures 2.44 and 2.45, for more complex STM molecular assembly). Drexel also points out that if steric constraints near the tool tip do make

it too difficult to manipulate particular individual atoms or small molecules with sufficient reliability, a simple alternative is to use conventional solution or gas phase chemistry for the bulk synthesis of somewhat larger nanoparts, consisting of 10–100 atoms. These larger nanoparts could then be bound to a positional mechanical device and assembled into larger structures without further significant steric constraints. This is the approach taken by the ribosome in the synthesis of proteins, where individual amino acids are sequentially assembled into an atomically precise polypeptide without the need to manipulate individual atoms. *Atomically precise* is a description of the precision of the final product, not a description of the manufacturing method. It must be argued though that once we are dealing with a cluster of atoms the original objective of atomic scale control has been abandoned and a ribosome still recognizes the individual base pairs to stitch the protein together from its amino acid components.

Smalley also advances the sticky fingers problem, which is the claim that "the atoms of the manipulator tool will adhere to the atom that is being moved. So it will often be impossible to release this minuscule building block in precisely the right spot."[48] Drexel's counterargument is that the sticky fingers problem is not a fundamental barrier to building mechanical assemblers since reactions exist that allow the synthesis of a useful range of precise molecular structures. As an example, he mentions that the application of a voltage between a manipulator tool and the workpiece that cause the target atom or moiety to move to the desired position. He points again to work by Ho and Lee[50] for experimental proof.

A final criticism by Smalley was that these mechanical assemblers would be far too slow to be ever practical.[48] Here, a counterargument could be the possible use of arrays of assemblers, as we saw in the IBM Millipede (see Volume II, Figure 2.4b).

All in all, the Smalley-Drexel debate is about how to make molecular assemblers, not whether they are possible or not. To push either extreme, i.e., to claim that only biology/chemistry can do it (Smalley) or to put too much emphasis on mechanical assembly (Drexel), is to ignore that the solution might well be one that incorporates a combination of both. Drexel is indeed pointing at chemical synthesis (up to a nanocluster) and then mechanical positioning for further assembling of these clusters as a possible solution.

Nanostructure Properties

Introduction

In this section we introduce all the different properties that nanostructures bring to the table that do differ from microstructure properties. On the broadest level the differences may be summarized as shown in Table 3.2. Nanostructures occupy very high surface area and have a vastly increased surface area per unit mass, e.g., upwards of 1000 m^2 per gram. In terms of suface-to-volume (S/V) ratio, a 3-nm iron particle has 50% of its atoms on the surface, a 10-nm particle has 20%, and a 30 nm particle only 5% (see Volume I, Figure 2.49). Nanoparticles are quantum mechanical, wave-like, with discrete energies and quantized currents. Quantum size effects result in unique mechanical, electronic, photonic, and magnetic properties. The chemical reactivity of nanoscale materials greatly differs from that of more macroscopic forms. Microfabricated microstructures, on the other hand, exhibit relatively small S/V, behave mostly classically (semiclassically) and

Table 3.2 Comparison of Microstructures and Nanostructures

	Microstructure	Nanostructure
Physics	Semiclassic	Quantum mechanical
Electron nature	Particle-like	Wave-like
E- or k-space	Continuous	Discrete
Current	Continuous	Quantized
Decision	Deterministic	Probabilistic
Fabrication method	Microfabrication	Nanofabrication
Surface area/volume	Small	Very large
Packing density	Low	Very high

are particle-like, with continuous energy and currents. All the properties of nanoparticles we mention below are ultimately a consequence of the more fundamental ones listed in Table 3.2.

Melting Temperature, Thermal Transport, and Heat Capacity

In Chapter 7, this volume, on scaling, we learn that the surface-to-volume (S/V) ratio of particles scales as $1/r$, where r is the characteristic dimension of the particles. The smaller a particle, the more of its atoms find themselves at it surface. A bulk solid material will typically have less than 1% of its atoms on its surface, whereas 10-nm particles have about 15% of surface atoms. The high S/V ratio of nanoparticles makes them more reactive as catalysts in chemical reactions and lowers their melting temperature (T_m). There is an inverse linear relationship between melting temperature and the surface-to-volume ratio. This makes sense since atoms on a surface are more easily accessed and rearranged than atoms in the bulk. As a consequence, the melting temperature of particles is always lower than the bulk. As observed from Figure 3.18 the melting point of gold particles decreases dramatically as the particle size gets below 5 nm. For nanoparticles embedded in a matrix, melting point may be lower or higher, depending on the strength of the interaction between the particles and the matrix.

Macroscopic thermal transport theory, with Fourier's law (Volume I, Equation 3.53) and the diffusion equation (Volume I, Equation 3.54), breaks down for applications involving thermal transport in NEMS. Whereas for bulk materials, interface effects (surface or grain boundaries) dominate only at low temperatures, in nanometer-scale thin films and particles, the room temperature thermal conductivity can be significantly modified. One can think about the reduction in thermal conductivity in these structures as corresponding to a lower mean free path for the energy flux. In the case of conductors, electrons dominate the thermal conductivity and the mean free path depends on electron scattering phenomena. At lower temperatures, the electron mean free path (λ_e) is constant and controlled by defect scattering, whereas at higher temperatures, phonon scattering takes over and the mean free electron path falls off, as shown in Volume I, Figure 3.161, where we plot mean free electron path (λ_e) versus temperature. From Volume I, Figure 3.162, plotting electronic thermal conductivity (κ_e) versus temperature for Cu and Al, we see that thermal conductivity increases with temperature in the defect scattering regime and then goes down with temperature increase when phonon conduction takes over at higher temperatures. Phonons dominate the thermal conductivity in insulators. Phonons, like electrons, scatter at boundaries, defects, and other phonons (phonon-phonon scattering). For nanodevices the device boundary might take over from defect control. Scattering of heat flow at any of these interfaces significantly reduces thermal conductivity. The size effect on the thermal conductivity can exceed two orders of magnitude for layers of thickness near 1 μm at $T < 10$ K. Figure 3.19 illustrates this for three different thin-film Si layers (thickness d_s) versus bulk Si.[52]

Heat capacity (C_v) of a material is related to its atomic structure or its vibrational and configurational

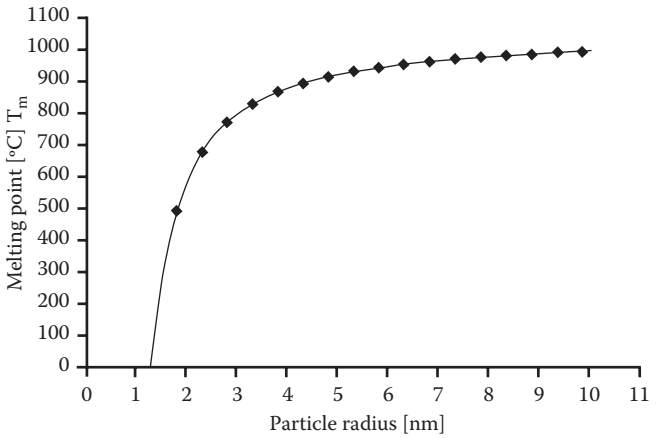


FIGURE 3.18 Melting temperature (T_m) of gold particles as a function of their radius.[51]

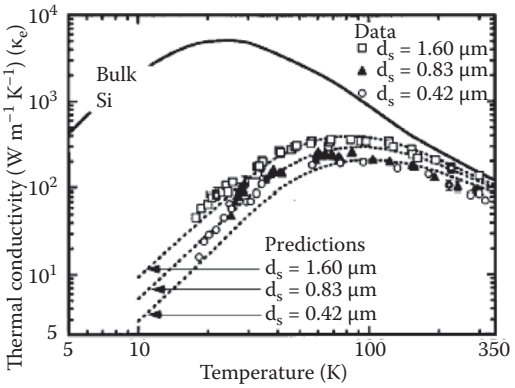


FIGURE 3.19 Thermal conductivity (κ_e) of thin Si films versus bulk Si. d_s is the Si layer thickness.[52]

entropy, which is controlled by all the possible nearest-neighbor configurations. Nanocrystalline materials feature ultrafine crystalline grains. A large fraction of atoms are located in the metastable grain boundaries in which the nearest-neighbor configurations are much different from those in the crystallites. The grain-boundary possesses an excess volume with respect to the perfect crystal lattice, therefore, heat capacities (C_v) of the nanocrystalline materials are expected to be higher than those of the corresponding coarse-grained polycrystalline counterparts.

For nanocrystalline materials, the C_v values, in practice, are indeed observed to be higher than their bulk counterparts. For example:

Pd: 48% up from 25 to 37 J/mol·K at 250 K for 6 nm crystalline
Cu: 8.3% up from 24 to 26 J/mol·K at 250 K for 8 nm
Ru: 22% up from 23 to 28 J/mol·K at 250 K for 6 nm

The phonon contribution to the molar specific heat capacity (C_v) of these polycrystalline materials, we saw in Volume I, Chapter 3, is given by the Dulong-Petit law (Volume I, Figure 3.15), with values for different materials approaching $C_v = 3R$ or 25 J/mol·K at high temperatures. As calculated by Debye, a constant value for C_v is reached at $T \gg \theta_D$ and for $T \ll \theta_D$ (where θ_D is the Debye temperature) he found:

$$C_v = \frac{12}{5}\pi^4 R\left(\frac{T}{\theta_D}\right)^3$$

(Volume I, Equation 3.331)

A T^3 temperature dependency for the specific heat stems from a T^4 dependency of the internal energy U ($C_v = \frac{dU}{dT}$). Since the density of phonon states, [$G(E)$], depends on the dimensionality of the structure, the specific heat of the lattice also depends on the dimensionality of the structure (3D, 2D, 1D, or 0D). Generalizing for different dimensions (D), one derives:

$$U \propto T^{D+1} \text{ and } C_{v,lat} \propto T^D$$

(Volume I, Equation 3.338)

Besides phonon contributions, if the material is an electronic conductor, the electronic contribution to the heat capacity must be counted as well. The electronic specific heat capacity for a 3D structure we introduced in Volume I, Chapter 3 as follows.

$$C_{v,el} = \frac{dU}{dT} = \frac{\pi^2}{3}k_B^2 TG(E = E_F)_{3D} = \frac{\pi^2}{2}\left(\frac{k_B T}{E_F}\right)n_{3D}k_B$$

(Volume I, Equation 3.307)

This expression must be modified again for different dimensionalities since the internal energy (U) depends on the density of electronic states $G(E)_{\#D}$ and $C_{v,el} = \frac{dU}{dT}$.

Thermoelectric Power

In the 1960s, semiconductor thermoelectrics were considered for replacement of mechanical refrigerators and diesel generators. The idea was that if transistors could replace vacuum tubes for electronics, then surely thermoelectrics could replace those outdated machines. The efficiency of cooling and heating is controlled by a dimensionless parameter Z given by the following equation:

$$ZT = \frac{S^2\sigma T}{\kappa} \tag{8.78}$$

where T is the absolute temperature; σ and κ are the electrical and total thermal conductivities (lattice and electrons) of the material, respectively; and S is the Seebeck coefficient. Consequently, requirements for a good thermoelectric material are a high figure of merit ZT; this in turn requires a high Seebeck coefficient (S), a high electrical conductivity (σ), and a low thermal conductivity (κ). For the best thermoelectric materials, one has a ZT of ~0.9, which remained the upper limit for many years. The biggest problem was that ZT had to be improved from ~0.9 to 3 or higher to become commercially attractive [ZT needs to improve to >1.3 at 300 K for a major impact in electronics cooling and to around 2.5 for a revolutionary impact in air-conditioning and power from waste heat (energy scavenging)]. Minimizing thermal conductivity and maximizing electrical conductivity has been the biggest dilemma for the last forty years. Dresselhaus et al. show that

lower dimensional structures such as nanowires based on Bi, BiSb, Bi_2Te_3, and SiGe nanotechnology give additional controls and can lead to a higher *ZT* (see Chapter 8, this volume).[53] Combining a superlattice with a nanowire structure enhances thermopower because of the sharper density of states than ordinary 1D nanowires (enhances *S*) and reduction of the lattice thermal conductivity by increasing the phonon scattering at the segment interfaces (decreases κ). The reason is that the higher density of electron states at the Fermi level increases σ, the increased phonon-boundary scattering lowers κ, and both factors contribute to making *ZT* higher. Figure 8.32 shows the thermoelectric figure of merit *ZT* of a Bi quantum well and Bi quantum wire as a function of size.

In 2008, two teams, one at Caltech and one at the University of California, Berkeley, separately reported that they could increase Si's ability to convert heat into electric current by as much as 100 times[54,55] (see also http://www.spectrum.ieee.org/jan08/5879). Bulk Si used to be a poor material for thermoelectric conversion, because its thermal conductivity is too high; heat travels across it too well so that it is difficult to create the necessary temperature gradient. Bulk Si at room temperature has a *ZT* of only 0.01, but the Berkeley team increased that to 0.4, and the Caltech team increased it to 0.6. This puts Si nanowires almost on par with bismuth telluride. Making thermoelectric devices out of Si, which is much more abundant, cheaper, and easier to handle than bismuth telluride, could help create a large market for these devices.

By fashioning Si into nanowires with diameters of 10–100 nm and introducing defects in the wires that slow the flow of phonons, both teams have found that they can decrease silicon's thermal conductivity without losing too much of its electronic conductivity, and in doing so increase *ZT*. Defects at three different length scales were engineered into the nanowires. First, with silicon in the shape of nanowires, the material is small enough compared with the phonons that the size of the wires themselves affects how the phonons move. Secondly, the surface of the nanowires is made rough, introducing a set of defects at a smaller scale. Finally, the silicon nanowires are doped with boron to introduce defects at the atomic level. Heath at Caltech induced a greater drop in thermal conductivity by making his nanowires even smaller than Yang's group at Berkeley—only 10–20 nm in diameter.[55] Normally, a wire can carry two types of phonons: one that causes the wire's diameter to expand or contract, and one that causes it to lengthen or shorten. But when the nanowires get small enough, these two vibration modes merge into a single type of phonon, which slows down the heat transport even more. Unfortunately, using wires only 10 nm wide may give the best results for thermal conductivity, but the electrical conductivity crucial to thermoelectric conversion also drops.

One of the applications for these silicon nanowires is for recycling waste heat from computer chips into electricity, i.e., energy scavenging (see Figure 3.20).

Mechanical Properties

Small grains, over a certain size range, lead to improved strength and hardness in lightweight nanocomposites and nanomaterials, altering bending and compression properties. This mechanical change is explained as follows. Dislocations are blocked by grain boundaries, so slip is blocked, and therefore, since the smaller the grain size the larger the surface of the grain boundaries, a larger elastic limit is obtained. The latter is expressed in the empirical Hall-Petch equation representing the maximum elastic yield strength (σ_Y = stress at which the material permanently deforms) of a polycrystalline material as a function of grain size. As we learned

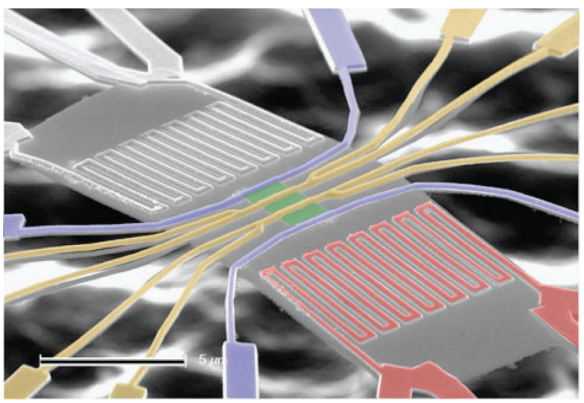


FIGURE 3.20 A nanofurnace: an array of nanowires (green) convert heat from the temperature difference between two slivers of a microchip. Current flowing in through a heater (red) causes the temperature difference. (From James Heath, http://www.spectrum.ieee.org/jan08/5879.)

in Volume I, Chapter 2, Hall and Petch, working independently, found that the yield strength of a polycrystalline material follows the following dependence:

$$\sigma_Y = \sigma_0 + \frac{K}{\sqrt{d}}$$

(Volume I, Equation 2.70)

where d is the average diameter of the grains in micrometers, with σ_0 the frictionless stress (N/m^2) and K a constant. The strength of a material thus depends on grain size (see Table 3.3). The strength of a material thus depends on grain size. In a small grain, a dislocation gets to the boundary and stops, i.e., slip stops. In a large grain, the dislocation can travel farther. So small grain size equates to more strength. For example, the elastic limit of copper doubles when the grain size falls from 100 µm to 25 µm. Instead of yield strength, one might also plot the hardness (H) of the material as a function of grain size, which creates a similar relationship.

The Hall-Petch relation holds for conventional grain sizes (from 0.1 to 100 µm) but for grains smaller than 20 nm, this relation breaks down (see Figure 3.21). For nanometer grains, a reverse Hall-Petch effect has actually been observed, i.e., the strength and hardness of materials from a small grain size on starts to decrease with decreasing grain size. In other words, an optimal grain size (d_c) exists, as suggested by the plot in Figure 3.21 (see also Volume II, Figure 7.54). The classic Hall-Petch relationship is based on the idea that grain boundaries act as obstacles to dislocation movement and since dislocations are carriers of plastic deformation, this manifests itself macroscopically as an increase in material strength. The Hall-Petch behavior breaks down when the smallest dislocation loop no longer fits inside a grain. The deformation mechanism for materials with very small grains (<20 nm)

TABLE 3.3 The Strength of a Material Depends on Grain Size

Material	Yield Strength
Cold worked copper	393 MPa
400-nm copper	443 MPa
100-nm nanograin copper	900 MPa
10-nm nanograin copper	2.9 GPa

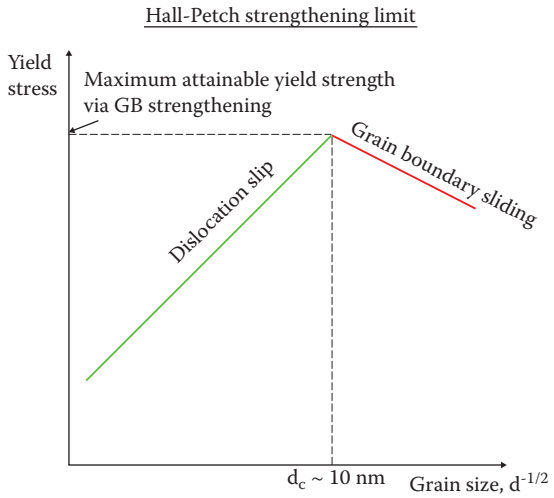


FIGURE 3.21 Hall-Petch reversal.

is different and it has been suggested that plastic deformation in this case is no longer dominated by dislocation motion and it is believed that individual atoms migrate, diffuse, and slide along grain boundaries and through triple junctions (Y-shaped grain boundary intersections).

The increase of the surface-to-volume (S/V) ratio in nanoparticles leads to an increasing dominance of the behavior of the surface atoms over those in the interior of the particle. This results in more interactions between the intermixed components in ceramic nanocomposites; those nanophase powders compact more easily in the sintering process, and successful sintering enhances their hardness.

Chemical Effects

Surface chemistry effects that are barely noticeable in large particle systems become overwhelming in nanoparticle systems that demonstrate enhanced reactivity due to their unusual adsorptive properties. In bulk solids, all molecules are surrounded by and bound to neighboring atoms, and forces are in balance. Surface atoms on the other hand are bound only on one side, leaving unbalanced atomic and molecular forces on the surface. These forces attract gases and molecules through the van der Waals force, and at low temperatures this leads to physical adsorption or physisorption. At high temperatures, electron sharing or valence bonding with gas atoms, resulting in chemical adsorption or chemisorptions, may satisfy those same unbalanced surface

forces. Chemisorption is a much stronger interaction than physisorption, with a heat of adsorption up to 800 KJ/mol. Adsorbing gas or vapor molecule splits into atoms, radicals, or ions, which form a chemical bond with the adsorption site. These processes are the basis for heterogeneous catalysis (key to production of fertilizers, pharmaceuticals, synthetic fibers, solvents, surfactants, gasolines, other fuels, and automobile catalytic converters). Nanostructured materials offer many advantages for this type of catalysis applications including the following:

- Huge surface area, high proportion of atoms on the surface
- Enhanced intrinsic chemical activity as size gets smaller, for example, likely due to changes in crystal shape
- When the shape changes from cubic to polyhedral, the number of edges and corner sites goes up significantly
- As crystal size gets smaller, anion/cation vacancies can increase, thus affecting surface energy; also, surface atoms can be distorted in their bonding patterns
- Enhanced solubility, sintering at lower *T*, more adsorptive capacity

All the properties discussed so far, unlike the quantum effects that we turn our attention to next, change gradually as the dimensions of the particles change.

Blue Shift in Semiconductor Nanoparticles

In Volume I, Chapter 3, we saw that whenever electrons are confined to a small space their behavior as tiny quantum wave packets takes precedence over their particle behavior. For a particle confined to a region of the *x*-axis of length Δx, the uncertainty in momentum is given by $\Delta p_x \sim \hbar/\Delta x$ (see Equation 3.106). If the particle has a mass *m*, the confinement gives it an additional kinetic energy of magnitude $E_c = (\Delta p_x)^2/2m \sim \hbar^2/2m(\Delta x)^2$ (Volume I, Equation 3.162). The confinement energy becomes significant if it is comparable to or greater than the kinetic energy of the particle due to its thermal motion, i.e., $E_c \sim \hbar^2/2m(\Delta x)^2 > \frac{1}{2}k_BT$, and the quantum size effect will be important if $\Delta x \sim \hbar/(mk_BT)^{1/2} \sim \hbar/p_x \sim \lambda_{DB}$ (Where λ_{DB} is the thermal de Broglie wavelength, i.e., roughly the average de Broglie wavelength of the particles in an ideal gas at the specified temperature, see Volume I, Equation 3.216). Thus, if the particle motion is confined in a region of the order of the particle de Broglie wavelength, one will observe effects of size quantization.

When the de Broglie wavelength of an electron is comparable to the size of the system in one or more directions, since only an integer number of half-waves of the electrons can be put in any finite system, one obtains a set of discrete electron states and energy levels, each of which is characterized by the corresponding number of half-wavelengths. This situation is summarized for 3D, 2D, 1D, and 0D degrees of freedom of electrons in Figure 3.22 (see also Volume I, Chapter 3).

A semiconductor quantum dot (QD) can be regarded as an artificial solid-state atom with delta-function-like density of states. Electrons in a quantum dot arrange themselves as if they were part of an atom, even though there is no nucleus for them to surround. The atom these assembled electrons emulate depends on the number of electrons and the exact geometry of the confining nanoprison. Here we calculate, as an example, the spacing between the lowest two energy levels, $\Delta E = E_{n+1} - E_n$, for a spherical

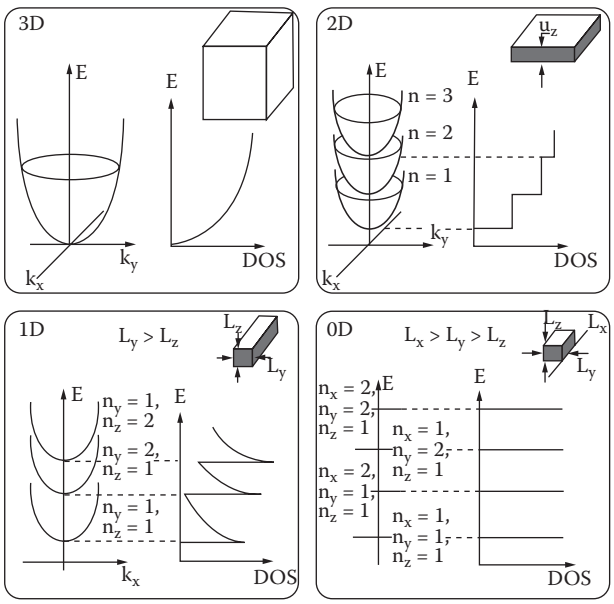


FIGURE 3.22 Quantization in zero (3D), one (2D), two (1D), and three (0D) directions.

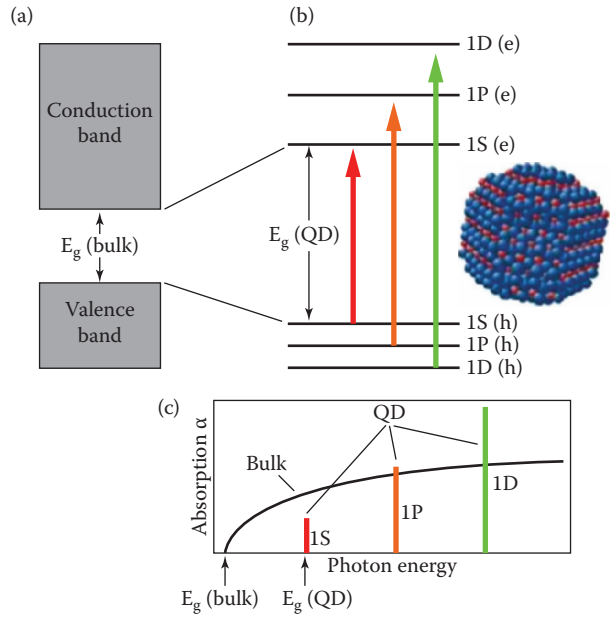


FIGURE 3.23 Blue shift in CdSe quantum dot. (a) CdSe bulk semiconductor. (b) CdSe quantum dot (QD). (c) Quantum dot photon energy absorption spectrum.

CdSe semiconductor quantum dot (see Figure 3.23). Such a spherical nanocrystal may be made from solution using colloid chemistry instead of the epitaxial methods used to make the cubic nanoparticle considered in Volume I, Figure 3.141. The dimensions of this nanocrystal are smaller than the electron de Broglie wavelength or the exciton Bohr radii (see also Volume I, Chapter 3). Semiconductor QDs possess high radiative efficiency, are compatible with semiconductor device technology, and are considered promising elements for implementing the coherent control of the quantum state, which is essential for quantum information processing.

In the case of an electron in a spherical quantum well, the Hamiltonian separates into angular and radial parts, and because of the spherical symmetry this results in the following eigen-energies and eigen-states:

$$E_{n,l,m} = E_{n,l} \quad \text{and} \quad \Psi(r,\theta,\phi) = Y_{l,m}(\theta,\phi)R_{n,l}(r) \qquad (3.1)$$

with $Y_{l,m}(\theta,\phi)$ the spherical harmonics and $R_{n,l}(r)$ the radial wave functions. The energy levels and wave functions do of course depend on the details of the confining potential. If that potential is an infinite spherical well, where $V = 0$ for $r < R$ and $V = \infty$ for $r > R$, then the solutions are the atom-like spherical Bessel functions from Volume I, Figure 3.81, described in the following expressions:

$$E_{n,1} = \frac{\hbar^2 \beta_{n,1}^2}{2m_e^* R^2} \text{ and} \qquad (3.2)$$

$$R_{n,1} = j_1\left(\frac{\beta_{n,1} r}{R}\right) \text{ for } r < R \qquad (3.3)$$

where $j_l(x)$ is the l^{th} spherical Bessel function and the coefficient $\beta_{n,l}$ is the n^{th} zero crossing of $j_l(x)$, with example values $\beta_{0,0} = \pi$ **(1S)**, $\beta_{0,1} = 4.5$ **(1P)**, $\beta_{0,2} = 5.8$ **(1D)**, $\beta_{1,0} = 2\pi$ **(2S)**, and $\beta_{1,1} = 7.7$ **(2P)**. In parentheses and boldface, we list the atomic notations for the orbitals involved. This outcome explains why these spherical nanocrystals are often called *artificial atoms*.

The spacing between the lowest two energy levels $\Delta E = E_{0,1} - E_{0,0}$, also called the *1S transition*, for a CdSe nanocrystal with a conduction band effective mass m_e^* of 0.13 m_e and a radius R of 2 nm is calculated, using Equation 3.2, as follows:

$$E_{n,1} = \frac{2.9 \text{ eV}}{R^2}\left(\frac{\beta_{n,1}}{\beta_{0,0}}\right)^2 \qquad (3.4)$$

or $\Delta E = E_{0,1} - E_{0,0} = 0.76$ eV, which is much larger than the thermal energy of the carriers (kT). This result is schematically represented in Figure 3.23, where in panel c we compare the absorption α of light by bulk CdSe with that by a nanocrystal of CdSe, clearly indicating that the bandgap is wider for the nanoparticle, which results in a blue shift. The probability of optical transitions between two quantum states is controlled by Fermi's Golden Rule (see Volume I, Chapter 3), which predicts that optical absorption must be proportional to the density of states, because a photon cannot be absorbed if there is no final state available for the electronic transition to go to. Bulk semiconductor and nanocrystal have the same overall absorption per unit volume when integrated over all frequencies. But in the case of a nanocrystal, the absorption spectrum consists of a series of discrete transitions with very high absorption transitions (sharp peaks), whereas a bulk semiconductor has a continuous absorption

spectrum. This spectral concentration of a quantum dot enhances all resonant effects and increases the energy selectivity. The sharpening of the density of states (DOS) at specific energies induced by quantum confinement is at the origin of many improvements in the properties of nanostructured materials compared with bulk materials.

The 1S (e) electron level shown in Figure 3.23b increases in energy with a decreasing R, while the 1S (h) energy level decreases for a decreasing R. As a consequence, the bandgap of CdSe grows over a wide range as one varies R, resulting in a dramatic shift in the optical spectra with changing particle size. This is called the *blue shift*, as the energies shift to higher values or shorter wavelengths for a decreasing particle size. The particle size sensitivity of optical spectra enables one to tune continuously those spectra from blue to infrared, for both absorption and emission. This can lead to applications ranging from new more intense light-emitting diodes to more stable fluorescent labels. As an important example, the very strong optical transitions at particular frequencies have motivated researchers to develop quantum dot lasers (the laser principle, which is explained in Volume I, Chapter 5). Arakawa and Sakaki, in 1982, predicted significant increases in the gain and decreases in the threshold current of semiconductor lasers utilizing quantum dots.[56] These predictions set off an intense effort worldwide to fabricate such structures.

In Volume I, Chapter 5, we compared the energy separation calculated above for a semiconductor particle to that of a metal nanoparticle of the same size (2 nm). We found that quantum effects are much more readily observed for semiconductor particles than for metal particles (see "Optical Properties of Metals").

Plasmonic Effects in Metal Nanoparticles

Surface plasmon (SP) waves are coherent electron oscillations at the interface between any two materials where the real part of the dielectric function changes sign across the interface (say a metal-dielectric interface). In the case of metal nanoparticles, we saw in Volume I, Chapter 5, that these surface plasmon waves are localized and are readily observed. These "localized plasmons" may be excited directly by light absorption. For metallic nanoparticles significantly smaller than the wavelength of light, light absorption then concentrates to a narrow wavelength range and results in brilliant colors (see Volume I, Figure 5.94). With changing particle size and shape, those colors change because of changes in the surface plasmon resonance frequency ω_p, the frequency at which conduction electrons oscillate in response to the alternating electric field of the incident electromagnetic radiation (see Equation 5.216). For a gold spherical particle, the frequency is about 0.58 of the bulk plasma frequency. Although the bulk plasma frequency for bulk gold is in the UV range, the SP frequency is in the visible range (close to 520 nm). The Au plasmon effects are illustrated by the historic stained glass which adorn medieval cathedrals (see Figure 3.1). Only materials with free electrons (essentially Au, Ag, and Cu and the alkali metals) possess plasmon resonances in the visible spectrum. Au and Ag often are used in nanophotonics because of their excellent chemical stability and very good conductivity. For nonspherical particles, such as rods, the resonance wavelength depends also on the orientation of the electric field. Confinement and quantization phenomena are visible in semiconductors already at dimensions of 200 nm, whereas in metals they typically are seen only at 1–2 nm. Metallic nanostructures exhibit major changes in their optical spectra due to light absorption by plasmonics waves at the surface of these structures. The plasmonic effects of metal nanoparticles often mask the metal nanoparticle quantum effects.

Transparent Ceramic Nanoparticles

Previously, we dealt with the fascinating quantum properties of semiconductor quantum dots and the plasmonic behavior of metal nanoparticles. Here we clarify the properties of insulating ceramic nanoparticles. Ceramics can be defined as inorganic, nonmetallic materials. They are typically crystalline in nature and are compounds formed between metallic and nonmetallic elements, such as aluminum and oxygen (alumina: Al_2O_3), calcium and oxygen (calcia: CaO), and silicon and nitrogen (silicon nitride: Si_3N_4). The two main categories of ceramic nanoparticles are metal oxide ceramics, such as titanium, zinc, aluminum, calcium, and iron oxides, to name just a few, and silicate nanoparticles (silicates, or silicon oxides), generally in the form of nanoscale

flakes of clays. It is the width of the forbidden bandgap (E_g) that determines whether these ceramics act as insulators ($E_g > 2$ eV) or as semiconductors ($E_g < 2$ eV). The fact that the dimensions of these ceramic nanoparticles are below the critical wavelength of light renders them transparent, which makes for a convenient property for their use in packaging, cosmetics, and coatings. Ceramic nanoparticles, like metal and semiconductor nanoparticles, also can be formed into coatings and bulk materials at lower temperatures than their non-nanocounterparts, reducing manufacturing costs.

Superhydrophobicity and the Lotus Effect

In Volume I, Chapter 6, we saw how nanofeatures on a surface improves the wettability for hydrophilic surfaces ($\theta < 90°$). A drop on such surface will seem to sink into the hydrophilic surface (see Volume I, Figure 6.45a). This phenomenon is called *superhydrophilicity*. For hydrophobic surfaces ($\theta > 90°$) the wettability decreases for a nanotextured surface compared with a polished surface of the same material. It is energetically unfavorable to wet a rough hydrophobic surface and this result in an increased water repellency or superhydrophobicity (Volume I, Figure 6.45b). Superhydrophic surfaces, like a lotus leaf, are also self-cleaning. As we saw in Volume I, Figure 6.46, a droplet on an inclined superhydrophobic surface does not slide off; it rolls off. When droplets roll over contaminant particles, the particles are removed from the surface if the force of absorption of the particles by the droplet is higher than the stiction force between the particles and the surface. Usually the force needed to remove particles is very low because of the minimized contact area between the contaminants and the surface. As a result, the droplet cleans the leaf by rolling off the surface. The combination of the chemistry, the nanostructures, and the adherence properties of dirt and water to the surface, is what Barthlott and Neinhuis named the lotus effect.[57] In the lotus plant, the effect arises because lotus leaves have a very fine surface structure and are coated with hydrophobic wax crystals of around 1 nm in diameter. In the lotus plant, the actual contact area is only 2–3% of the droplet-covered surface.

Several surface treatments have been developed for roof tiles, fabrics, wood, paper, etc., that can stay dry and clean themselves in the same way as the lotus leaf (see Figure 3.24). Water does not adhere to the wood surface shown in Figure 3.24 but rolls off the paint, picking up and washing away debris in the process. By remaining dry, the coating also resists better mold, mildew, and algae. The lotus effect can be achieved, for example, by using special fluorochemical or silicone treatments on structured surfaces or with compositions containing microscale and nanoscale particulates. In one method, by Guo et al.,[58] an aluminum surface is made superhydrophobic by immersing it in sodium hydroxide for several hours (which roughens the surface) followed by spin coating a layer of perfluorononane to a thickness of 2 nm. This procedure increases the water contact angle from 67° to 168°. Electron microscope images show that the aluminum surface resembles that of a lotus surface, with a porous microstructure containing trapped air. Any surface treatment that creates micro- or nanoscale roughness with features having a height to width ratio greater than 1 that is subsequently coated with a thin hydrophobic coating will exhibit the lotus effect. Clothing that repels water has already been developed and marketed by brands such as Gap and Dockers; it uses a fabric named Nano-Care http://www.usato day.com/tech/columnist/cckev060.htm.

The BASF's lotus-effect aerosol spray, demonstrated in Figure 3.24, combines nanoparticles with hydrophobic polymers such as polypropylene,

FIGURE 3.24 Water-repellant wood: BASF's lotus-effect aerosol spray combines nanoparticles with hydrophobic polymers such as polypropylene, polyethylene, and waxes. As it dries, the coating develops a nanostructure through self-assembly. BASF says that the spray particularly suits rough surfaces such as paper, leather, textiles, and masonry.

polyethylene, and waxes. It also includes a propellant gas. As it dries, the coating develops a nanostructure through self-assembly.

Magnetic Properties of Nanoparticles

Magnetic behavior of nanostructured materials was investigated from the early days of nanotechnology. One of the key findings so far is that the coercivity H_c, the intensity of the magnetic field needed to reduce the magnetization of a ferromagnetic material to zero after it has reached saturation, as a function of particle size goes through a maximum (see Figure 8.52).[59,60] The coercivity first increases linearly with decreasing grain size. The reason for the coercivity H_c increasing proportional to $1/d$ is that as particles decrease in size the magnetic domain size starts to coincide more and more with the grain size. As an example, for Fe the $1/d$ scaling is observed from 100 μm down to 40 nm. In the single-domain region a change of magnetization can only be produced by rotation of the magnetization of a single grain or crystallite as a whole, which requires strong magnetic fields, depending on the shape and anisotropy of the particle. As the size of the magnetic elements is further scaled down the coercivity goes over a peak and then decreases very fast (with a d^6 dependence). In this regime, a transformation from ferromagnetic to superparamagnetic behavior takes place. Typical particle sizes for the ferrro- to superparamagnetic phase transformation are between 10 and 20 nm for oxides and 1–3 nm for metals. In the superparamagnetic state of the material, the room temperature thermal energy, kT, overcomes the magnetostatic energy of the domain or particle. The resulting fluctuations in the direction of magnetization cause the magnetic field to average to zero. Thus, the material behaves in a manner similar to paramagnetism, except that instead of each individual atom being independently influenced by an external magnetic field, the magnetic moment of the entire crystallite tends to align with the magnetic field. Superparamagnetic particles are attracted to a magnetic field but retain no residual magnetism after the field is removed (no hysteresis). These superparamagnetic nanoparticles, feature a high saturation magnetization, high relative permeability, low coercivity, high resistivity, and low hysteresis loss, and all of this benefits a large number of applications,* including in ferrofluids (see Figure 8.53).

The quest for a high-throughput manufacturing method for nanoparticles—be they semiconductors, metals, or insulators—is driven by their many extraordinary properties, some of which are summarized in Table 3.4.

Nanoparticle Manufacture

Introduction

Generally, nanoparticle manufacturing techniques sort out in either "bottom-up" methods or "top down" methods. In bottom-up methods, as we saw above, the nanoparticles are synthesized directly to the desired shape and size. One such bottom-up method is chemical synthesis in a liquid. An example of a top-down technique is a mechanical method such as ball milling (also called *mechanical attrition*). A further way of categorizing nanoparticle production techniques is based on the phase in which the process takes place, i.e., gas phase, liquid phase, and solid phase methods. The solid phase method is limited in terms of producing nanometer scale particles. The liquid phase method and the gas phase method are more suitable. Yet another way of classifying nanoparticle manufacturing techniques, the method we adapt in this chapter, is as follows:

1. In situ fabrication: These techniques incorporate lithography, vacuum coating [e.g., molecular beam epitaxial growth (MBE) and metallorganic chemical vapor deposition (MOCVD) epitaxial growth] and spray coating on solid surfaces.
2. Mechanical manufacturing: This is a top-down method that reduces the size of particles by attrition, for example, ball milling or planetary grinding.
3. Liquid phase synthesis (wet chemistry): This is fundamentally a bottom-up technique; i.e., it starts with ions or molecules and builds up into larger structures. These nanoparticle

* On the other hand, it is also because of superparamagnetism that hard drives are expected to stop growing once they reach a density of 150 gigabits per square inch.

Table 3.4 Semiconductor, Metal, and Insulator Nanoparticle Properties

Semiconductor
Quantum confinement at dimensions ~10–100 nm. Blue shift when the nanoparticle is of the size of the exciton's Bohr radius.
Light absorption is determined by the band-edge transition, which is strongly size- and shape-dependent, making the absorption highly tunable.
Density of states: electronic energy levels around the Fermi level are strongly affected for sizes in the 10–100-nm range.
Luminescence is very intense and can be tuned.
The I-V behavior is characterized by the Coulomb blockade.
Quantum efficiency (~0.1–0.5), spectral range of emission (400–2000 nm), and cross-section ~10^{-19} m^2.
A very high S/V ratio leads to very high catalytic properties and lowers the melting temperature T_m.
Metal
Quantum confinement at smaller dimensions ~2–3 nm. Electron motion is limited by the size of nanoparticle.
Light absorption is determined by the surface plasmon resonance, which is size- and shape-dependent.
Density of states: electronic energy levels around the Fermi level are strongly affected for sizes in the 1–2-nm range.
Luminescence is very weak.
The I-V behavior is characterized by the Coulomb blockade.
Quantum efficiency (~10^{-4}), spectral range of emission (500–9000 nm), and cross-section ~10^{-13} m^2.
A very high S/V ratio promotes very high catalytic properties and lowers melting temperature T_m.
Superparamagnetic nanoparticles feature a high saturation magnetization, high relative permeability, low coercivity, high resistivity, and low hysteresis loss, and all of this benefits ferrofluids.
Insulator
Improved sintering and hardness. Nanophase powder compacts more easily in the sintering process. Successful sintering enhances hardness of materials. Nanophase powders compact more easily than their analogous submicrometer particles.
Strength and hardness. Better for materials with smaller crystal grains, as dictated by the Hall-Petch relation (Volume I, Equation 2.70).
The specific heat capacity (C_v) of nanocrystalline materials is higher than that of their bulk counterparts.
A very high S/V ratio promotes very high catalytic properties and lowers melting temperature T_m.

manufacturing techniques historically can be placed under the title of colloid chemistry and involve classic "sol-gel" processes or other aggregation processes.

4. Gas phase synthesis: These methods include plasma vaporization, chemical vapor synthesis, and laser ablation.

Terminology one encounters frequently when discussing manufacture of nanoparticles includes:

- Cluster: A collection of units (atoms or reactive molecules) of up to about fifty units
- Colloids: A stable liquid phase containing particles in the 1–1000 nm range
- Nanoparticle: A solid particle in the 1–100 nm range that could be noncrystalline, an aggregate of crystals, or a single crystal
- Nanocrystal: A solid particle that is a single crystal in the nanometer range

In Situ Fabrication

Lithography Techniques

In a lithographic top-down approach, electron beam or ion beam lithography may be used to etch in situ a pattern of columns in a quantum well heterostructure [e.g., a GaAs/AlGaAs quantum well (Figure 3.25)]. The quantum well heterostructure provides

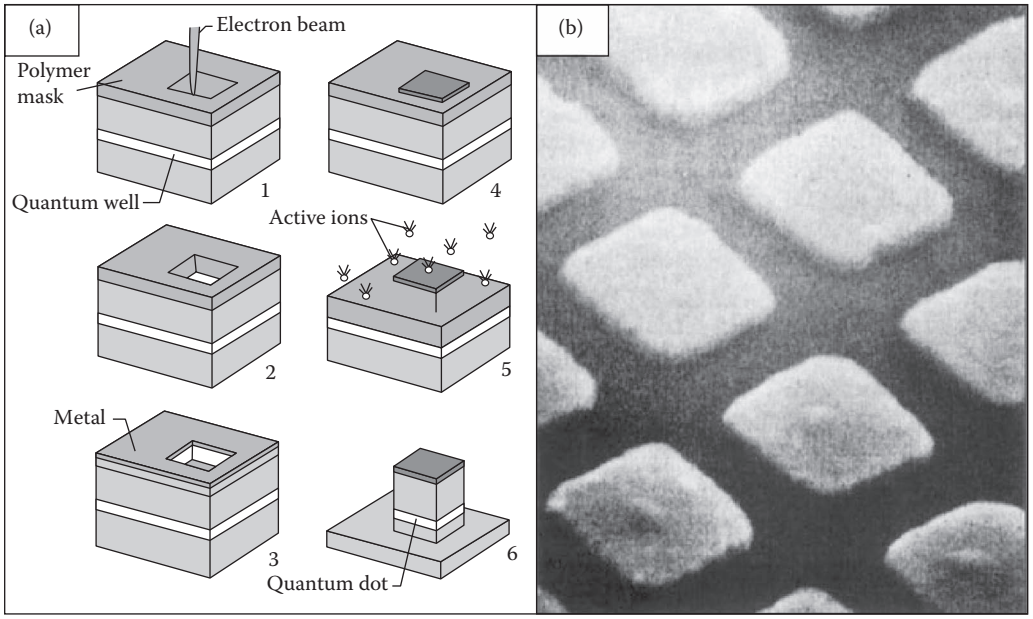


FIGURE 3.25 Fabrication of small, laterally patterned multiple quantum dot columns by e-beam lithography in a GaAs/AlGaAs. (a) Process of quantum dot etching. (b) Etched quantum dots in a Gas/AlGaAs well (scanning electron microscope picture).[61]

1D electron confinement in a plane because of the mismatch of the bandgaps of the two composing materials resulting in the formation of a potential energy well at their interfaces (see also Volume I, Figures 3.122 and 3.128). In the example in Figure 3.25, a positive resist is used as electron or ion beam mask and in the exposed areas the mask is removed. Next, the entire surface is covered with a thin metal layer. After a lift-off development, the polymer film and the protective metal layer are removed, and a clean surface of the sample is obtained, except for the previously exposed areas, where the metal layer remains (4). Next, by reactive ion etching, the areas not protected by the metal mask, slim pillars are created, containing the cut-out fragments of quantum wells (5–6). In this way, the motion of electrons, which initially is confined in the plane of the quantum well, is further restricted to a small pillar with a diameter of the order of 10–100 nm. Owing to the simplicity of the production of thin, homogeneous quantum wells, GaAs is the most commonly used material for creating dots by means of etching. Disadvantages of this technique are that it is slow, low density, and prone to defect formation.

As pointed out in Volume I, Chapter 3, confinement in lateral directions may also be implemented using electrostatic confinement with metal electrodes, lithographically patterned on the surface of a quantum well structure (see Volume I, Figure 3.138, for a quantum wire and Figure 3.144 for a quantum dot).

Self-Organized Epitaxial Growth

Self-organized epitaxial growth is another example of in situ nanoparticle manufacture. Oriented growth is referred to as epitaxy and, as we saw in Volume II, Chapter 7, it is subdivided into homoepitaxy and heteroepitaxy. Homoepitaxy is epitaxy performed with only one material: a crystalline film is grown on a substrate or film of the same material. Heteroepitaxy is epitaxy performed with materials that are different from each other: a crystalline film grows on a crystalline substrate or film of another material. This technology may be applied for growing crystalline films of materials of which single crystals cannot be obtained and for the integration of crystalline layers of different materials. Examples include gallium nitride (GaN) on sapphire or aluminum gallium indium phosphide (AlGaInP) on gallium arsenide (GaAs). Here, we will use the heteroepitaxy technique to form islands (quantum dots) of one material on the surface of another. To understand this method better, we first gather some

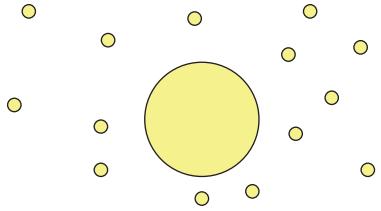


FIGURE 3.26 Nanoparticle growth in a vapor.

background on nucleation and growth of particles in air and on a solid surface.

Consider a droplet of radius r in contact with a vapor (Figure 3.26). The Gibbs free energy change ΔG for this system is given by the following expression:

$$\Delta G = \frac{4}{3}\pi r^3 \Delta G_V + 4\pi r^2 \gamma \tag{3.5}$$

where γ = surface tension, the first term on the right is the free energy change per unit volume, and the second term on the right is the free energy change per unit area of surface. The first term can be calculated as follows:

$$\Delta G_V = \frac{kT}{\Omega}\ln\left(\frac{P_V}{P_S}\right) = \frac{kT}{\Omega}\ln(1+S) \tag{3.6}$$

with S defined as supersatuation $S = \frac{P_v - P_s}{P_s}$, Ω = the atomic volume (volume per atom), P_v the actual pressure in the gas phase, and P_s the saturated vapor pressure at equilibrium. When $P > P_s$, one defines supersaturation, $P = P_s$ means equilibrium is attained, and at $P < P_s$ the gas phase is undersaturated. The presence of surface tension, as illustrated in Figure 3.27, always produces an activation barrier to the nucleation of condensed phases. This is true for solids and for liquids (see, for example, the case of etching a single crystal in a solution as illustrated in Volume II, Figure 4.60).

The critical nucleus size, marked r_c in Figure 3.27, can be calculated from the following:

$$\frac{\partial \Delta G}{\partial r} = 4\pi r^2 \Delta G_V + 8\pi r \gamma \tag{3.7}$$

The maximum of the curve is at $\frac{\partial \Delta G}{\partial r} = 0$, so that $r_{critical} = \frac{2-\gamma}{\Delta G_v}$ and, with $\Delta G_v = \frac{kT}{\Omega}\ln\left(\frac{P_v}{P_s}\right) = \frac{kT}{\Omega}(1+S)$ (Equation 3.6) we obtain:

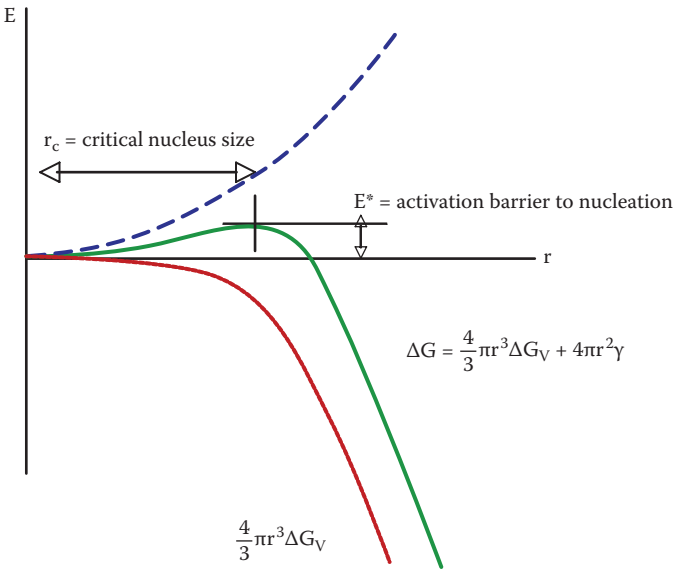


FIGURE 3.27 The Gibbs free energy as a function of particle radius.

$$r_{critical} = \frac{-2\gamma\Omega}{kT\ln(1+S)} \tag{3.8}$$

and also:

$$\Delta G^*_{critical} = \frac{16\pi\gamma^3}{3(\Delta G_v)^2} \tag{3.9}$$

From Equation 3.8, as supersaturation S increases, the critical nucleus size decreases, so that for very small particles one needs high super-saturation and rapid nucleation.

The deposition rate of atoms on the nanoparticle is then given as follows:

$$\frac{dN}{dt} = hn_s A^* \exp^{\frac{-\Delta G^*_{critical}}{kT}} \tag{3.10}$$

where h = the Arrhenius pre-exponential
n_s = the number density of atoms in the gas phase
A* = the area of the critical nucleus

The Schwöbel barrier ($\Delta G^*_{critical}$) determines the kinetics (dN/dt) of the system.

For condensation of atoms on a solid surface (Figure 3.28) instead of on a spherical particle, Equation 3.5 is modified to the following expression:

$$\Delta G = l_1 l_2 h \Delta G_V + l_1 l_2(\gamma) + l_1\beta_1 + l_2\beta_2 \tag{3.11}$$

Surface tension — Line tension

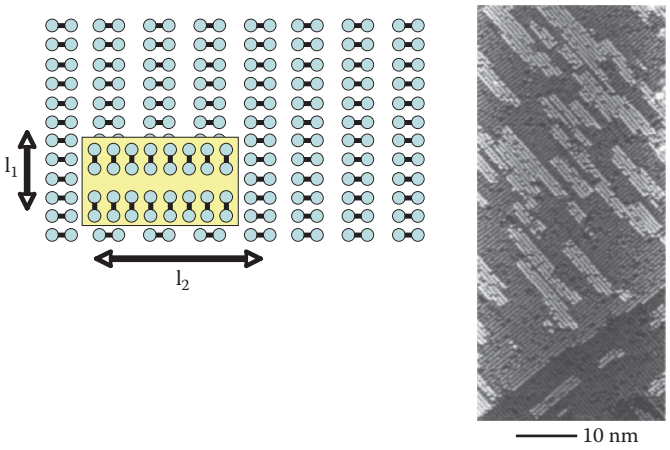


FIGURE 3.28 Island formation on flat solid surfaces.[62]

To identify l_1 and l_2 see Figure 3.28. The product $l_1 l_2$ corresponds to the area of the adsorbate coverage. Where γ is the surface tension and β_1 and β_2 are line tensions. A line tension represents an additional contribution to the surface tension due to the presence of steps. The step line tension may be calculated as the difference between the surface tension of stepped and flat surfaces per step and step length. When an adsorbate coverage as shown in Figure 3.28 exceeds one monolayer, one speaks of thin-film growth. Thin-film growth is controlled by the interplay of thermodynamics and kinetics, and the following elementary processes take part: adsorption, surface diffusion, re-evaporation, capturing by defects, and combination with other adatoms to form clusters (nucleation). Small clusters are metastable, but they become stable at a critical island size, determined by the energy gain for condensation and the energy cost to form new surfaces. Depending on the surface diffusion of atoms, different growth modes result, governed by the bond strength between the atoms in the layer and the atom-substrate bonds. The occurrence of these different growth modes can be understood qualitatively in terms of the surface tension, γ (surface energy) as depicted in Figure 3.29.

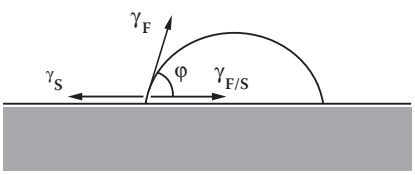


FIGURE 3.29 Surface tension of a film forming on a solid surface (see also Volume I, Figure 6.43).

The surface tension, represented by the symbol γ (see Volume I, Equation 6.88), is the force along a line of unit length, where the force is parallel to the surface but perpendicular to the line. Surface tension is therefore measured in force per unit length. Its SI unit is Newton per meter, but the cgs unit of dynes per centimeter is most commonly used. The wetting angle φ of a drop-like island, illustrated in Figure 3.29, is determined by the surface tensions of substrate (γ_S), film (γ_F), and film-substrate interface (γ_{FS}). Balance of these forces changes during growth. Flat surfaces exhibit three canonical growth modes, depending on the surface and interface energies. Beyond that, one finds more exotic growth modes at stepped surfaces.

In the case of $\gamma_S > \gamma_F + \gamma_{FS}$, typically, the first layer wets the surface but subsequent layers do not. The first atomic layer tries to coat the whole surface to provide optimal free energy reduction but with subsequent layers the situation changes: the first atomic layer reduces the surface energy of the substrate and new atoms do not experience the same interface energy as the first atomic layer. In addition, the added layers have to absorb the misfit strain energy, which grows with increasing film thickness (see below). Therefore, one can have either a continued layer-by-layer growth mode (Frank-van der Merwe mode), which is relatively rare, or the formation of islands on top of a flat first layer (Stranski-Krastanov growth mode). If the energy balance at the interface is tipped the opposite way, i.e., $\gamma_S < \gamma_F + \gamma_{FS}$, the overlayer has a tendency to nucleate three-dimensional islands right away and leave the low-energy substrate exposed (Volmer-Weber growth). This situation usually occurs when growing more reactive materials on top of an inert substrate, such as in the case of transition metals on a noble metal or an oxide.

The Stranski-Krastanov growth (SK growth; also spelled Stransky-Krastanov or Stranski-Krastanow) is one of the three primary modes by which thin films grow epitaxially at a crystal surface or interface. Also known as *layer-plus-island growth*, the SK growth mechanism was first noted by Ivan Stranski and L. Von Krastanov in 1939. It was not until 1958, however, in a seminal work by Ernst Bauer published in *Zeitschrift für Kristallographie*, that the

SK, Volmer-Weber, and Frank-van der Merwe mechanisms were systematically classified as the primary thin-film growth processes. Since then, SK growth has been the subject of intense investigation, not only to better understand the complex thermodynamics and kinetics at the core of thin-film formation, but also as a route to fabricating novel nanostructures for application in the microelectronics industry.

The range of possible growth modes becomes richer when going down in dimensionality. For example, when growing one-dimensional stripes on a two-dimensional, stepped surface. In addition to an extension of the three growth modes described above, one finds additional phenomena that are related to the interaction with the one-dimensional steps. As we saw in Equation 3.11 when stepped surfaces are involved line tensions due to the presence of steps contribute to the surface tension. Layer-by-layer growth turns into row-by-row growth, which is often called *step flow growth* because the step edge propagates downhill. There are other modes as well, such as the reverse step flow, where the material accumulates at the top of the step edge and propagates in the opposite direction. For manipulating these different growth modes, it is important to understand and control the surface and interface energies. Broadly speaking, one distinguishes three classes of solids where three different terms dominate the surface energy. Metals try to achieve the highest possible number of neighbors. Therefore, the surface with the lowest free energy is close-packed. Ionic solids, such as alkali halides, have the constraint that a stable surface needs to be charge neutral. Otherwise the Coulomb energy of a sheet of charge diverges. Covalent solids, such as semiconductors, have very directional bonds that rearrange themselves extensively in order to minimize the density of broken bonds at the surface. Bond counting is a first approximation, but the intricate rearrangement of atoms at semiconductor surfaces often defies our expectations. The absolute values of the surface energy are typically highest for transition metals and rare earths (2–3 Jm^{-2}); intermediate for covalent semiconductors, noble metals, and simple metals (1–2 Jm^{-2}); and lowest for ionic insulators (0.3–1 Jm^{-2}). The three growth modes are summarized in Figure 3.30.

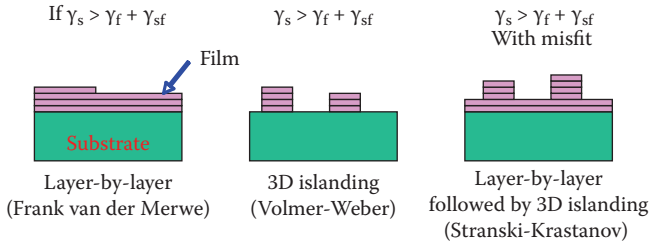


FIGURE 3.30 Summary of the three film growth mechanisms.

1. $\gamma_S > \gamma_F + \gamma_{FS}$. At the highest surface diffusion rates, layer-by-layer or Frank-van der Merwe growth mode is observed. The growing layer reduces the substrate surface very fast, i.e., it wets the surface.
2. $\gamma_S < \gamma_F + \gamma_{FS}$. High rate of incoming atoms and high Ehrlich-Schwöbel barrier, leads to island growth or Volmer-Weber growth. The growing layer wants to minimize interface energy and its own surface energy and the layer "balls up" on the surface.
3. $\gamma_S > \gamma_F + \gamma_{FS}$ with misfit. Layer plus island or SK growth. SK growth is possible in a heteroepitaxial system (misfit). This is the case of most interest for nanomanufacturing.

If the crystal structure (lattice constants) is different for substrate and film, strain appears at the interface because of this misfit. The strain depends on the size of this misfit: $\varepsilon = (b - a)/a$, with *a* and *b* the lattice constants for substrate and film (see also Vegard's law, Volume I, Equation 4.74). The elastic strain of a lattice mismatch can be accommodated either by a pseudomorphic growth or by misfit dislocations, as illustrated in Figure 3.31.

In Figure 3.32, this is illustrated for the case of Ge on Si, where we see how the island structure forms (see also Volume I, Figure 4.59). In this case, since the Ge lattice is 4% bigger than Si, bonding of Ge to Si compresses the Ge lattice by 4%. As the deposited film gets thicker, atoms continue to be forced to adapt to a different lattice constant, and the total strain energy increases linearly with the thickness of the growing film:

$$E_{strain} = \lambda \left(\frac{\Delta a}{a} \right)^2 At \qquad (3.12)$$

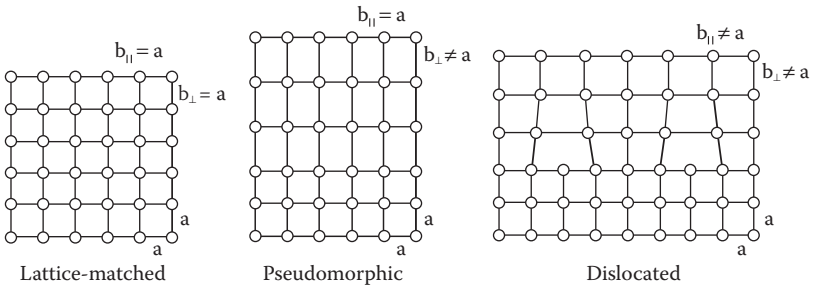


FIGURE 3.31 The deposited atoms try to adopt to the lattice constant. A lattice mismatch can be accommodated by pseudomorphic growth or by misfit dislocations.

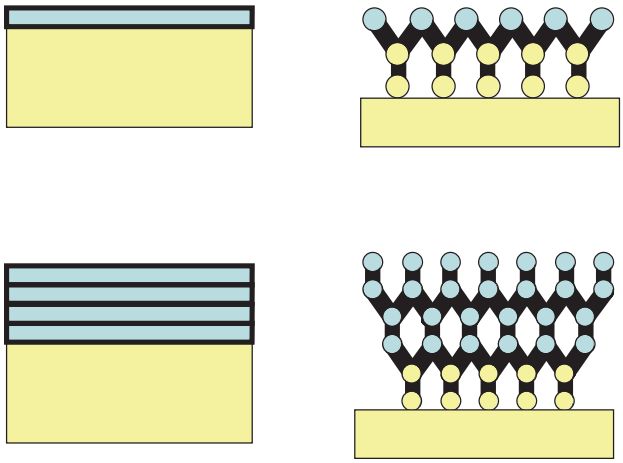


FIGURE 3.32 As the germanium lattice is 4% bigger than that of Si, bonding of Ge to Si compresses the Ge lattice by 4%. As the deposited film gets thicker, atoms continue to be forced to adapt to a different lattice constant, and the total strain energy increases linearly with the thickness of the growing film.

where

λ = elastic modulus
Δa = forced change in lattice constant
a = "free" lattice constant
A = area
t = thickness

In lattice-mismatched systems, self-organized quantum dots can be achieved by utilizing the Stranski-Krastanov growth mode (in an MBE or MOCVD reactor). In this mode, isolated islands form spontaneously above a certain critical thickness to relieve the mismatch strain energy. This is illustrated in Figure 3.33.

Depending on the growth conditions, one can get compact islands or ramified (fractal) islands. The compactness of the islands depends on the diffusion probability along its edges: this in turn depends on the surface temperature. In the extreme case of hit-and-stick, at low temperatures, where surface diffusion is limited, a diffusion limited aggregation model holds, which predicts fractal islands. This is demonstrated for the example of Pt growth on Pt(111) in Figure 3.34a (left-most image). The equilibrium shapes are hexagons (at 700 K) (Figure 3.34 panel c). The size distribution of the islands is affected by critical island size, coverage, and substrate temperature. According to scaling theory, one can derive a scaling relation for the size distribution, which depends only on the critical size.[63]

Certain surfactants may change the growth modes. Adding surfactants, a modification of the kinetics takes place rather than a change of the energetics. For example, the adatom diffusivity on the terraces might be influenced, the number of nucleation centers can be changed, the Schwöbel barrier may be changed, etc. The classic example is the growth of Ge on Si(111). On the bare surface there is a Stranski-Krastanov (SK) growth. With Sb as a surfactant, the film grows in a layer-by-layer fashion. Another example is the influence of H on Si(111) on the growth of Ag.

A successful method for quantum dot laser fabrication is based on the self-organized SK growth described here. SK growth occurs, we just saw, when a few monolayers of QD material (e.g., InGaAs) are deposited on a lattice mismatched material (e.g., GaAs). First, a quantum well-like wetting layer forms, from which quantum dots emerge. Strain, surface energy, and interaction with surrounding dots determine the dot shape, size, and distribution. Self-organized quantum dots are distributed on a semiconductor surface in a random way, and their size is subject to variations of a few percent (see

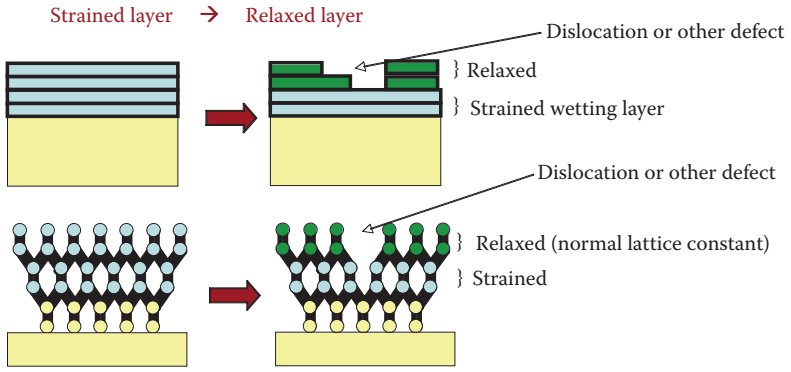


FIGURE 3.33 Stranski-Krastanov growth mode (in an MBE growth or MOCVD reactor): isolated islands form spontaneously above a certain critical thickness to relieve the mismatch strain energy.

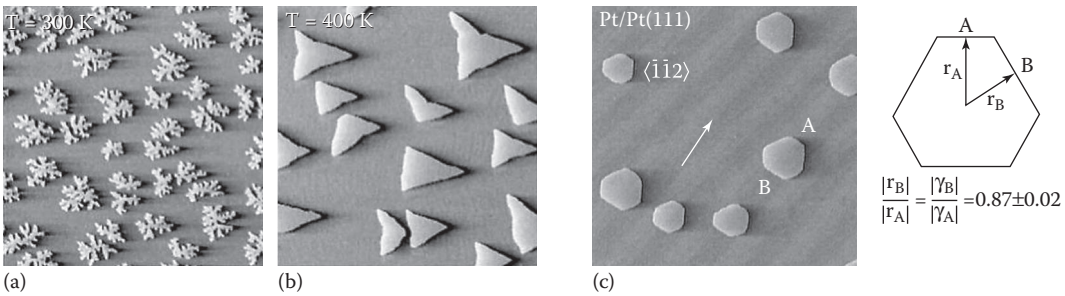


FIGURE 3.34 STM image of Pt growth on Pt(111). At 300 (a), 400 (b), and 700 K (c).

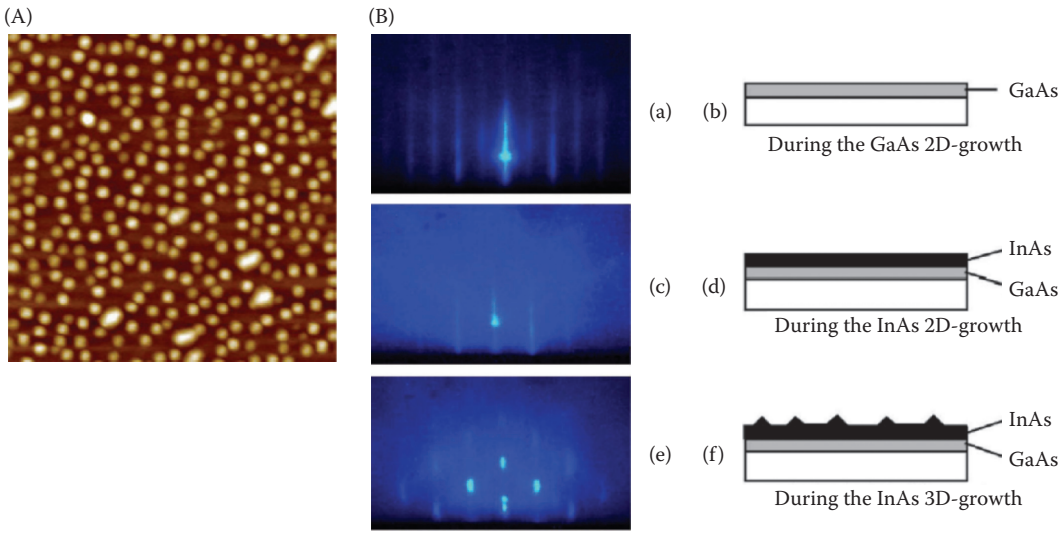


FIGURE 3.35 InAs quantum dots on a GaAs surface. (A) Atomic force microscope (AFM) picture of InAs/GaAs quantum dots. (B) Reflection high-energy electron diffraction (RHEED) patterns (a, c, e) during the two-dimensional (2D) deposition of GaAs (b) and the 2D deposition of InAs (d), and the 3D deposition of InAs (f) on a (100)GaAs buffer layer.[64]

Figure 3.35 with InAs quantum dots on a GaAs surface). While single quantum dots have sharp, atom-like emission lines, ensembles of quantum dots show an inhomogeneously broadened spectrum. Size, distribution, composition, and shape of the quantum dots can be chosen by different growth conditions such as temperature and precursor pressure. Larger quantum dots emit longer wavelengths, and smaller dots emit shorter wavelengths. The main advantages of self-organized growth are its integration into a single growth run, the low defect density, and the high dot density. The main limitations of this method are the cost of fabrication and the lack of control over positioning of individual dots. Some claim that Stranski-Krastanov methods for growing self-assembly quantum dots has rendered the

lithographic approach to semiconductor quantum dot fabrication virtually obsolete.

Reflection high-energy electron diffraction (RHEED) is very sensitive to surface changes, because of either structural changes or adsorption, and is a good method by which to monitor island growth. In RHEED, a high-energy electron beam (3–100 keV) is directed at the sample surface at a grazing angle. The distance between the streaks in a RHEED pattern (see Figure 3.35) gives us the surface lattice unit cell size. For an atomically flat surface, sharp RHEED patterns result and with rougher surfaces one gets diffused RHEED patterns. In layer-by-layer growth mode, RHEED oscillations are observed (Figure 3.35Ba). These are a direct measure of growth rates in MBE: the oscillation frequency corresponds to the monolayer growth rate. The incident angle dependence of the oscillations suggest that interference between electrons scattering from the underlying layer and the partially grown layer contribute to these oscillations. The magnitude of the RHEED oscillations damps out (see Figure 3.35Be) because as the growth progresses, islands nucleate before the previous layer is finished.

Epitaxial growth on patterned substrates can organize the quantum dots on a surface into a regular pattern, as illustrated in Figure 3.36 where the growth of AlGaAs in pyramid-shaped recesses is shown. The recesses are formed by selective ion etching. The main disadvantage of this technique is that the density of QDs is limited by the mask pattern.

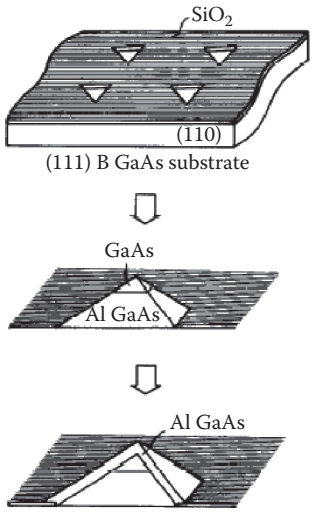


FIGURE 3.36 Growth of quantum dots in pyramid-shaped recesses on a GaAs substrate.[65]

Solid-State Processes

Solid-state processes, involving top-down grinding or milling to produce ceramic nanoparticles, are also referred to as mechanical attrition methods. In mechanical attrition, macro- or microscale particles are ground in a ball mill or in other size-reducing mechanisms. The resulting particles are air classified to recover the nanoparticles. Ball milling involves using fixed and moving elements to reduce larger particles into micron-sized and nanosized particles. The material to be ground is reduced in size because of the interaction of the fixed elements, the moving elements, and the material itself. The milling materials, the milling time, and the atmosphere all affect the nanoparticle properties. The kinetic energy transfer from balls to the powder in the milling bowl accomplishes the reduction in grain size. Since the kinetic energy of the balls is a function of their mass and velocity, dense materials such as steel or tungsten carbide are preferred. Other materials used as balls are agate, sintered corundum, zirconium dioxide, Teflon, chrome nickel, silicon nitride, etc. In order to prevent excessive abrasion, the hardness of the grinding bowl used and of the grinding balls must be higher than that of the materials that are being ground. Normally, grinding bowls and grinding balls should be made of the same material.

In the initial stage of milling, a fast decrease of grain size occurs that slows down after extended milling. Initially the kinetic energy transfer leads to the production of a myriad of dislocations in the powder grains resulting in atomic level strains. At a certain strain level, the dislocations annihilate and recombine to form small-angle grain boundaries, separating the individual grains. This way, subgrains with reduced grain size are formed. During further milling, this process extends throughout the entire sample. To maintain this reduction in size, the material must experience very high stresses. However, extended traditional ball milling cannot maintain the required high stresses, and hence reduction of grain size is limited. The other two parameters also causing a limit to grain size reduction in traditional ball milling are the local temperature developed due to ball collisions and the overall temperature in the bowl. A rise in temperature also results from

Table 3.5 Traditional Ball Milling Compared with High-Energy Ball Milling

	Ball Milling	High-Energy Ball Milling
Milling time	<1 h	20–200 h
Impact energy (W/g per ball)	0.001	0.2
Particle size	μm	μm clusters
Structural changes	No	Yes
Chemical changes	No	Yes
Atmosphere	No control	Vacuum, inert gas
Temperature	No control	Liquid nitrogen, –700°C

collisions between balls and balls, balls and powder, and balls and wall. Traditional ball milling thus produces only fine powders. In addition, traditional ball milling requires an increasing amount of energy and time to produce incrementally smaller particles. To produce nanomaterials, high-energy ball milling must be used (see Table 3.5).

High-energy ball milling uses steel balls to transfer mechanical energy into materials and creates structure changes and chemical reactions at ambient temperature (mechanochemistry). The impact energy is considerably higher, longer milling times are employed, and the process is carried out at liquid nitrogen temperatures in vacuum or in an inert atmosphere. High-energy ball milling can produce nanosized oxides and nitrides (<100 nm) via mechanical chemistry and has been recognized as one important synthesis method for nanomaterials, including nanoparticle, nanotube, nanocrystalline, nanoporous, and nanocomposite materials.

High-energy ball milling (see Figure 3.37 for a typical high-energy ball mill) is used for the commercial manufacture of magnetic, catalytic, and structural nanoparticles. Unfortunately, ball milling usually produces nanoparticles that are nonuniform in size and shape. The attrition approach is used to produce nanoparticles from materials that do not readily lend themselves to gas or liquid phase preparation.

FIGURE 3.37 High-energy ball milling in a rotating steel ball mill.

Liquid Phase Synthesis

Several liquid phase nanoparticle synthesis methods were reviewed in Volumes I and II. Here, we briefly review the most salient points about those different liquid phase synthesis methods.

An important example of liquid phase nanoparticle synthesis is colloidal processing as detailed in Volume I, Chapter 7, where the Derjaguin, Landau, Verwey, and Overbeek theory (the DLVO theory) was used to describe the stability of a colloid (see "Intermolecular Forces"). A colloidal system, we remember, consists of two separate phases: a dispersed phase (or internal phase) and a continuous phase (or dispersion medium) that can be a solid, a liquid, or a gas. The colloids of interest here are sols, in which solid particles are dispersed in a liquid. Chemical reactions can be fine-tuned into forming dispersed nanoparticles, and capping the dispersed solid particles, formed by these chemical reactions with appropriate ligands, stabilizes them. This method is widely used for preparation of quantum dots (semiconductor nanoparticles) and was pioneered by Columbia University's Louis Brus (Volume I, Figure 3.68) in the late 1970s when he was at Bell Labs and it was carried on by two of his post docs, Paul Alivisatos and Moungi Bawendi. As we learned in Volume I, Chapter 3, before 1993, quantum dots (QDs) were often prepared from colloidal aqueous solutions with added stabilizing agents to avoid colloid precipitation. In 1993, Bawendi and coworkers synthesized the best luminescent CdSe quantum dots made up to that time by using a high-temperature organometallic procedure instead.[66,67]

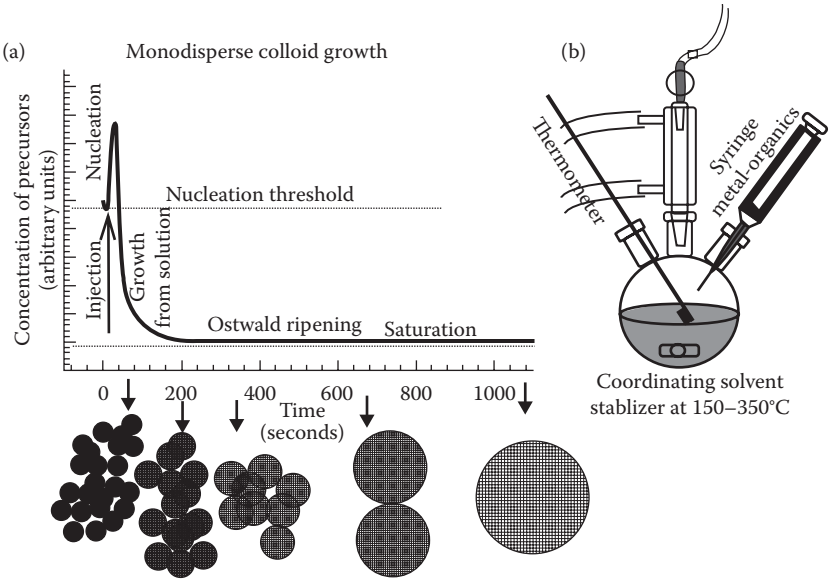


FIGURE 3.38 (a) Schematic depicting the stages of nucleation and growth for the preparation of monodisperse quantum dots (QDs) in the framework of the La Mer and Dinegar model. As nanocrystals grow over time, a size series of QDs may be isolated by periodically removing aliquots from the reaction vessel. (b) Representation of a typical apparatus used in monodisperse QD synthesis.[68]

Crystallites from –12 Å to –115 Å in diameter with consistent crystal structure, surface derivatization, and a high degree of monodispersity were prepared in a single reaction based on the pyrolysis of organometallic reagents by their injection into a hot coordinating nonaqueous solvent (see Figure 3.38b and Volume I, Figures 7.102 and 7.103).

A schematic to explain the formation of uniform nanoparticles in solution is shown in Figure 3.38a. In homogeneous precipitation, a short single burst of nucleation occurs when the concentration of constituent species reaches critical supersaturation. LaMer and Dinegar (1950) applied classic nucleation theory to qualitatively describe the kinetics of this type of burst nucleation. They proposed that from a strong initial supersaturation, a rapid nucleation or burst of particles occurs, followed by the slower absorption of diffusing atomic matter onto these nucleated particles. Their key observation was that the reduction in supersaturation after the nucleation burst strongly depletes the local availability (concentration) of constituent species, thus decreasing the rate of further nucleation, leading ultimately to a narrow size distribution of the nucleated, diffusively growing nanocrystals. Ostwald ripening may broaden the resulting distribution of nanocrystals. In Figure 3.38a this process starts at about 300s after injection. In Ostwald ripening, many small crystals initially form in a system, but they slowly disappear, except for a few that grow larger at the expense of the small crystals. The smaller crystals act as "nutrients" for the bigger crystals. This is a spontaneous process that occurs because larger crystals are more energetically favored than smaller crystals. While the formation of many small crystals is kinetically favored (i.e., they nucleate more easily), large crystals are thermodynamically favored. Thus, from a standpoint of kinetics, it is easier to nucleate many small crystals. However, small crystals have a larger ratio of surface area to volume than large crystals. Molecules on the surface are energetically less stable than the ones already well ordered and packed in the interior. Large crystals, with their greater volume to surface area ratio, represent a lower energy state.

Colloidal synthesis is by far the cheapest method and has the advantage of being carried out in benchtop conditions. For an example of a typical set-up see the apparatus in Figure 3.38b. This method is simple and can fabricate large quantities of quantum dots.

In Volume II, Chapter 8, we introduced sol-gel processing, which produces particles (typically metal oxides) by a series of hydrolysis and condensation reactions to form a sol of particles suspended

in a liquid phase (Volume II, Figure 8.2). The sol-gel technique constitutes a chemical forming process for making very small particles—say, 20–40 nm in diameter—a particle size that is virtually impossible to make by conventional mechanical grinding (Volume II, Figure 8.1). The method is used to synthesize glass, ceramic, and glass-ceramic nanoparticles. From Volume II, Figure 8.2, ultrafine and uniform ceramic powders may be formed either by grinding of a xerogel or by slowly drying the sol. The disadvantages of sol-gel processing include the toxic chemicals required for the reactions and the lengthy preparation time.

Additional liquid phase nanoparticle synthesis techniques include sonochemical processing, cavitation processing, and microemulsion processing. In sonochemistry, an acoustic cavitation process generates a transient localized hot zone with extremely high temperature and pressure gradients. Such sudden changes in temperature and pressure assist the destruction of the sonochemical precursor (e.g., an organometallic compound) and the formation of nanoparticles. The technique can be used to produce large volumes of materials for industrial applications.

In hydrodynamic cavitation, nanoparticles are generated through creation and release of gas bubbles inside a sol-gel solution.[69,70] The sol-gel solution is mixed by rapidly pressurizing it in a supercritical drying chamber and exposing it to cavitational disturbance and high-temperature heating. The erupted hydrodynamic bubbles are responsible for nucleation, growth, and quenching of the nanoparticles. Acoustic cavitation in liquids actually leads to two different types of major effects, i.e., physical phenomena such as streaming, turbulence, microjet, shear, etc. and chemical phenomena such as radical production. The collapse of cavities can indeed generate highly reactive radicals, which may induce specific chemical reactions. One easily recognizes how in the case of gas-solid reactions, the dissolution of solids is enhanced due to the turbulent mixing generated by hydrodynamic cavitation. The vigorous mixing enhances the transport of gas molecules to the solid surface resulting in an increase in the mass transfer and hence the overall reaction rate. The particle size can be controlled by adjusting the pressure and the solution retention time in the cavitation chamber. Gedanken[69] and Gogate[70] reviewed the use of sonochemistry for the fabrication of inorganic nanomaterials of various shapes, size, structure, and phases.

Microemulsions (liquid in a liquid) and micelles have been used for synthesis of metallic, semiconductor, silica, barium sulfate, magnetic, and superconductor nanoparticles.[71] In contrast to ordinary emulsions, microemulsions form upon simple mixing of the components and do not require the high shear conditions generally used in the formation of ordinary emulsions. The two basic types of microemulsions are direct [oil dispersed in water (o/w)] and reversed [water dispersed in oil (w/o)]. The technique is useful for large-scale production of nanoparticles using relatively simple and inexpensive hardware.

The diagram in Figure 3.39 explains the basic scheme for microemulsion nanoparticle synthesis. Basically, each microemulsion type (A and B) contains a reactant of the synthesis, and when they are forced together (in a reactor) exchange occurs causing the reactants to combine. Because there is limited reactant in each droplet, the size of the nanoparticle is controllable. The ability to synthesize nanoparticles in a water/supercritical fluid microemulsion was also realized.[72–75] This method promises to be a highly useful route to controlled nanoparticle synthesis because of the added control variables afforded by the tunability of the solvent quality (density) through pressure and temperature.

Liquid phase synthesis techniques in general are better than gas phase techniques (see next) for controlling the final shape and size of the nanoparticles. The ultimate size is controlled, as in the case of gas

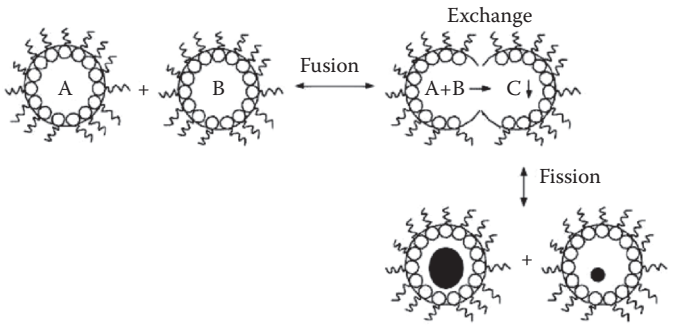


FIGURE 3.39 Microemulsion nanoparticle synthesis.

phase synthesis, by either stopping the process when the desired size is reached or by choosing chemicals that form particles that are stable, and stop growing at a certain size. These approaches are relatively low cost and high throughput, but contamination from precursor chemicals may be a problem.

Gas Phase Synthesis

Introduction

In gas phase nanoparticle synthesis a gas phase mixture is rendered thermodynamically unstable relative to the formation of the desired nanoparticles. Supersaturation is achieved by vaporizing material into a background gas and condensation to form particles is induced by cooling, e.g., by mixing with a cool gas or expansion through a nozzle. Once nucleation occurs, the remaining supersaturation is relieved by additional condensation on virgin substrate areas, or reaction of vapor phase molecules continues on deposited particles. This initiates the particle growth phase and rapid quenching after nucleation prevents particles that are too large.

In Table 3.6 we list several types of gas phase nanoparticle synthesis techniques grouped into those that use solid precursors and those that employ liquid or vapor precursors. The latter category is called *chemical vapor synthesis* (CVS), a modified chemical vapor deposition (CVD) method where the process parameters are adjusted to form nanoparticles instead of a thin film.

Gas Phase Synthesis with Solid-State Precursors

Inert-gas aggregation frequently is used to make nanoparticles from solid metals with low melting points (e.g., Bi). The metal is vaporized in a vacuum chamber and then supercooled with an inert gas stream. The supercooled metal vapor condenses into nanometer-sized particles that can be entrained in the inert gas stream and deposited on a substrate or studied in situ. Inert gases are used to avoid oxidation when creating metal nanoparticles, whereas a reactive oxygen atmosphere is used to produce metal oxide ceramic particles. The main advantage of this approach is the low contamination level. Final particle size is controlled by varying temperature, gas environment, and evaporation rate.

Pulsed laser ablation, like ion cluster beam deposition, is a nanoparticle gas phase synthesis method that uses solid precursors. Both techniques were reviewed in Volume II, Chapter 7. Pulsed laser ablation is used to vaporize a plume of material. Generally, the method can only produce small amounts of nanoparticles, but it can vaporize materials that are not easily evaporated, e.g., in the synthesis of Si, MgO, titania, and hydrogenated-silicon nanoparticles. There is a strong dependence of particle formation dynamics on the background gas. The use of pulsed laser ablation shortens the reaction time and allows for the preparation of very small particles. In 1995, at Rice University, it was reported that synthesis of carbon nanotubes could be accomplished by laser ablation.[76] A pulsed or continuous laser was used to vaporize a graphite target placed in an oven at 1200°C (Volume II, Figure 7.18). Contact with a cooled copper substrate caused the carbon atoms to be deposited in the form of nanotubes. This method has a carbon nanotube yield of around 70% and produces primarily single-walled carbon nanotubes with a controllable diameter determined by the reaction temperature. However, it is more

Table 3.6 Gas Phase Synthesis Methods for Nanoparticles

Methods Using Solid Precursors	Methods Using Liquid or Vapor Precursors
Inert and reactive gas condensation	Spray pyrolysis using different types of atomizers (e.g., pressure electrostatic, ultrasonic) and salt-assisted spray pyrolysis (SASP)
Pulsed laser ablation	Laser pyrolysis/photochemical synthesis
Ion cluster beam deposition	Low-temperature reactive synthesis
Spark discharge deposition	Thermal plasma synthesis
	Flame decomposition
	Flame spray pyrolysis

expensive than either arc discharge or chemical vapor synthesis.

Ion cluster beam technology also qualifies as a gas phase nanoparticle synthesis method. In ion cluster beam technology the cooling occurs when the vapor exits a special evaporation cell. As shown in Volume II, Figure 7.14, the heating of the evaporant in an evaporation cell with a small opening causes an adiabatic expansion from more than 100 to 10^{-5} or 10^{-7} mbar, upon exiting that cell. The expansion causes a sudden cooling, inducing the formation of atom clusters.

In spark discharge deposition, electrodes made of the metal to be vaporized in the presence of an inert background gas are polarized until the breakdown voltage is reached and an arc is formed across the electrodes that vaporizes a small amount of metal (e.g., Ni). This is the very technique that led to the discovery of carbon nanotubes in 1991 by NEC's Sumio Iijima (Volume I, Figure 3.64). Carbon nanotubes were found in the carbon soot of graphite electrodes during an arc discharge, by using a current of 100 amps. During this process, the carbon contained in the negative electrode sublimates because of the high temperatures caused by the discharge. Because nanotubes initially were discovered using this technique, it has been the most widely used method of nanotube synthesis. The yield for this method is up to 30% by weight, and it produces both single-walled and multiwalled nanotubes with lengths of up to 50 μm.

Gas Phase Synthesis with Liquid and Vapor Precursors

The entire range of reaction regimes and corresponding microstructures (epitaxial, polycrystalline, columnar, granular films, and aerogel coatings, as well as nanopowders) in going from chemical vapor deposition (CVD) to chemical vapor synthesis (CVS) of nanoparticles are illustrated in Figure 3.40.

In chemical vapor synthesis (CVS as opposed to CVD), vapor phase precursors are brought into a hot-wall reactor under nucleating condition. In CVS, vapor phase nucleation of particles is favored over film deposition characteristic for CVD. CVS constitutes a very flexible technique that can produce a wide range of materials and takes advantage of the huge database of precursor chemistries developed for CVD processes (see Volume II, Chapter 7). Examples include oxide-coated Si nanoparticles for high-density nonvolatile memory devices and tungsten (W) nanoparticles by decomposition of tungsten hexacarbonyl. The method allows for the formation of doped or multicomponent nanoparticles by the use of multiple precursors. For example, nanocrystalline europium-doped yttria is synthesized from organometallic yttrium and europium precursors. Other multicomponent particles that have been made by CVS include erbium in Si nanoparticles and zirconia particles doped with alumina.

Both in CVD and CVS, precursors are metalorganics, carbonyls, hydrides, chlorides, and other volatile compounds. The energy for the conversion

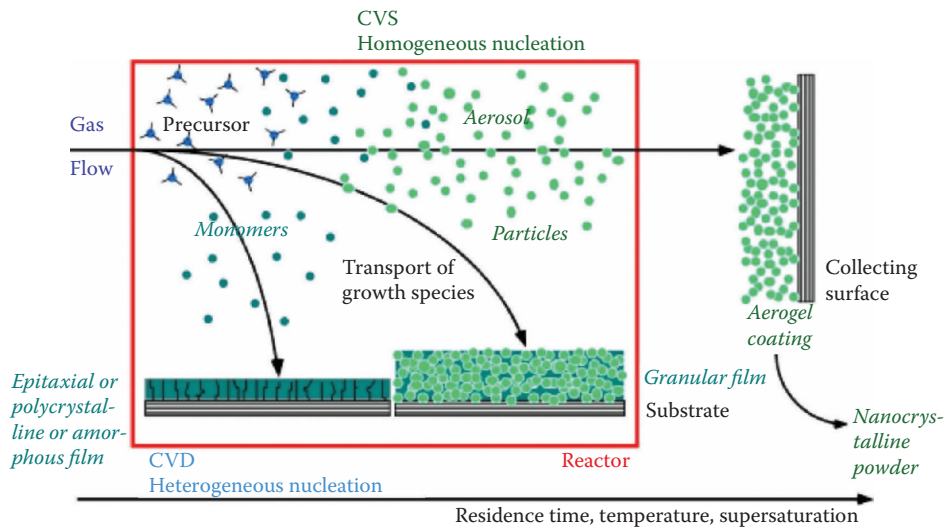


FIGURE 3.40 From chemical vapor deposition (CVD) to chemical vapor synthesis (CVS).

of the reactants into nanoparticles is supplied in hot wall (external furnace), flame (reaction enthalpy), plasma (microwave or radiofrequency), and laser (photolysis or pyrolysis) reactors.

The most important process parameters determining the quality and usability of the nanopowders are the total pressure (typical range from 100 to 100,000 Pa), the precursor material (decomposition kinetics and ligands determining the impurity level), the partial pressure of the precursor (determining the production rate and particle size), the temperature or power of the energy source, the carrier gas (mass flow determining the residence time), and the reactor geometry. The nanoparticles are extracted from the aerosol by means of filters, thermophoretic collectors, electrostatic precipitators, or scrubbing in a liquid. A typical laboratory reactor consists of a precursor delivery system, a reaction zone, a particle collector, and a pumping system. Modifications of the precursor delivery system and the reaction zone allow the synthesis of pure oxides, doped oxides, coated nanoparticles, functionalized nanoparticles, and granular films. Several CVS methods, including spray pyrolysis, salt-assisted spray pyrolysis (SASP), flame spray and electrospray pyrolysis for nanoparticle synthesis were detailed in Volume II, Chapter 9. For further reading on nanoparticle fabrication, a good resource is Jacak et al.[77] You may also want to consult Green and O'Brien.[78]

Carbon Nanostructures: Fullerenes

Introduction

Carbon, as we learn in Chapter 5, this volume, comes in an amazing number of varieties that may be grouped as crystalline, amorphous, nanoparticle, and engineered. The chemical genius of carbon is that it can combine with itself in different ways to create structures with entirely different properties. Two prominent examples of carbon structures are diamond and graphite. In pure sp^3 hybridization, four valence electrons are equally shared between neighboring carbon atoms to yield hard diamond, whereas in pure sp^2 hybridization, three in-plane covalent bonds leave a fourth delocalized electron to produce slippery graphite (see Figure 3.41 and Chapter 5, this volume: the carbon phase diagram in Table 5.9, Inset 2). In Volume I, Chapter 3, using the semiconductor band model, we explained how this same material acts as an insulator (diamond), a metal (graphite), or a superconductor (buckminsterfullerenes).

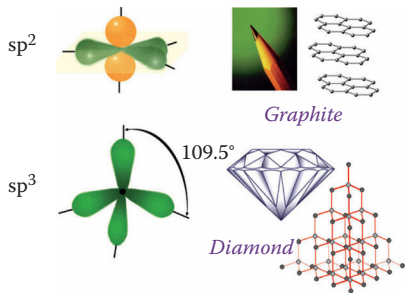


FIGURE 3.41 From graphite to diamond with carbon.[79]

In this section we only focus on fullerenes, carbon nanotubes, and graphene. Fullerenes are large, closed-cage, carbon atom clusters (Figure 3.42 and Volume I, Figure 1.29) and have several special properties that, until their discovery, were not found in any other material. Before the first synthesis and detection of the smaller fullerenes C60 and C70, it generally was assumed that these large spherical molecules were unstable. However, some Russian scientists already calculated that C60 in the gas phase should be stable and comes with a relatively large band gap. As is the case of many, important scientific breakthroughs, fullerenes were discovered accidentally. In 1985, Curl, Kroto, Smalley, and coworkers found strange results in mass spectra of evaporated carbon samples.[80] Herewith, the stability of C60 in the gas phase was proven, starting off the search for other fullerenes.

Since their discovery in 1991 by NEC's Sumio Iijima (Volume I, Figure 3.64), multiwalled carbon nanotubes (MWNTs) (Figure 3.43 and Volume I,

FIGURE 3.42 Sixty carbon atoms in the shape of a sphere—a buckyball. Every molecule combines twenty hexagons with twelve pentagons, resulting in a sphere shaped like the Buckminster Fuller geodesic sphere.

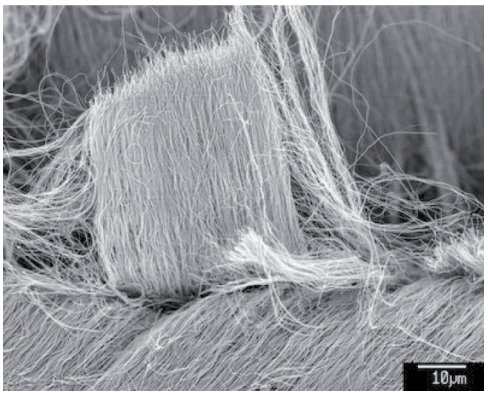


FIGURE 3.43 Multiwalled carbon nanotubes produced by the chemical vapor deposition (CVD) method. (Courtesy of the Department of Materials Science and Metallurgy, University of Cambridge.)

Figure 1.30) have been investigated by researchers all over the world.[81] With their large length (up to several millimeters) and small diameter (a few nanometers), they may be regarded as nearly one-dimensional fullerenes. Therefore, these materials possess additional interesting electronic, mechanical, and molecular properties. Carbon nanotubes, like fullerenes and graphene, profit from the intrusion of quantum phenomena and can be formed into a cylinder one atom thick in so-called "single-walled carbon nanotubes" (SWNTs) to give a wire that is perfect and without defects, much like the single-crystal whiskers we discussed in Volume I, Chapter 2.

Graphene is a flat monolayer of carbon atoms in a 2D honeycomb lattice. It is the basic building block for graphite materials of all other dimensionalities: 0D (buckyballs, C60), 1D (nanotubes), 2D (graphene), and 3D (graphite) (see Figure 3.44). Graphene is the two-dimensional crystalline form of carbon: a single layer of carbon atoms arranged in hexagons, like a sheet of chicken wire with an atom at each nexus. The familiar pencil-lead form of carbon, graphite, consists of layers of carbon atoms tightly bonded in the plane but only loosely bonded between planes; because the layers move easily over one another, graphite is a good lubricant. In fact, these graphite layers are graphene, although they had never been observed in isolation before 2004. As free-standing objects, such two-dimensional crystals were actually believed to be impossible to create—even to exist—until Russian-born physicists Andre Geim and Konstantin Novoselev at the

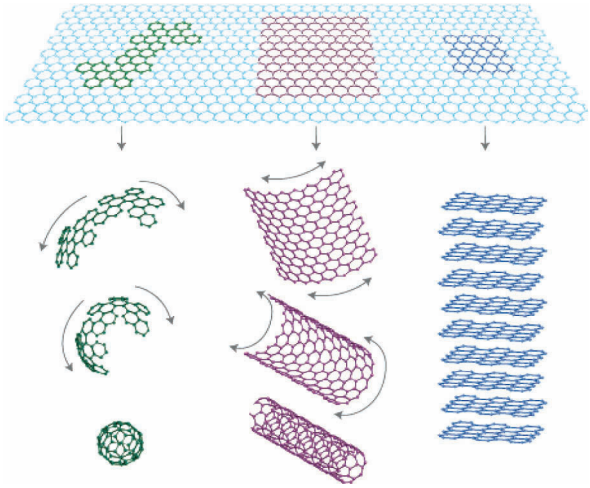


FIGURE 3.44 Graphene as the building block for graphite materials of all other dimensionalities.[94]

University of Manchester actually made free-standing graphene sheets in 2004.[82] Graphene sheets actually are easy to make but hard to see (Figure 3.45). It is now presumed that tiny fragments of graphene sheets are produced (along with quantities of other debris) whenever graphite is abraded, such as when a line is drawn with a pencil (Figure 3.45).

A summary of the many fullerene discoveries made since the discovery of C60 is presented in Table 3.7.

Buckyballs: 0D

A buckyball (short for *buckminsterfullerene*) is a 60-carbon atom molecule named for its resemblance to the geodesic spheres of architect Buckminster Fuller (Figure 3.42). Electrons in such a structure are confined in 3 dimensions and such a zero dimensional (0D) molecule or quantum dot comes with a de Broglie wavelength $\lambda_{DB} > d$ (the diameter of the particle) (see also Volume I, Table 3.11).

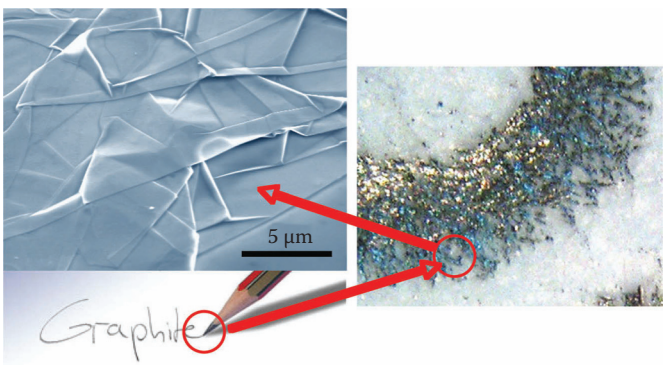


FIGURE 3.45 Easy-to-make, hard-to-see graphene sheets.[79]

Table 3.7 Short History of Fullerenes

1985: Finding of evidence for C60 in a mass spectrometry analysis of evaporated carbon samples by Kroto et al.[80]
1991: Discovery of multiwalled carbon nanotubes by Iijima[81]
1992: Conductivity of carbon nanotubes established by Mintmire et al.[83]
1993: Structural rigidity of carbon nanotubes measured by Overney et al.[84]
1993: Synthesis of single-walled nanotubes by Iijima and Ichihashi[85]
1995: Nanotubes used as field emitters by Rinzler et al.[86]
1997: Hydrogen storage in nanotubes demonstrated by Dillon et al.[87]
1998: Demonstration of the first carbon nanotube transistor by Dekker et al.[88]
1998: Synthesis of nanotube peapods by Smith et al.[89,90]
2000: Thermal conductivity of nanotubes measured by Tománek et al.[91]
2001: Integration of carbon nanotubes for logic circuits by Collins et al.[92]
2001: Intrinsic superconductivity of carbon nanotubes demonstrated by Kociak et al.[93]
2004: Discovery of stable graphene sheets by Novoselov et al.[82]

This soccer ball-shaped molecule measures just under a nanometer across and was discovered by Robert F. Curl, Jr. (Rice University), Harold W. Kroto (University of Sussex), the late Richard E. Smalley (Rice University), and some graduate students in August of 1985.[80] The professors (but not the graduate students—surprise!) received the 1996 Nobel Prize in Chemistry for this discovery. Spherical molecules like the one depicted in Figure 3.42, called *fullerenes*, come in various sizes and are identified by the letter C followed by the number of carbon atoms, for example, C60, C70, and C80.

Smalley made his early C60 buckyballs by vaporizing carbon with a laser and allowing the carbon to condense in a cold zone (see laser ablation in Figure 3.56, and Volume II, Figure 7.18).

Until C60 could be created in somewhat larger quantities, scientists were stumped in their attempts to study the new material. In 1990, Dr. Donald R. Huffman, a physicist at the University of Arizona, and Dr. Wolfgang Kratschmer of the Max Planck Institute in Heidelberg, Germany, solved this production problem with the development of the so-called "Huffman-Kratschmer reactor."[95,96] By adjusting the gas pressure in an apparatus basically consisting of an arc welder with carbon rods, they made large quantities of soot rich in buckyballs. The arc discharge technique in general was detailed in Volume II, Chapter 9 (see Volume II, Figure 9.25), and a carbon-arc discharge setup is shown in Figure 3.51 below. A graduate student of Dr. Kroto, Jonathan Hare, succeeded in separating buckyballs from the soot produced in the carbon arc by dissolving it in benzene; ordinary graphite does not dissolve, but buckyballs form a red solution. After the benzene is evaporated, the orange-brown crystals that remain are pure fullerene carbon.

Ever since their discovery, expectations have been high that these hollow, pure carbon nanostructures could be used as drug delivery agents, as environmental tracers, as antioxidants, in medical imaging, to make stronger polymers, for more efficient fuel cell membranes, as light detectors, in superconductors, and in diamond-hard coatings for hard disks. As an antioxidant in health and medicine, a buckyball can supply an electron from one of its carbon bonds and neutralize a radical's unpaired electron. In the polymer area, DuPont and Exxon Mobil are using buckyballs to develop stronger polymers. Sony is developing a more efficient fuel-cell membrane. German industrial giant Siemens developed a buckyball-based light detector, and Seagate patented buckyballs to develop diamond-hard coatings for computer disc drives. Of the many tantalizing applications involving buckyballs, imagine buckyballs filled with a specific atom or a molecule, analogous to the ferretin liver protein caging iron oxide molecules. Because of the extreme stability of these molecules, it was difficult at first to open and close buckyballs. However, by rendering one of the thirty double bonds making up a C60 cage more reactive, Wudl at the University of California, Santa Barbara, succeeded

in opening and closing the cage, contributing to a widening class of container molecules (endohedrals) tailored to fit a variety of target molecules.[97-99]

At MIT a fullerene mass production method was developed (together with a startup company, Nano-C, Inc.) in which hydrocarbons and oxygen are mixed and the hydrocarbons are burned at a low pressure. Nano-C has a prototype burner in its headquarters that can produce 1 ton of fullerenes a year. The standard atmosphere in the burner is a low-pressure mixture of oxygen and benzene, heated to over 3140°F. By manipulating conditions inside the reactor, different types of fullerenes can be produced. Nano-C claims that they obtain a yield in fullerenes of 95% obviating the usual separation step of buckyballs from soot (http://www.nano-c.com/fullerenes.html). C60 fullerenes are the least expensive fullerene and used to fetch $25 per gram. Larger fullerenes can cost thousands of dollars per gram. Nano-C says they can now make C60 fullerenes for about $4 or $5 per gram. They expect the cost to tumble below $1 with larger-scale equipment. The Frontier Carbon Corporation (FCC), headquartered in Tokyo, Japan, is the first company in the world to mass-produce fullerenes at a commercial scale, and it increased its fullerene production capacity to 40 tons/yr in April 2003 (http://www.f-carbon.com). FCC now sells fullerenes at a price one-tenth that of the existing market and has plans to decrease its price even further.

Because of the decrease in price and unique characteristics of fullerenes, some face creams and antiaging creams based on them have already entered the market (http://www.buckyusa.com). Other near-term projected industrial applications include batteries, lubricants, grinding and polishing material, polymer additives, and rubber additives.

Very few buckyball applications are financially lucrative yet. There also might be an increase in environmental concerns as, although fullerenes are found to be bactericidal, they also seem to cause brain damage in fish.[100]

Nanotubes: 1D

Carbon nanotubes (CNTs) had been observed before their "invention" in 1991. Carbon nanofibers (diameter range, 3–100 nm; length range, 0.1–1000 mm) were known for a long time as a nuisance often emerging during catalytic conversion of carbon-containing gases. In 1959, Roger Bacon produced images of carbon nanotubes,[101] and in the 1980s, Howard Tennant even applied for a patent on a method to produce them. But it was only in 1991 that Sumio Iijima, at NEC's Fundamental Research Lab, found that through carbon self-organization it is possible to deliberately fabricate those tiny graphite tubes, perhaps the smallest electrical wires ever made.[81,85] These nanotubes are called *buckytubes*, because they are structurally similar to the carbon buckyballs. CNTs can be described as a sheet of graphite, i.e., graphene rolled into a cylinder, constructed from hexagonal rings of carbon and with caps at the ends, making the shorter ones look like pills. It is possible to open those caps through oxidation: the pentagons in the end caps of nanotubes are more reactive than the sidewall, so during oxidation, the caps easily are removed while the sidewall stays intact. The first buckytubes discovered consisted of several nested, concentric cylinders with nanometer-scale diameters, called *multiwalled nanotubes* (MWNTs) (Figure 3.46). In the nested configuration MWNTs, inner and outer tubes slide and rotate easily against each other. Like Russian nested dolls, these type of structures typically feature about ten nested tubes. By 1993, Iijima and Donald Bethune at IBM independently created single-walled nanotubes consisting of just one layer of carbon atoms (SWNT) (Figure 3.47).[102] A typical

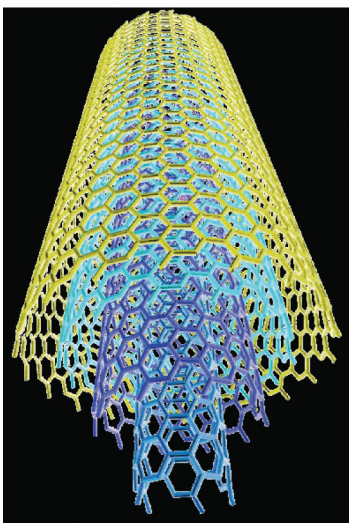


FIGURE 3.46 Artistic rendition of multiwalled nanotubes. (From A. Rochefort, Nano-CERCA, University of Montreal.)

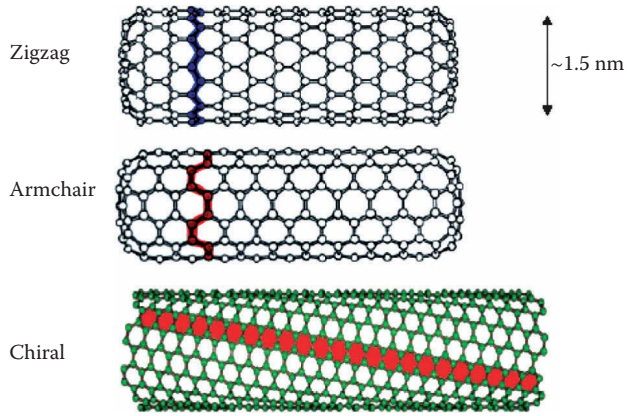


FIGURE 3.47 A single-walled carbon nanotube diameter is ~1.5 nm. Three types of lattice orientations can be identified: zigzag, armchair, and chiral nanotubes.

SWNT diameter is 1–1.5 nm, and tube lengths of several millimeters have been achieved.

Although carbon nanotubes are well ordered, STM images rarely show the sixfold symmetry of their graphene lattice. The images instead reveal the backscattering of electron waves from defects on the walls.

Nanotubes are very stiff and very stable and can be built with their length exceeding their thickness thousands of times. They fold and buckle but do not snap, and their hollow interior enables one to put materials inside.

Linear nanotubes may behave as metals or as semiconductors, depending on their diameter and the exact details of the 2D lattice orientation. Three types of lattice orientations can be identified. In armchair nanotubes, there is a line of hexagons parallel to the axis of the nanotube; in zigzag nanotubes, there is a line of carbon bonds down the center; and chiral nanotubes exhibit a twist or spiral around the nanotube (see Figure 3.47). It is the amount of twist of the lattice that determines whether a nanotube is metallic or semiconducting. Armchair nanotubes are all metallic, whereas about two-thirds of zigzag and chiral nanotubes are semiconducting (the remaining one-third are, again, metallic).

A bit more quantitative the geometry of a CNT is determined by its chiral vector (**c**) defined as follows (see Figure 3.48):

$$\mathbf{c} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 \qquad (3.13)$$

where $\mathbf{a}_1$ and $\mathbf{a}_2$ are primitive lattice vectors of graphene and n_1 and n_2 are integers (chiral numbers). The basis of this 2D lattice contains two atoms and

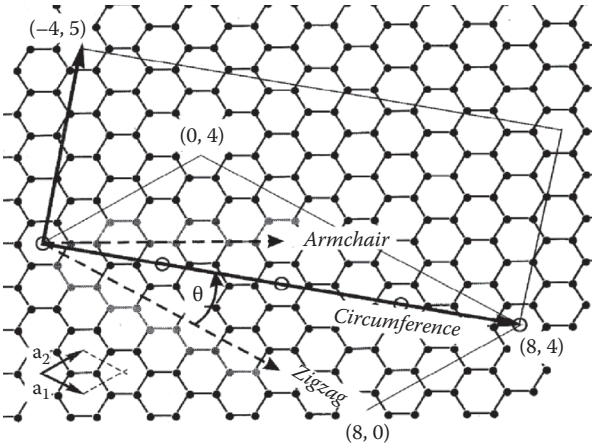


FIGURE 3.48 Graphene sheet and chiral vector.

the side length d of a hexagon in the lattice is 0.14 nm long. The chiral angle is the angle θ between the chiral vector **c** and the unit vector $\mathbf{a}_1$. If the chiral angle is 0°, $n_2 = 0$, and the nanotube is classified as zigzag. If the chiral angle is 30°, $n_1 = n_2$, and the nanotube is classified as armchair. If **c** is not along a mirror line (see Figure 3.48) the structure is chiral and these tubes, with angles between 0° and 30°, are referred to as chiral nanotubes. Chiral structures have no mirror symmetry and sort out into "left-handed" and "right-handed" versions. Due to the symmetry and unique electronic properties of graphene layers, the structure and dimensions of a nanotube strongly determines its electrical properties. For a given (n_1, n_2) nanotube, if $2n_1 + n_2 = 3q$ (where q is an integer), then the nanotube is metallic, otherwise the nanotube is a semiconductor. Thus all armchair ($n_1 = n_2$) nanotubes are metallic, and nanotubes such as (5,0), (6,4), (9,1), are all semiconducting. An alternative representation of this condition is that if $(n_1 - n_2)/3 = q$, then the SWNT is metallic. In theory, metallic nanotubes can have an electrical current density more than 1,000 times stronger than metals such as copper. The energy gap for the semiconducting fibers decreases with increasing fiber diameter d and in the limit of d_∞, we obtain the 2D case of a zero-gap semiconductor. The chirality of a SWNT can be determined by Raman spectroscopy and STM and transmission electron microscopy (TEM); STM and AFM can determine its diameter.

Also from Figure 3.48, if n is the greatest common divisor (4) of the chiral numbers n_1 (8) and n_2 (4), then the number of lattice points (open circles)

along the chiral vector is $n + 1$ (there are five open circles in Figure 3.48).

In an ordinary wire, the charge transport is diffusive, and the diffusive length of the electrons, l, is much smaller than the wire length ($l \ll L$) (see Figure 3.49a). The unique electronic properties of carbon nanotubes are due to the quantum confinement of electrons perpendicular to the axis of the tube. In that direction, the monolayer thick graphene sheet confines electrons. Around the circumference of the nanotube, periodic boundary conditions hold. For example, if a nanotube has twelve hexagons around its circumference, the thirteenth hexagonal will coincide with the first. In other words, going around the cylindrical tube just once introduces a phase difference of 2π. Because of this quantum confinement, electrons propagate along the nanotube axis, and their wave vectors point in that direction. The actual number of one-dimensional conduction and valence bands of a nanotube depends on the number of standing waves that are set up around the circumference of the nanotube. The electrons in an ideal nanotube then behave like waves traveling down a smooth channel with no atoms to bump into, and they do not undergo scattering. This allows the nanotube to pass current without resistance along its length and permits very high current densities. This type of quantum movement of an electron within a nanotube is called *ballistic transport* ($l >>> L$) (see Figure 3.49b). The contacts are part of the device and all the scattering occurs at the contacts; hence, all the voltage drops also at the contacts. The resistance, as we saw in Volume I, Chapter 3, is quantized in amounts of R = 25.8 kΩ (see also the Landauer equation: Volume I, Equation 3.274).

In metallic nanotubes, electrons can easily move from the valence to the conduction band. There is no gap between the valence and conduction band in armchair nanotubes, but as we just saw, an energy gap exists between the valence and conduction bands in about two-thirds of zigzag and chiral nanotubes. The bandgap of the semiconducting carbon nanotubes can be "tuned" from 0.5 to 1.5 eV by adjusting the diameter of the nanotube; as the nanotube diameter increases, more wave vectors are allowed in the circumferential direction and the band gap narrows. The carbon nanotube band gap is inversely proportional to the tube diameter and approaches zero at large diameters, just as for a graphene sheet (at a nanotube diameter of about 3 nm, the band gap becomes comparable to thermal energies at room temperature). The fact that the electronic structure of a carbon nanotube depends only on its physical geometry, without any doping, is unique to solid-state physics and is the basis for many of its proposed applications in electronics.

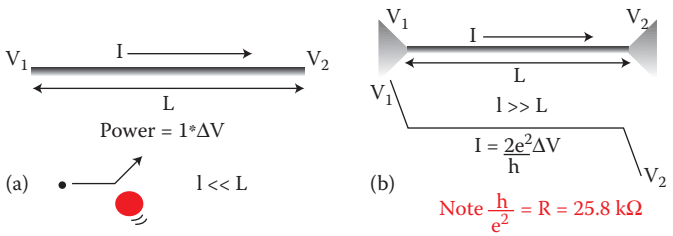


FIGURE 3.49 (a) Diffusive transport in an ordinary wire. (b) Ballistic transport in a quantum wire.

Nanotubes, we explained earlier, are made up of graphene sheets, and the unusual electronic properties of graphene arise from the fact that the carbon atom has four electrons, three of which are tied up in bonding with its neighbors. But the unbound fourth electrons are in orbitals extending vertically above and below the plane, and the hybridization of these spreads across the whole graphene sheet (see also the section below on graphene). The energy vs wavevector (**E-k**) surfaces and band diagrams of a metallic and a semiconducting carbon nanotube are compared in Figure 3.50A.

As we saw above carbon nanotubes can be viewed as a rolled up graphene sheet (Figure 3.50Aa and b). The periodic boundary condition only allows quantized wave vectors around the circumferential direction, which generates one-dimensional bands for carbon nanotubes. The nanotube can be either metallic (Figure 3.50Ac) or semiconducting (Figure 3.50Ad). If $(n_1 - n_2)/3 = q$ (with q an integer), then the SWNT is metallic. So with $n_1 - n_2 = 0$ we are dealing with a metallic tube and the bandgap is zero. And if $n_1 - n_2/3$ is not an integer we have a semiconducting tube that exhibits a forbidden band gap in the **E-k** diagram.

As we learned in Volume I, Chapter 3, carbon nanotubes have peaks in their density of states (DOS) function that are called *criticalities* or *Van Hove singularities*. These singularities in the DOS function lead to sharp peaks in optical spectra (see Figure 3.50B) and can be observed directly with scanning tunneling microscopy

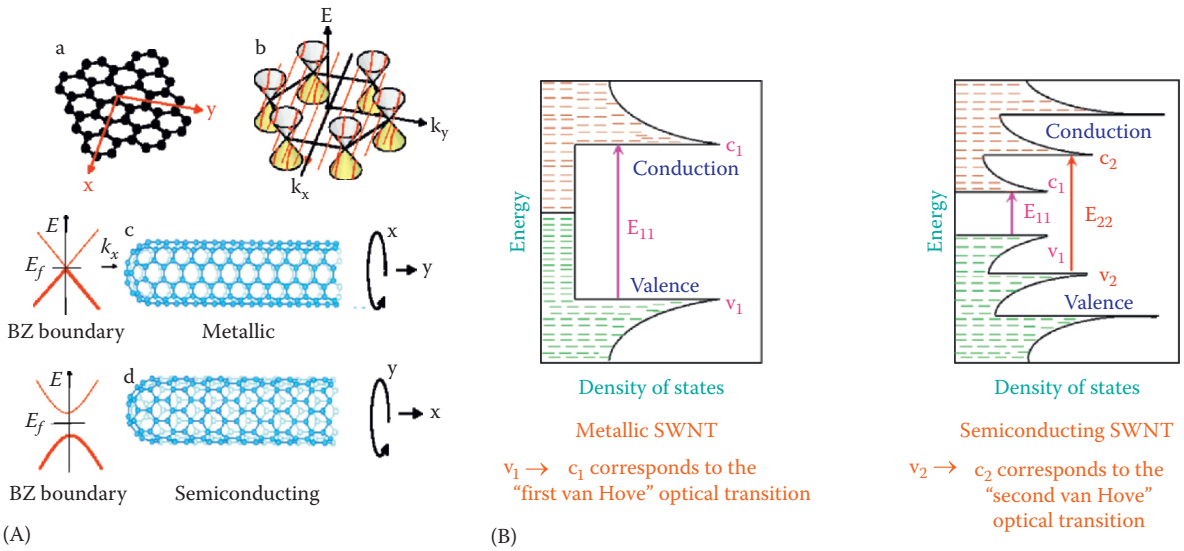


FIGURE 3.50 The E-k surfaces and band diagrams of metallic and semiconducting carbon nanotubes compared. (A) Energy vs wave vector plot. Carbon nanotubes can be viewed as rolled up graphene sheets (a,b). The periodic boundary condition only allows quantized wave vectors around the circumferential direction, which generates one-dimensional bands for carbon nanotubes. The nanotube can be either metallic (c) or semiconducting (d), depending on whether $(n_1 - n_2)$ is the multiple of 3. (B) Sharp optical transitions due to Van Hove singularities. Optical transitions between the $v_1 - c_1$, $v_2 - c_2$, etc., states of semiconducting or metallic nanotubes are traditionally labeled as S_{11}, S_{22}, M_{11}, etc. The energies between the Van Hove singularities depend on the nanotube structure. Thus by varying this structure, one can tune the optoelectronic properties of carbon nanotubes. Optical transitions are sharp (~10 meV) and strong. Consequently, it is relatively easy to selectively excite nanotubes having certain (n_1, n_2) indices, as well as to detect optical signals from individual nanotubes.

(STM) (see Volume I, Figure 3.135). At low temperatures, the derivative of the STM current with respect to voltage (dI/dV) is proportional to the density of states at the tunneling electron energy, weighted by the density of those states at the STM tip position. Since dI/dV ~ local density of states (LDOS), the LDOS is directly observed from an STM scan. In practice, during scanning tunneling spectroscopy measurements, a constant tip-sample separation is first established by fixing the set-point current at a given bias voltage. Next, the feedback of the STM is turned off and a small sinusoidal modulation is added to the DC bias voltage. The resulting current modulation (dI/dV) is then read off from a lock-in amplifier, into which the tunneling current signal from the current preamplifier is fed. The output signal from the lock-in amplifier is plotted as a function of the DC bias voltage and the tip position to generate a tunneling spectrum dI/dV (V, r), and hence the LDOS. By taking dI/dV along a line or by taking a two-dimensional dI/dV map, changes in LDOS around a impurity, or among nanotubes with different crystalline orientations can be revealed and compared with theoretical predictions. The density of states (DOS) functions for a metallic and a semiconducting carbon nanotube are illustrated in Figure 3.50B. The peaks in these DOS functions are the Van Hove singularities. In Volume I, Figure 3.135, the same DOS functions are shown in the top two plates, and the differential conductance is illustrated in the bottom two plates. Metallic carbon tubes have a constant DOS around E = 0, whereas for semiconducting ones, there is a gap at E = 0 (DOS = 0).

Carbon Nanotube Manufacture

Like fullerenes, nanotubes may be formed in macroscopic quantities in an arc discharge apparatus as described in Volume II, Chapter 9 (see Figure 9.25). In the case of a carbon arc, one refers to such a setup as a Huffman-Kratschmer reactor.[95] In a carbon arc as illustrated in Figure 3.51, two carbon rods are placed end to end with a separation distance of about 1 mm and are used as electrodes. In an enclosed volume of an inert gas (He, Ar) at a low pressure (50–70 mbar), a current of 50–100 A, driven by 20 V, creates a high-temperature arc between the electrodes. The discharge vaporizes one rod forming a deposit on the other. Producing nanotubes in high yield depends upon the uniformity in the plasma arc and also upon

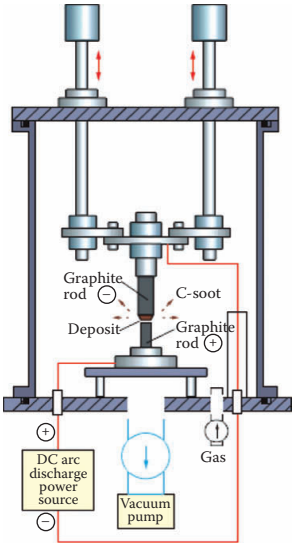


FIGURE 3.51 In the case of a carbon arc, one refers to such a setup as a Huffman-Kratschmer reactor.[95]

the temperature of the deposit on the second carbon electrode. At the extremely high temperatures of the arc plasma, a wide range of carbon materials is formed. This may include amorphous carbon, fullerenes, nanotubes, endohedrals (carbon structures with enclosures), nanocrystallites, and graphite crystals. A typical yield for nanotubes is around 30% by weight. It is the most common and arguably the easiest way to produce carbon nanotubes; however, as pointed out, it produces a mixture of items such as "soot" and catalytic metals in the end product, and the tubes are deposited in random sizes and directions.

There are a few ways of improving the process of arc discharge for producing either SWNT or MWNT (see Figure 3.52). In order to favor the production of SWNT, the anode in the apparatus is doped with a metal catalyst. The most important metals to catalyze the growth of graphitic carbon nanofibers are iron, cobalt, and nickel; chromium, vanadium, and molybdenum have also been studied. The metals have been used both as bulk particles (size typically 100 nm) and as supported particles (10–50 nm). MWNTs do not need any catalysts to grow.

By controlling the mixture of helium and argon gas, one can also control the diameter distribution of the produced nanotubes. Supposedly, this is due to the fact that the different mixtures have different diffusion coefficients and thermal conductivities.

Metal-filled carbon nanotubes (endohedrals) also have been produced with the arc-discharge method, for example, by doping a 99.4% graphite anode with a transition metal such as Cr or Ni, a rare earth such as Yb or Dy, or a covalent element such as S or Ge. A growth mechanism based on a catalytic process involving three elements, i.e., carbon, a metal, and sulfur has been proposed. The excellent protective nature of the outer graphitic cages against oxidation of the inner materials was demonstrated.[103] These carbon-coated nanoparticles are of potential importance in the study of nanowire protection against oxidation, magnetic data storage, magnetic toner xerography, magnetic inks, and ferrofluids.

To favor multiwalled carbon nanotube (MWNT) production, both electrodes in the arc discharge apparatus should be graphite (Figure 3.52). Synthesis

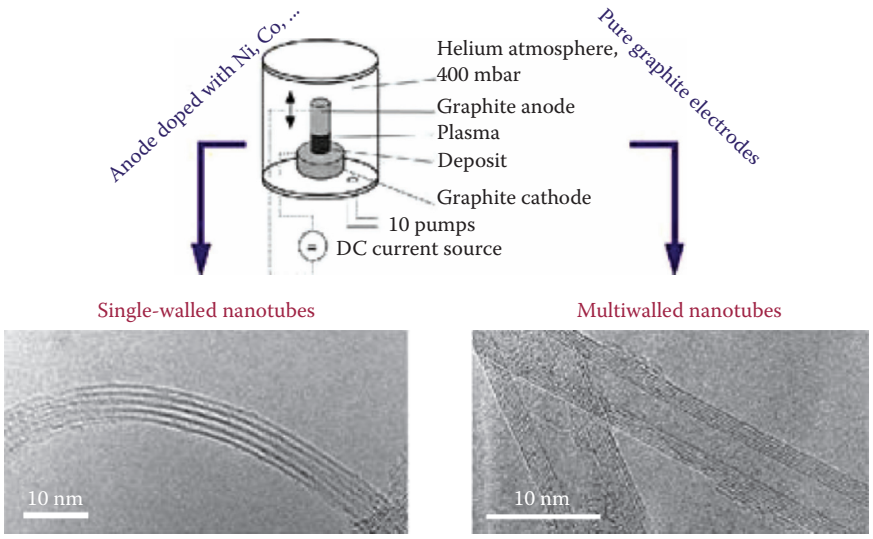


FIGURE 3.52 A carbon arc can be used to obtain single-walled and multiwalled nanotubes.

in a magnetic field and with a rotating arc discharge also improves the MWNT yield. Magnets are placed around electrodes and help align nanotubes during formation. Purification is needed in this case, which may lead to loss of structure and can cause disorder in the nanotube walls. Purification of carbon nanotubes is quite difficult because they are insoluble in organic solvents and metals are often encased in a carbon coating. Possible solutions entail acid oxidation, gas or water oxidation, filtration, and chromatography.

Besides carbon plasma arcing, other established methods to produce nanotubes are chemical vapor deposition (CVD) and laser ablation. Morinubo Endo of Shinshu University was the first to make nanotubes using CVD.[104] Pushing the field one step further, Hongjie Dai et al. at Stanford University devised a novel method for growing vertical single-walled carbon nanotubes (SWNTs) on a large scale. This group modified a plasma-enhanced chemical vapor deposition station and achieved ultrahigh-yield growth of SWNTs, thus increasing their application into commercial products.[105] By patterning a catalyst on a planar substrate they were able to control where the tubes form, for example, to make carbon nanotube arrays on a silicon wafer.[106] Typically, methane, carbon monoxide, synthesis gas (H_2/CO), ethyne, and ethene in the temperature range 700–1200 K are employed to provide the carbon atoms in the process. Dai and his colleagues discovered that the key to attaining SWNTs is to add oxygen to the CVD reaction. While carbon atoms try to form the nanotube's planar structure, hydrogen radicals are eating away at the carbon tube. Adding oxygen remedies the problem by scavenging the hydrogen radicals, creating a carbon-rich and hydrogen-deficient environment. These conditions jumpstart growth and spawn a vertical forest of nanotubes. (Dai and his colleagues were able to create 4-in. wafers blanketed with SWNTs. In addition, they devised a method for lifting the nanotubes off the original growth substrate and transferring them onto a variety of more desirable substrates such as plastics and metals—materials incompatible with the high temperatures required for nanotube growth.)

A typical CVD setup, we saw in Volume II, Chapter 7, consists of a target substrate held in a quartz tube placed inside a furnace (see Figure 3.53). The yield of this process ranges from 20% to 100%, the process is easy to scale up, and nanotubes of great length can be produced. For example, with fast-heating CVD, CNTs that are millimeters long can be grown while maintaining control over their orientation. Fast heating minimizes interaction of the growing tubes with the substrate. Using this process, Huang et al. grew a square 2D network of SWNTs guided by flow of carrier gas (Fe/Mo catalyst and CO_2 gas) (see Volume II, Figure 7.41).[108] The longest carbon nanotubes, as long as 18.5 cm, were reported in 2009. They were grown on Si substrates using an improved chemical vapor deposition (CVD).[109]

FIGURE 3.53 Plasma-enhanced chemical vapor deposition setup for nanotubes: Firstnano EasyTube 3000.

The general picture that has emerged for nanofiber formation is shown in Figure 3.54, where we contrast base growth and tip growth. In base growth, the catalyst particle can be found at the base of the nanotube, whereas in tip growth the nanoparticle is seen at the very tip of the tube. The latter seems to be the most frequent growth mode.

It is not exactly known yet how the carbon atoms in any of the manufacturingtechniques detailed here condense into tubes. Hoogenraad has proposed the following model for nucleation and growth in the case of carbon nanowires grown on Ni catalyst particles (Figure 3.55).[107] The process starts when methane decomposes into carbon and hydrogen atoms at the nickel catalyst surface (1–2). Next, H_2 molecules desorb, and carbon dissolves and forms a (substoichiometric) nickel carbide (3). This nickel carbide is metastable with respect to nickel metal and graphite so that after a short time (e.g., 10 min), the carbide phase decomposes into metallic nickel

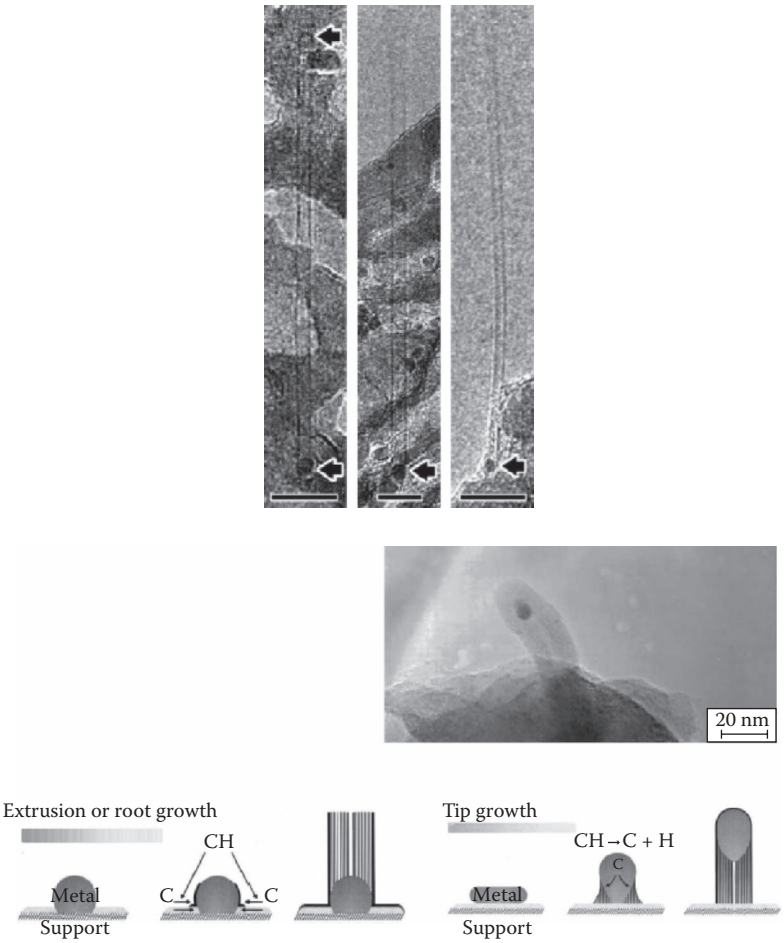


FIGURE 3.54 Base and tip growth of carbon nanotubes. Arrows point to the catalyst particle. (Hoogenraad, M.S. 1995. Utrecht University, PhD thesis. Utrecht, the Netherlands.[107])

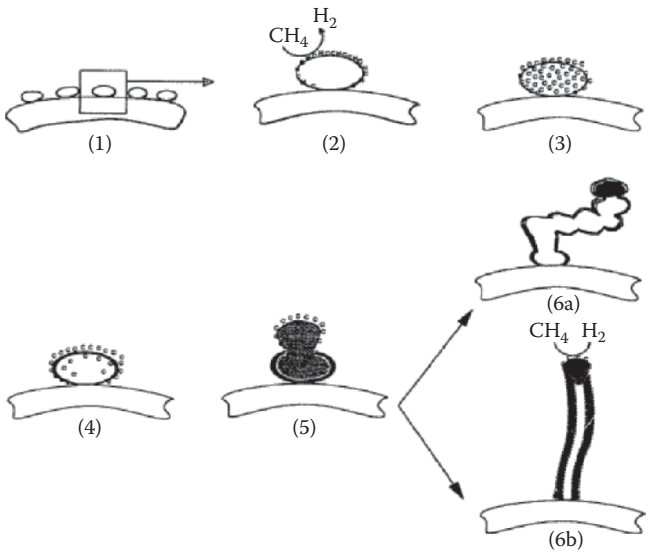


FIGURE 3.55 Mechanism for the nucleation and growth of a carbon nanofiber from methane catalyzed by a supported metal particle. (Based on the work of Hoogenraad, M.S. 1995. Utrecht University, PhD thesis. Utrecht, the Netherlands.[107])

and graphite with the graphite encapsulating the nickel particle in question (4). Hoogenraad now speculates that the metal particle is squeezed out of its encapsulation because of pressure buildup due to the formation of graphite layers at the internal surface of the graphite envelope, on the one hand, and liquid-like behavior of the metal under these conditions, on the other (5). As soon as the metal is pushed out, a fresh Ni surface is exposed to the methane and growth continues. Finally, a steady-state process occurs with either pulsed growth (6a) or smooth growth of a straight fiber (6b). This model explains why, more often than not, metal particles are found at the tip of the carbon fiber: the graphite fiber pushes the metal particle from the support and continues to grow at the back of the particle.

This type of mechanism corresponds to the vapor-liquid-solid (VLS) model as originally proposed in 1964 by R.S. Wagner and W.C. Ellis[110,140]

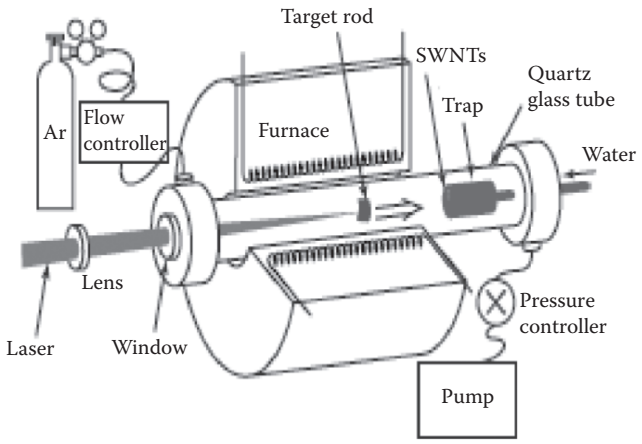


FIGURE 3.56 Laser ablation station for carbon nanotube production.[112] (With kind permission from Elsevier.)

for the Au-catalyzed Si whisker growth (see Figures 3.60 and 3.61).

Smalley made his early buckyballs by laser ablation in a furnace, as shown in Figure 3.56 (see also Volume II, Figure 7.18). The same setup can be utilized to make carbon nanotubes.

Smalley and coworkers at Rice University pioneered the use of intense laser pulses to bombard graphite rods to generate the hot carbon gas from which nanotubes settle on a copper collector. Typical yield is 70%, and the method tends to produce single-walled nanotubes. The method is quite expensive because of the use of the lasers.

In Table 3.8 we compare the pros and cons of carbon-arc discharge with chemical vapor deposition and pulse laser vaporization for the fabrication of carbon nanotubes.

Carbon Nanotube Applications

Electronic Properties

Some of the amazing properties of nanotubes are summarized in Table 3.9 below. These properties could well make carbon nanotubes (CNTs) the "Si" of this century.

It is possible for the carbon nanotube to supply the basic building blocks for an integrated circuit, as

TABLE 3.8 Comparison of the Pros and Cons of Carbon Nanotube Synthesis by Carbon-Arc Discharge with Chemical Vapor Deposition and Pulse Laser Vaporization

Method:	Carbon-arc discharge	Chemical vapor deposition	Pulse laser vaporization
Who:	Ebbesen and Ajayan, NEC, Japan 1992[113]	Endo, Shinshu University, Nagano, Japan[114]	Smalley, Rice, 1995[76]
How:	Connect two graphite rods to a power supply, place them a few millimeters apart, and throw the switch; at 100 amps, carbon vaporizes and forms a hot plasma	Place substrate in oven, heat to 600°C, and slowly add a carbon-bearing gas such as methane; as gas decomposes it frees up carbon atoms, which recombine in the form of NTs	Blast graphite with intense laser pulses; use the laser pulses rather than electricity to generate carbon gas from which the NTs form try various conditions until hitting upon one that produces prodigious amounts of SWNTs
Yield:	30–90%	20–100%	Up to 70%
SWNT:	Short tubes with diameters of 0.6–1.4 nm	Long tubes with diameters ranging from 0.6–4 nm	Long bundles of tubes (5–20 μm), with individual diameter from 1–2 nm
MWNT:	Short tubes with inner diameter of 1–3 nm and outer diameter of approximately 10 nm	Long tubes with diameters ranging from 10–240 nm	Not very much interest in this technique because it is too expensive, but MWNT synthesis is possible
Pros:	Can easily produce SWNT and MWNTs; SWNTs have few structural defects; MWNTs without catalyst, not too expensive, open-air synthesis possible	Easiest to scale up to industrial production; long length, simple process; SWNT diameter controllable, quite pure	Primarily SWNTs, with good diameter control and few defects; reaction product quite pure
Cons:	Tubes tend to be short with random sizes and directions; often needs a lot of purification	NTs are usually MWNTs and often riddled with defects	Costly technique because it requires expensive lasers and high-power equipment, but improving

Table 3.9 Some Physical Properties of Carbon Nanotubes

	Property	Remark/Comparison	Applications
Size	The average diameter of an SWNT is 1.2–1.4 nm	E-beam lithography produces lines 30 nm wide	Nanotube-tipped AFMs can trace an individual strand of DNA. The hollow structures may store hydrogen, lithium ions, or even drug molecules for slow release devices
High specific surface	>1000 m^2/g	For comparison carbon aerosol gels have a specific surface area (200–350 m^2/g)	Materials with large specific surface areas are used for energy storage as in capacitors, batteries, and fuel cells. They are also used for storage, purification, and separation of chemicals
Density	1.33–1.40 gcm^{-3} Lowest for armchair, highest for chiral	Al is 2.7 g cm^{-3}	The applications for low density materials are far ranging from consumer products such as food containers, to packaging materials, to crash protection energy absorbers
Tensile strength	Estimated to be 150 GPa	High-strength carbon steel alloys, 1 GPa. Carbon nanotubes have the greatest tensile strength of any material. It is also the stiffest known material, with a tremendously high elastic modulus (see Young's modulus)	More rugged tips for AFM and STM, superstrong composites
Resilience	Can be bent at large angle without damage	Metals and carbon fibers fracture at grain boundaries	Mechanical memory switches and nanotweezers, superstrong materials
Young's modulus	1–6 TPa (SWNT) 0.27–0.95 Tpa (MWNT)	The Young's modulus of the best nanotubes can be higher than 1000 GPa which is approximately 5x higher than steel. Reported numbers vary widely	This high Young's modulus combined with an excellent tensile strength and the lightness of carbon nanotubes, gives them great potential in applications such as aerospace (see "space elevator" in the heading of this chapter)
Current carrying capacity	Estimated at 1×10^{13} Am^{-2}	Cu wires will carry no more than 1×10^{6} A cm^{-2}	Potential for terahertz FETs
Thermal conductivity and phonon mean free path	Thermal conductivity for SWNT is as high as 3000 W/mK along the nanotube axis. The phonon mean free path length is ~100 nm	Compare this to copper, a metal well-known for its good thermal conductivity, which transmits 385 W/mK. The thermal conductivity of carbon nanotubes is dependent on the temperature and the large phonon mean free paths	Efficient heat sinking in electronics and optoelectronics
Electrical resistivity	10^{-4} ohm.cm for ropes of SWNTs at room temperature		Wiring in electronic circuitry
Field emission	Activates phosphors at 1–2 V with electrodes spaced 1 μm apart	Mo tips (Spindt cathodes, see Volume II, Figure 2.42) require fields of 50–100 V/μm	Flat-panel displays, nanovacuum tubes
Temperature stability	In vacuum up to 2800°C and 750°C in air	Metal in ICs melts between 600°C and 1000°C	High-temperature environments

Table 3.9 Some Physical Properties of Carbon Nanotubes (*Continued*)

	Property	Remark/Comparison	Applications
Quantum effects	For every nanotube circumference there is a unique band gap	Easy to manipulate electronic properties	New generation of quantum devices
Chemical sensitivity	Oxygen, halogens, and alkalis cause a drastic resistance change at room temperature	Most solid-state gas sensors only work at higher temperatures (e.g., >350°C)	Supersensitive, room temperature chemical sensors (selectivity might be an issue)
High aspect ratio	Easily 1000 : 1		The high aspect ratio of CNTs imparts electrical conductivity to polymers at lower loadings, compared to conventional additive materials, such as carbon black, chopped carbon fiber, or stainless steel fiber
Cost	~$100/g	Au, ~$10/g	High cost will make many of the above applications impossible until prices come down
Band gap	0 for metallic and 0.5 eV for semiconductors		Semiconductive type can be used as a transistor and metallic as an interconnect

For a good introductory reading on fullerenes, we suggest *The Chemistry of the Fullerenes* by Hirsch.[119] For a good introduction to nanotubes, we refer to the book *Carbon Nanotubes* by Dresselhaus et al.[120] and the special section "Carbon Nanotubes" in *Physics World*.[121]

the metallic nanotubes can act as highly conductive wires and the semiconducting nanotubes can be fabricated into transistors. Cees Dekker in Delft,[88] Paul McEuen at Cornell,[112] Phaedron Avouris at IBM,[115] and Charles Lieber at Harvard[116] all have demonstrated that single nanotubes can act as transistors. A field-effect transistor—a three-terminal switching device—can be constructed of only one semiconducting SWNT. By applying a voltage to a gate electrode, the nanotube can be switched from a conducting to an insulating state. In Volume I, Figure 1.30b, we show a single semiconducting nanotube contacted by two electrodes. The Si substrate, which is covered by a layer of SiO_2 300 nm thick, acts as a back gate. In 2002, researchers built the world's first array of transistors out of carbon nanotubes.[117]

Another carbon nanotube electronic application involves field emission. When a solid is subjected to a sufficiently high electric field, electrons near the Fermi level can be extracted from the solid by tunneling through the surface potential barrier (see Volume II, Figure 2.17). This emission current depends on both the strength of the local electric field at the surface and the material's work function (which denotes the energy necessary to extract an electron from its highest bounded state to the vacuum level). The applied electric field must be very high in order to extract an electron, a condition easily fulfilled for carbon nanotubes, because their elongated shape ensures a very large electric field concentration. For technological applications, the emissive material should also have a low threshold emission field and remain stable at very high current densities. Furthermore, an ideal emitter should be nanometer size, have high structural integrity and high electrical conductivity, exhibit a small energy spread, and demonstrate excellent chemical stability. Carbon nanotubes possess all these properties. The growth of aligned arrays of CNTs inside the pores of self-organized nanoporous alumina templates is an especially intriguing application in this regard (see Chapter 5, this volume). Govyadinov et al.[118] synthesized arrays of CNTs (deposited with catalytic CVD) in alumina pores with diameters in the sub-100-nm range and densities from 10^8 to 10^{11} tips per cm^2. These prototype CNT arrays have been used as field emitters (carbon has a low work function) and shown to exhibit field emission at a low threshold of 3–5 V/μm and emission current densities up to 100 mA/cm^2 in low-vacuum conditions.

However, a bottleneck in the use of nanotubes for field emitter application is the dependence of the conductivity and emission stability of the nanotubes

on the fabrication process and synthesis conditions. Examples of applications for nanotubes as field emitting devices are flat-panel displays (see Volume II, Figure 7.90), gas-discharge tubes in telecom networks, electron guns for electron microscopes, AFM tips, and microwave amplifiers.

Mechanical Properties

Estimates vary as to precisely how strong a single-carbon nanotube can be, but experimental results have shown their tensile strength to be in excess of 60 times stronger than high-grade steel. Depending upon the specific paper at hand, the reported numerical values for the Young's modulus range from 1–6 TPa. This scatter is known as "Yakobson's Paradox" after Boris Yakobson (Rice University) who first discussed this.[122] In a recent paper, Young Huang and co-workers address this issue and provide a resolution.[123] According to Huang et al. the errors originate in the wide scatter of assumptions about "h" the thickness of the single-walled nanotube or the graphene sheet. Theoretically, SWNTs could have a Young's modulus of between 1 and 6 TPa. MWNTs are less strong because the individual cylinders slide with respect to each other. Ropes of SWNTs are also less strong. The individual tubes can pull out by shearing, and eventually the whole rope will break. This happens at stresses far below the tensile strength of individual nanotubes. Nanotubes sustain large strains in tension without showing signs of fracture, and in other directions, nanotubes are highly flexible. Because of their flexibility, nanotubes are used, for example, in scanning probe instruments. CNTs are not nearly as strong under compression than under tension. Because of their hollow structure and high aspect ratio, they tend to undergo bucking when placed under compressive, torsional, or bending stress. Since MWNT tips are conducting, they can be used in STM and AFM instruments (Volume II, Figure 2.40). Advantages are the improved resolution in comparison with conventional Si or metal tips, while the tips do not suffer from crashes with the surfaces because of their high elasticity. However, nanotube vibration due to the large length remains an important issue, until shorter nanotubes can be grown controllably. Because of the stiffness of carbon nanotubes, they should be ideal candidates for structural applications. For example, they may be used as reinforcements in high-strength, low-weight, and high-performance composites (see below).

Energy Storage and Conversion

See Chapter 9, this volume.

Chemical Sensors and Nanoreactors

Nanotube tips can be modified chemically through attachment of functional groups. Because of this, nanotubes can be used as molecular probes, with potential applications in chemistry and biology. Because of the small channels, strong capillary forces exist in nanotubes. These forces are strong enough to hold gases and fluids in nanotubes. In this way, it may be possible to fill the cavities of the nanotubes to create nanowires. The critical issue here is the wetting characteristic of nanotubes. Because of their smaller pore sizes, filling SWNTs is more difficult than filling MWNTs. If it becomes possible to keep fluids inside nanotubes, it could also be possible to perform chemical reactions inside their cavities. In particular, organic solvents wet nanotubes easily. In this case we could speak of a nanoreactor. Table 3.10 summarizes some chemical and biological sensor applications of carbon nanotubes.

Carbon Nanotubes in Composites

One of the most important applications of carbon nanotubes is projected to be as reinforcement in composite materials. However, there have not been many successful experiments that show that nanotubes are better fillers than the traditionally used carbon fibers. The main problem is to create a good interface between the nanotubes and the polymer matrix, as nanotubes are very smooth and have a small diameter that is nearly the same as that of a polymer chain. Secondly, nanotube aggregates, which are very common, behave differently in response to loads than individual nanotubes do. Limiting factors for good load transfer could be sliding of cylinders in MWNTs and shearing of tubes in SWNT ropes. To solve this problem, the

Table 3.10 Chemical and Biological Applications of Carbon Nanotubes

System	Target Species	Salient Feature
SWNT	NH_3 and NO_2	Sensitive to 200 ppm of NO_2 and 1% of NH_3.[124]
SWNT	N_2, He, O_2, and Ar	Gas concentrations as low as 100 ppm can be detected.[125]
MWCNT	NH_3	Gas concentrations as low as 10 ppm can be detected. The sensor showed a reversible response of few minutes.[126]
Poly(*o*-anisidine)-coated CNT	HCl	Nine times increase in sensitivity compared with uncoated CNT.[127]
MWNT-SiO_2	CO_2, O_2, and NH_3	Sensor response times are approximately 45 s, 4 min, and 2 min for CO_2, O_2, and NH_3, respectively. The sensor response is reversible for O_2 and CO_2 but irreversible for NH_3.[128]
SWNT	D-Glucose	Senses D-glucose in solution phase by two distinct mechanisms of signal transduction: fluorescence and charge transfer.[129]
Poly(vinylferrocene)-derivatized MWCNT	Glucose	Glucose concentration in real blood sample can be determined.[130]

aggregates need to be broken up and dispersed or cross-linked to prevent slippage. A main advantage of using nanotubes for structural polymer composites is that nanotube reinforcements will increase the toughness of the composites by absorbing energy during their highly flexible elastic behavior. Other advantages are the low density of the nanotubes, an increased electrical conduction, and better performance during compressive load. Another possibility, which is an example of a nonstructural application, is filling of photoactive polymers with nanotubes. Poly-*p*-phenylenevinylene filled with MWNTs and SWNTs is a composite that has been used for several experiments. These composites show a large increase in conductivity with only a small loss in photoluminescence and electroluminescence yields. Another benefit is that the composite is more robust than the pure polymer. Of course, nanotube-polymer composites could be used also in other areas. For instance, they could be used in the biochemical field as membranes for molecular separations or for osteointegration (growth of bone cells). However, these areas are less well explored. The most important thing we have to know about nanotubes for efficient use of nanotubes as reinforcing fibers is knowledge of how to manipulate the surfaces chemically to enhance interfacial behavior between the individual nanotubes and the matrix material.

Conclusion

Carbon nanotubes continue to hold a tremendous promise as a structural and electronic material. Some types of nanotubes might present a danger similar to that posed by asbestos, which has tiny fibers that can get into lung tissues and cause cancerous lesions.[131] Researchers are suggesting more study and caution in handling the material.

The Rise of Graphene:*[138] 2D

Graphene is the name given to a monolayer of carbon atoms arranged into benzene rings; it is used to describe properties of many carbon-based

* Based on the paper of the same title by Geim and Novoselov[119] (*Nature Materials* 2007;6:183–91).

materials, including graphite, fullerenes, and nanotubes (see Figures 3.44 and 3.45). The graphene theory was first developed by Philip Wallace in 1947 as an approximation to help understand the electronic properties of more complex, three-dimensional graphite.[132] Landau and Peierls, 70 years ago, claimed that 2D crystals are thermodynamically unstable and that divergent contribution of thermal fluctuations in 2D crystals lattices would lead to displacements of atoms comparable to the interatomic distances and thus destroy the 2D layer. They also concluded that the melting temperature of these thin films would rapidly decrease with decreasing thickness. Furthermore, in quantum field theory and statistical mechanics, the Mermin-Wagner theorem (also known as *Mermin-Wagner-Hohenberg theorem* or *Coleman theorem*) says that crystalline order is impossible in two-dimensional structures at non non-zero temperatures, and this has been interpreted as saying that a freestanding, two-dimensional crystal would be disrupted by thermodynamic forces (hence "2D materials were presumed not to exist").[133] It was believed that because of instabilities, a 2D graphene structure would fold up, or crumple, to form a more stable 3D structure. In summary, planar graphene was presumed not to exist in the free state, being unstable with respect to the formation of curved structures such as soot, fullerenes, and nanotubes.

However, in 2004, Novoselov et al. obtained relatively large graphene sheets (eventually, up to 100 μm in size) by mechanical exfoliation (repeated peeling) of 3D graphite crystals.[82] Their motivation allegedly was to study the electrical properties of thin graphite films. As purely two-dimensional crystals were unknown before and presumed not to exist, their discovery of individual planes of graphite may be viewed as accidental. The Manchester experimental discovery of 2D crystal matter was at first put in doubt until 2005, when the groups of Andre Geim and Philip Kim of Columbia University proved that the obtained graphitic layers exhibited the expected theoretical electronic properties of graphene.[134,135] On October 5, 2010, the Royal Swedish Academy of Sciences awarded the Nobel Prize in Physics to Andre Geim and Konstantin Novoselov University of Manchester, UK, "for groundbreaking experiments regarding the two-dimensional material graphene," (http://nobelprize.org/nobel_prizes/physics/laureates/2010/press.html). Single layers of graphite were previously (starting from the 1970s) grown epitaxially on top of other materials. Epitaxial graphene also consists of a single-atom-thick hexagonal lattice of sp^2-bonded carbon atoms, as in free-standing graphene, but there is significant charge transfer from the substrate to the epitaxial graphene, and, in some cases, hybridization between the *d* orbitals of the substrate atoms and π orbitals of the graphene layer. This interaction significantly alters the electronic structure of the epitaxial graphene. The Manchester free-standing graphene on the other hand is a single-atom-thick crystal of a macroscopic size that is either suspended or interacting only very weakly with a substrate.

The electronic structure of graphene approaches the 3D limit of graphite at ten layers of graphene. Today, graphene sheets are prepared by chemical exfoliation (solution-based), epitaxial growth by CVD (same as nanotubes), epitaxy on catalytic surfaces (Ni, Pt), and mechanical cleavage. Single graphene layers are not that difficult to make (see pencil in Figure 3.45), but they are more difficult to identify. With AFM it is possible, but the throughput is very low on large areas, and SEM does not reveal the number of atomic layers. A solution involves the placement of graphene on top of a Si wafer with a carefully chosen thickness of SiO_2; in this way graphene becomes visible even in an optical microscope. An interference-like contrast appears with an SiO_2 thickness of 300 or 90 nm (see Figure 3.57).

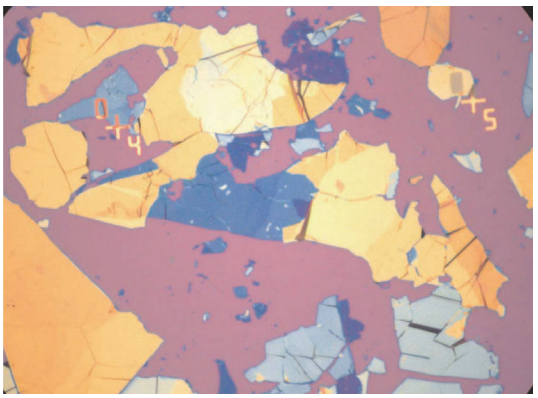


FIGURE 3.57 Graphene on SiO_2 observed in an optical microscope.[117]

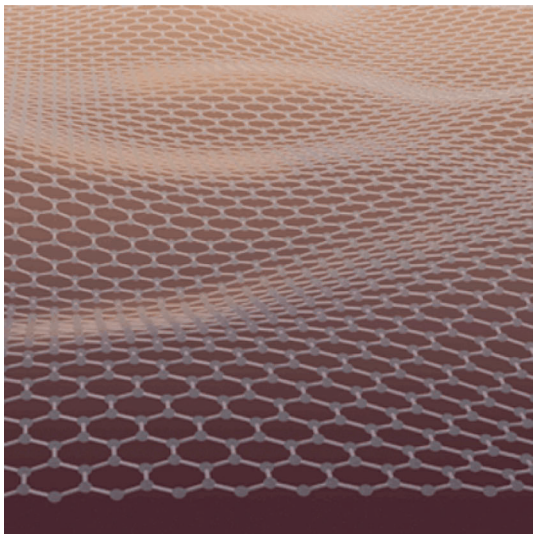


FIGURE 3.58 Graphene sheet with ripples.[119]

The atomic structure of isolated, single-layer graphene has been studied by transmission electron microscopy (TEM) on sheets of graphene suspended between bars of a metallic grid.[136] These electron diffraction patterns showed the expected hexagonal lattice of graphene but also revealed "rippling" of the flat sheet, with an amplitude of about 1 nm (see Figure 3.58). These ripples may be intrinsic to graphene as a result of the instability of two-dimensional crystals referred to above, or they may be extrinsic, originating perhaps from the ubiquitous dirt seen in all TEM images of graphene. While an infinitely large single layer of graphene would be in direct contradiction to the Mermin-Wagner theorem (see above), a finite-size 2D crystal of graphene could be stable. The observed ripples in suspended layers of graphene, it has been proposed, are caused by thermal fluctuations in the material; graphene adjusts to the thermal fluctuations, which could threaten to destroy the structure, by adjusting its bond length to accommodate the fluctuations. Within this context, it is actually debatable whether graphene is truly 2D or not, because of this natural tendency to ripple.[137]

Because graphene is somewhat tricky to make and even trickier to see and handle, most experiments have been performed not on free-standing monolayers but on two types of graphene samples: exfoliated graphene, deposited on a silicon-oxide/silicon substrate (see Figure 3.57), and epitaxial graphene, a layer of carbon atoms chemically deposited on (and chemically bonded to) a substrate of silicon carbide.

The surge of interest in the electronic properties of graphene has largely been motivated by the discovery that electron mobility in graphene is ten times higher than in commercial-grade silicon (up to 100,000 $cm^2\ V^{-1}\ s^{-1}$, limited by impurity scattering, and weakly dependent on temperature), ballistic transport is observed on the submicron scale (up to 0.3 μm at 300 K), and the quantum Hall effect can be observed at room temperature (extending the previous temperature range by a factor of 10). All these properties raise the possibility of high-efficiency, low-power, carbon-based electronics. The surprising electronic properties of graphene are ascribed to the presence of charge carriers that behave as if they are massless, "relativistic" quasiparticles called *Dirac fermions*.

A 2D graphene crystal is very different from a 3D silicon single crystal such as Si. In a semiconductor such as Si, electrons and holes interact with the periodic field of the atomic lattice to form quasiparticles. For example, in Volume I, Chapters 3 and 4, we saw how an electron moving through a conventional solid is described as having an effective mass (m_e^*) that takes into account the drag on its momentum from the surrounding crystal lattice as well as from interactions with other particles. The energy of quasiparticles in a solid depends on their momentum, a relationship described by the energy bands. In a typical 3D semiconductor, the energy bands are "parabolic."

Between them is the forbidden band gap, representing the amount of energy it takes to promote an electron from the valence band to the conduction band. The energy (E) of an electron in such a crystal depends quadratically on its momentum (p), as given by the following equation.

$$E = \frac{\hbar^2 \mathbf{k}^2}{2m_e^*} \left(= \frac{p^2}{2m_e^*} \right)$$

(Volume I, Equation 3.213)

While electron transport in most solids is accurately described by the nonrelativistic Schrödinger

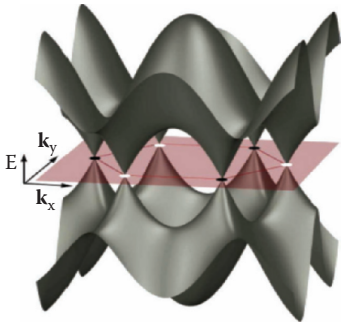


FIGURE 3.59 *E*-**k** surface of graphene. Intrinsic graphene is a semimetal or zero-gap semiconductor.

equation, graphene is quite different from most conventional three-dimensional materials. Unlike in ordinary semiconductors, graphs of the valence band and the conduction band in an *E*-**k** diagram for graphene are smooth-sided cones that meet at six points, called the *Dirac points* (see Figure 3.59). One interesting consequence of this unique band structure is that the electrons in graphene are "sort of free." The *E*-**k** relation near the Dirac points is linear for low energies, leading to zero effective mass for electrons and holes. Charge carriers in this energy range thus behave like relativistic particles (they behave like photons) described by the Dirac equation for spin 1/2 particles. The linear dispersion relationship in the low energy region is simply given by the following equation:

$$E = v_F \mathbf{k} \tag{3.14}$$

where the wave number (**k**) represents momentum and the Fermi velocity (v_F) stands in for the speed of light. Dirac fermions* travel at a constant speed—a small fraction of the speed of light.[138] Graphene has all the ideal properties to become an almost ideal component for integrated circuits. It has a high carrier mobility, as well as low noise, allowing it to be utilized as the channel in a FET. In 2008, the smallest transistor so far, one atom thick, ten atoms wide, was made of graphene.[139] IBM researchers are reporting progress in circuits based on narrow ribbons of planar graphene. The narrow graphene ribbons confine electrons in the narrow, single-planed graphene ribbons, which open up the possibility of building quantum wires and quantum wells. Narrow ribbons of graphene behave much like nanotubes and may be more amenable to manufacturing techniques such as etching. The current technology limits how thin graphene ribbons can be made. It does not allow them to be made thin enough to push to room temperature, so the work is carried out at lower temperatures.

Besides its electronic properties, graphene also is researched for its mechanical, thermal, and optical properties.

Despite all the excitement, don't expect "graphenium" microprocessors just yet. Think about carbon nanotubes, observed in 1991: they have created tremendous excitement, but today, there are no commercial CNT-based breakthrough products on the market—although they supposedly make your golf balls go straighter (a breakthrough for some golfers at least). Moreover, graphene currently is one of the most expensive materials on Earth, with a sample that can be placed on the cross-section of a human hair costing more than $1000 (as of April 2008). The price may fall dramatically, though, if and when commercial production methods are developed in the future.

Noncarbon Nanotubes and Nanowires

A nanotube, strictly speaking, is any tube with nanoscale dimensions, but generally the term is used to refer to carbon nanotubes. With a thin shell that is only one atom wide, nanotubes are hollow, giving them characteristics significantly different from those of solid nanowires made of silicon or other materials. Nanowires can be defined as structures that have a lateral size constrained to tens of nanometers or less and an unconstrained longitudinal size. At these scales, quantum mechanical effects are important. Such wires are also known as *quantum wires*. Many different types of nanowires exist, including metallic (e.g., Ni, Pt, Au), semiconducting (e.g., Si, InP, GaN, etc.), and insulating (e.g., SiO_2, TiO_2). Molecular nanowires are composed of repeating molecular units

* Dirac fermions have also been invoked to explain the quantum Hall effect in graphene, the magnetic-field-driven metal–insulator-like transition in graphite, superfluidity in ^{3}He, and the exotic pseudogap phase of high-temperature superconductors.

either organic (e.g., DNA and protein) or inorganic (e.g., $Mo_6S_{9-x}I_x$).

A major drawback to the utility of carbon nanotubes is the limited control over their chirality, which we saw above directly affects their electronic properties. At present no synthesis methods capable of controlling this feature are available, and selective synthesis of carbon nanotubes having only semiconducting characteristics remains elusive. A commonly studied noncarbon nanotube variety is made of boron nitride (BN). BN is the III-V homolog to graphene, and in this 2D lattice B and N occupy different sublattices—this lowers the symmetry and leads to new physical effects. BN nanotubes are all insulating, with a band gap of about 5 eV. Unlike carbon nanotubes, the conductive properties of BN nanotubes are more or less independent of chirality, diameter, and even the number of walls in the tube. Theoretically, the existence of silicon nanotubes with an sp^2 atomic arrangement is also predicted: creation of this type of nanotube is considered an important challenge for the future.

By making clever choices of seed crystals and growing conditions, it is possible to cause crystals to assume unusual shapes. As originally proposed by R.S. Wagner and W.C. Ellis, from Bell Labs, for the Au-catalyzed Si whisker growth, a vapor-liquid-solid (VLS) mechanism still is mostly used (Figure 3.60).[140]

As illustrated in Figure 3.61, (111)-oriented Si "whiskers" grow from a small Au particle island on a Si(111) surface when heated at 950°C and exposed

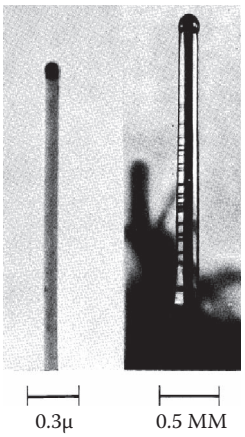


FIGURE 3.60 Si whisker growth: Au-catalyzed Si whisker growth based on a vapor-liquid-solid (VLS) mechanism.

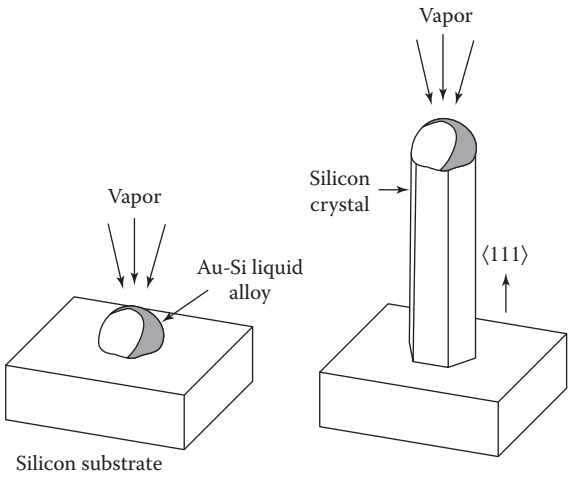


FIGURE 3.61 The vapor-liquid-solid mechanism for single-crystal Si growth.

to a flow of $SiCl_4$ and H_2 in a CVD reactor. Similar results are obtained with Pt, Ag, Pd, Cu, and Ni. When the substrate is heated up, a gold monolayer de-wets (unbinds) from the substrate in liquid form. From there, the liquid gold coalesces into nanoclusters and naturally self-assembles into gold nanoclusters. The gold acts as the catalyst in the VLS model. The liquid droplet is the preferred site for CVD deposition, and once the percentage of source material has supersaturated the nanoclusters, it begins to solidify and grow outward in a perfect single-crystal nanowire. In this growth mechanism the "impurity" is essential and remains in evidence from a small "globule" that is present at the tip of the whiskers during the growth (see Figure 3.61). The role of the impurity is to form a liquid alloy droplet at relatively low *T*. The selection of the impurity is important, as the diameter of the nanowire roughly is controlled by the diameter of the nanoclusters. The length of a nanowire is controlled by the amount of time the source is left on.

The VLS model corresponds to the model for the nucleation and growth of a carbon nanofiber from methane catalyzed by a supported metal particle shown in Figure 3.55 (based on the work of Hoogenraad[107]).

Nanoscale catalysts have been used to seed long, wire-like single crystals of compounds such as indium phosphide or gallium arsenide. The nanowire field got a shot in the arm (a rebirth, so to speak) with efforts by Charles Lieber (Harvard) (Volume I, Figure 3.65), Peidong Yang (http://www.cchem.berkeley.edu/

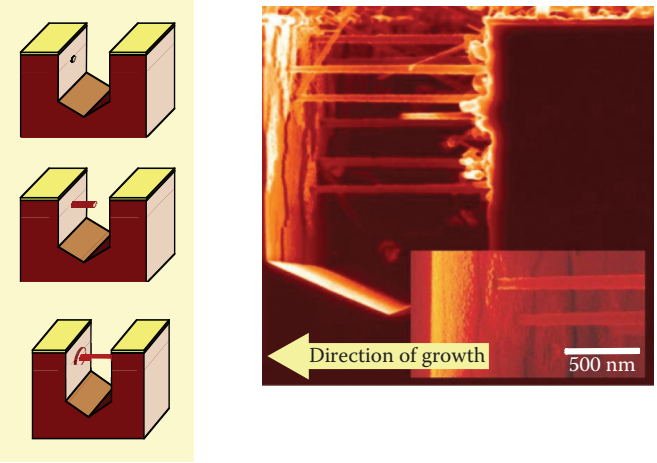


FIGURE 3.62 Lateral growth of nanowires between Si(111) surfaces. (From M. Saif Islam, S. Sharma, T.I. Kamins, and R.S. Williams, *Nanotechnology* 15;L5–L8, 2004. With permission.[141])

pdygrp/main.html), James Heath (http://www.its.caltech.edu/~heathgrp), and Hongkun Park (http://www.people.fas.harvard.edu/~hpark). Lieber's group at Harvard (http://cmliris.harvard.edu) reported arranging indium phosphide semiconducting nanowires into a simple configuration that resembled the lines in a tic-tac-toe board. The team used electron beam lithography to place electrical contacts at the ends of the nanowires in order to show that the array was electronically active.

Saif Islam et al. came up with the idea of connecting vertical Si(111) walls by growing Si nanowires laterally. Starting with a good catalyst on one wall, if there is another vertical surface in its close vicinity, the nanowire hits that surface and makes a bridge.[141] This is illustrated in Figure 3.62.

Electrochemical deposition of nanowires into a porous medium is another popular technique for making metallic nanowires. The porous medium may be porous silica, alumina, polycarbonate membranes, and even polyaniline nanotubules. Nanowire diameters as low as ~10 nm have been reported, and wire length may range from ~1 μm to >20 μm depending on the wire's diameter. Using nanoporous alumina as a template for the deposition of nanostructures of a material of choice, such as metals, semiconductors, and polymers, will be explained in Chapter 5, this volume (in Figure 5.10, the electrochemical deposition of the nanowire thermoelectric material Bi_2Te_3 is illustrated). Dresselhaus et al. show that lower-dimensional structures, such as nanowires based on Bi, BiSb, Bi_2Te_3, and SiGe nanotechnology, give additional controls that could lead to a higher ZT (see Chapter 8, this volume).[54]

TABLE 3.11 Nanowires and Their Applications

Material	Application
Si	Electronics, sensors
Cobalt	Magnetic
Germanium	Electronics, infrared detectors
Tin oxide	Chemical sensors
Indium oxide	Chemical sensors, biosensors
Indium tin oxide	Transparent conductive film in display electrodes, solar cells, organic light-emitting diodes
ZnO	UV lasers, field emission devices, chemical sensors
Copper oxide	Field emission device
Wide bandgap nitrides (GaN)	High temperature electronics, UV detectors and lasers, automotive electronics and sensors
Boron nitride	Insulator
Indium phosphide	Electronics, optoelectronics
Zinc selenide	Photonics (Q-switch, blue-green laser diode, blue UV photodetector)
Bi_2Te_3	Thermoelectric
Copper, tungsten	Electrical interconnects

Table 3.12 Nanowire Chemical Sensor Applications

Metal Oxide	Target Species	Salient Features
V_2O_5 nanofibers	1-Butylamine, toluene, propanol	Extremely high sensitivity was measured for 1-butylamine (below 30 ppb) and moderate sensitivity for ammonia. In contrast, only very little sensitivity was observed for toluene and 1-propanol vapors.[142]
SnO_2 nanobelts	CO, NO_2, ethanol	Sensitivity at the level of a few parts per billion.[143]
In_2O_3 nanowires	NH_3, NO_2	The response times have been determined to be 5 s for 100-ppm NO_2 and 10 s for 1% NH_3, and the lowest detectable concentrations are 0.5 ppm for NO_2 and 0.02% for NH_3.[144]
ZnO nanowires	Ethanol	Sensitive to ethanol concentration is in the range of 1–100 ppm. Sensitivity increases sharply as the temperature is raised from 200 to 300°C.[145]
MoO_3 nanorods	Ethanol and CO	The detection limit for ethanol and CO is lower than 30 ppm.[146]
Cd-doped ZnO nanowire	Relative humidity	Cd-doped ZnO nanowires show a clear positive temperature coefficient of resistance effect, which is quite abnormal compared with pure ZnO nanowires.[147]
SnO_2	Dimethyl methylphosphonate (DMMP)	Sensitive to 53 ppb DMMP; can be improved via doping nanobelts with catalytic additives.[148]

Advanced lithographic techniques (electron beam) are also used to pattern catalyst seeds for nanowire growth. In Table 3.11 we list some typical nanowires and their applications, and in Table 3.12, we review nanowire chemical sensor applications.

Questions

Questions by Dr. Marc Madou assisted by Mr. George O'Quinn, UC Berkeley

3.1: Explain the "fat fingers" and "sticky fingers" controversy, and suggest some possible ways that the problems associated with fat fingers and sticky fingers may be overcome.

3.2: Explain the difference between self-assembly, self-organization, and self-replication.

3.3: Why are fractal structures used consistently in nature? What benefits do these structures entail?

3.4: Make a listing of material properties that change in the nanoscale and explain why they change.

3.5: Explain the process of Stransky-Krastanov (SK) growth and how it can be used.

3.6: What effects do the chiral, zigzag, and arm-chair structures have on the properties of carbon nanotubes?

3.7: What are some important distinctions between top-down and bottom-up manufacturing (nanochemistry)?

3.8: Give some examples of bottom-up manufacturing methods.

3.9: Characterize the guiding principles involved in the growth of living organisms.

3.10: What does irreversible thermodynamics have to do with self-organization?

3.11: In the last few years, applications of quantum dots have become increasingly popular among analytical chemists. Quantum dots are now commercially available for a variety of applications, and many examples of their uses can be found in the literature. Briefly describe what quantum dots are and describe some of their applications.

3.12: Nanotechnology: hype or reality? Write one page.

References

1. Hall, J. S. 2005. *Nanofuture: What's next for nanotechnology*. Amherst, NY: Prometheus Books.
2. Royal Society. 2004. The Royal Society and The Royal Academy of Engineering.
3. Maasden, S. 2006. The assembled self of nanotechnology: the career of self-assembly as a metaphor. *EASST Conference*. Lausanne, Switzerland.
4. Rothemund, P. 2006. Folding DNA to create nanoscale shapes and patterns. *Nature* 440:297–302.
5. Prigogine, I. 1961. *Thermodynamics of irreversible processes*. New York: Wiley.
6. John, K., and M. Bar. 2005. Alternative mechanisms of structuring biomembranes: self-assembly versus self-organization. *Phys Rev Lett* 95:198101.
7. Ball, P. 1997. *Made to measure*. Princeton, NJ: Princeton University Press.
8. Jones, R. 2007. *Soft machines: Nanotechnology and life*. Oxford, United Kingdom: Oxford University Press.
9. Sellinger, A., P. M. Weiss, A. Nguygen, Y. Lu, R. A. Assink, W. Gong, and C. J. Brinker. 1998. Continuous self-assembly of organic-inorganic nanocomposite coatings that mimic nacre. *Nature* 394:256–59.
10. Achilles, T., and G. Von Kiedrowski. 1993. A self-replicating system from three starting materials. *Agnew Chem Int Ed Engl* 32:1198–201.
11. Lipson, H. 2005. Robotics: Self-reproducing machines. *Nature* 435:163–64.
12. Lehn, J. M. 1988. Supramolecular chemistry: Scope and perspectives (Nobel lecture). *Agnew Chem Int Ed Engl* 27:89–112.
13. Lehn, J. M. 1993. Supramolecular chemistry. *Nature* 260:1762–63.
14. Ringsdorf, H., B. Schlarb, and J. Venzmer. 1988. Molecular architecture and function of polymeric oriented systems: Models for the study of organization, surface recognition, and dynamics. *Agnew Chem Int Ed Engl* 27:113–58.
15. Whitesides, G. M., J. P. Mathias, and C. T. Seto. 1991. Molecular self-assembly and nanochemistry: A chemical strategy for the synthesis of nanostructures. *Science* 254:1312–18.
16. Whitesides, G. M. 1995. Self-assembling materials. *Sci Am* 146–49.
17. Pederson, C. J. 1988. The discovery of crown ethers. *Science* 241:536–40.
18. Stoddart, J. F., and D. B. Amabilimo. 1995. Interlocked and intertwined structures and superstructures. *Chem Rev* 95:2725–28.
19. Willner, I., and S. Rubin. 1996. Control of the structure and functions of biomaterials by light. *Agnew Chem Int Ed Engl* 35:367–385.
20. Stoddart, J. F., and S. A. Nepogodiev. 1998. Cyclodextrin-based catenanes and rotaxanes. *Chem Rev* 98:1959–76.
21. Stoddart, J. F., D. S. Hecht, R. J. A. Ramirez, M. Briman, E. Artukovic, K. S. Chichak, and G. Gruner. 2006. Bioinspired detection of light using a porphyrin-sensitized single-wall nanotube field effect transistor. *Nano Lett* 6:2031–36.
22. Saha, S., L. E. Johansson, A. H. Flood, H.-R. Tseng, J. I. Zink, and J. F. Stoddart. 2005. Powering a supramolecular machine with a photoactive molecular triad. *Small* 1:87–90.
23. Tomalia, D. A., H. Baker, J. Dewald, M. Hall, G. Kallos, S. Martin, J. Roeck, J. Ryder, and P. Smith. 1985. A new class of polymers: starburst-dendritic macromolecules. *Polym J* 17:117–32.
24. Newkome, G. R., Z. Yao, H. Baker, and V. K. Gupta. 1985. Cascade molecules: A new approach to micelles. *J Org Chem* 50:2003–04.
25. Newkome, G. R., C. N. Moorefield, and F. Vögtle. 1996. *Dentric molecules: Concepts, syntheses, perspectives*. Weinheim, Germany: VCH.
26. Fréchet, J. M. J., and Donald A. Tomalia. 2001. *Dendrimers and other dendritic polymers*. Weinheim, Germany: Wiley-VCH.
27. Grayson, S. M., and J. M. J. Fréchet. 2001. Convergent dendrons and dendrimers: From synthesis to applications. *Chem Rev* 101:3819–68.
28. Zeng, F., and S. C. Zimmerman. 1997. Dendrimers in supramolecular chemistry: From molecular recognition to self-assembly. *Chem Rev* 97:1681–712.
29. Zimmerman, S. C., F. W. Zeng, D. E. Reichert, and S. V. Kolotuchin. 1996. Self-assembling dendrimers. *Science* 271:1095–98.
30. Ward, M. D. 2000. Supramolecular coordination chemistry. *R Soc Chem Annu Rep A* 96:345–85.
31. Denti, S., S. Campagna, S. Serroni, M. Ciano, and V. Balzani. 1992. Decanuclear homo- and heterometallic polypyridine complexes. Synthesis, absorption spectra, luminescence, electrochemical oxidation, intercomponent energy transfer. *J Am Chem Soc* 114:2944–50.
32. Balzani, V., A. Juris, M. Venturi, S. Campagna, and S. Serroni. 1996. Luminescent and redox-active polynuclear transition-metal complexes. *Chem Rev* 96:759–833.
33. Percec, V., W. D. Cho, P. E. Mosier, G. Ungar, and D. J. P. Yeardley. 1998. Structural analysis of cylindrical and spherical supramolecular dendrimers quantifies the concept of monodendritic shape control by generation number. *J Am Chem Soc* 120:11061–70.
34. Hudson, S. D., H.-T. Jung, V. Percec, W.-D. Cho, G. Johansson, G. Ungar, and V. S. K. Balagurusamy. 1997. Direct visualization of individual cylindrical and spherical supramolecular dendrimers. *Science* 278:449–52.
35. Enomoto, M., A. Kishimura, and T. Aida. 2001. Coordination metallacycles of an achiral dendron self-assemble via metal-metal interaction to form luminescent superhelical fibers. *J Am Chem Soc* 123:5608–09.
36. Fréchet, J. M. J. 2002. Supramolecular chemistry and self-assembly special feature dendrimers and supramolecular chemistry. *Proc Natl Acad Sci USA* 99(8):4782–87.
37. Cagin, T., G. Wang, R. Martin, and W. A. Goddard, III. 1999. Molecular modeling of dendrimers for nanoscale applications. *Seventh Foresight Conference on Molecular Nanotechnology*. Santa Clara, CA.
38. Tully, D. C., K. Wilder, J. M. J. Fréchet, A. R. Trimble, and C. F. Quate. 1999. Dendrimer-based self assembled monolayers as resists for scanning probe lithography. *Adv Mater* 11:314–18.
39. Hawker, C. J., K. L. Wooley, and J. M. J. Fréchet. 1993. Unimolecular micelles and globular amphiphiles: Dendritic macromolecules as novel recyclable solubilization agents. *J Chem Soc Perkin Trans* 1:1287–97.
40. Kono, K., M. Liu, and J. M. J. Fréchet. 1999. Design of dendritic macromolecules containing folate or methotrexate residues. *Bioconjugate Chem* 10:1115–21.

41. Wells, M., and R. M. Crooks. 1996. Interactions between organized, surface-confined monolayers and vapor-phase probe molecules. 10. Preparation and properties of chemically sensitive dendrimer surfaces. *J Am Chem Soc* 118:3988–89.
42. Crooks, R. M., and A. J. Ricco. 1998. New organic materials suitable for use in chemical sensor arrays. *Acc Chem Res* 31:219–27.
43. Yoon, H. C., M. Y. Hong, and H. S. Kim. 2000. Affinity biosensor for avidin using a double-functionalized dendrimer monolayer on gold electrode. *Anal Biochem* 282:121–28.
44. Yoon, H. C., M. Y. Hong, and H. S. Kim. 2001. Reversible association/dissociation reaction of avidin on the dendrimer monolayer functionalized with a biotin-analog for a reagentless affinity-sensing surface. *Langmuir* 17:1234–39.
45. Tully, D. C., A. R. Trimble, J. M. J. Fréchet, K. Wilder, and C. F. Quate. 1999. Synthesis and preparation of ionically bound dendrimer monolayers and application towards scanning probe lithography. *Chem Mater* 11:2892–98.
46. Drexler, K. E. 1981. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc Natl Acad Sci USA* 78:5275–78.
47. Drexler, K. E. 1986. *Engines of creation: The coming era of nanotechnology.* New York: Doubleday.
48. Smalley, R. E. 2001. Of chemistry, love and nanobots. *Sci Am* 285:76–77.
49. Drexler, K. E., D. Forrest, R. A. Freitas, J. S. Hall, N. Jacobstein, T. McKendree, R. Merkle, and C. Peterson. 2001. *On physics, fundamentals, and nanorobots: a rebuttal to Smalley's assertion that self-replicating mechanical nanorobots are simply not possible*. Palo Alto, CA: Institute for Molecular Manufacturing.
50. Ho, W., and H. Lee. 1999. Single bond formation and characterization with a scanning tunneling microscope. *Science* 286:1719–22.
51. Klabunde, K. J. 2001. *Nanoscale materials in chemistry.* New York: Wiley-Interscience.
52. Asheghi, M., Y. S. Ju, and K. E. Goodson. 1997. Thermal conductivity of SOI device layers. *Proceedings of the IEEE International SOI conference, Tenaya Lodge at Yosemite*. Fish Camp, CA.
53. Dresselhaus, M. S., X. Sun, and Y. Lin. 2000. Theoretical investigation of thermoelectric transport properties of cylindrical bi nanowires. *Phys Rev B* 62:4610–23.
54. Yang, P., A. I. Hochbaum, R. Chen, R. D. Delgado, W. Liang, E. C. Garnett, M. Najarian, and A. Majumdar. 2008. Rough silicon nanowires as high performance thermoelectric materials. *Nature* 451:163–67.
55. Heath, J. R., A. Boukai, Y. Bunimovich, J. Tahir-Kheli, J. Yu, and W. A. Goddard. 2008. Silicon nanowires as efficient thermoelectric materials. *Nature* 451:168–71.
56. Arakawa, Y., and H. Sakaki. 1982. Multidimensional quantum well laser and temperature dependence of its threshold current. *Appl Phys Lett* 40:939.
57. Barthlott, W., and C. Neinhuis. 1997. The purity of sacred lotus or escape from contamination in biological surfaces. *Planta* 202:1–8.
58. Guo, Z., F. Zhou, J. Hao, and W. Liu. 2005. Stable biomimetic super-hydrophobic engineering materials. *J Am Chem Soc* 127:15670–15671.
59. Herzer, G. 1997. Nanocrystalline soft magnetic alloys. *Handbook of magnetic materials*, ed. K. H. J. Buschow, 417–61. Amsterdam: Elsevier Science B.V.
60. Herzer, G., and H. Warlimont. 1992. *Nanostruct Mater* 1:292.
61. Jacak, L. 2000. Semiconductor quantum dots: Towards a new generation of semiconductor devices. *Eur J Phys* 21:487–97.
62. Hashizume, T., R. J. Hamers, J. E. Demuth, and K. Markert. 1990. Initial stage deposition of Ag on the Si(001)-(2X1) surface studied by STM/STS. *J Vac Sci Tech A* 8:249.
63. Blackman, J. A., and A. Wilding. 1991. Scaling theory of island growth in thin films. *Europhys Lett* 16:115–20.
64. Franchi, S., G. Trevisi, L. Seravalli, and P. Frigeri. 2003. Quantum dot nanostructures and molecular beam epitaxy. progress in crystal growth and characterization of materials. *Progress in Crystal Growth and Characterization of Materials* 47:166.
65. Fukui, T. E. A. 1991. GaAs tetrahedral quantum dot structures fabricated using selective area metalorganic chemical vapor deposition. *Appl Phys Lett* 58:2018–20.
66. Murray, C. B., C. R. Kagan, and M. G. Bawendi. 1995. Self-organization of CdSe nanocrystallites into three-dimensional quantum dot superlattices. *Science* 270:1335–38.
67. Murray, C. B., D. J. Norris, and M. G. Bawendi. 1993. Synthesis and characterization of nearly monodisperse CdE (E=S, Se, Te) semiconductor nanocrystallites. *J Am Chem Soc* 115:8706–15.
68. Murray, C. B., C. R. Kagan, and M. G. Bawendi. 2000. Synthesis and characterization of monodisperse nanocrystals and close-packed nanocrystal assemblies. *Annu Rev Mater* 30:545–610.
69. Gedanken, A. 2004. Using sonochemistry for the fabrication of nanomaterials. *Ultrason Sonochem* 11(2):47–55.
70. Gogate, P. R. 2008. Cavitational reactors for process intensification of chemical processing applications: A critical review. *Chem Eng Process* 47(4):515–27.
71. Arturo Lopez-Quintela, M. 2003. Synthesis of nanomaterials in microemulsions: Formation mechanisms and growth control. *Curr Opin Coll Int Sci* 8:137–44.
72. Eckert, C. A., B. L. Knutson, and P. G. Debenedetti. 1996. Supercritical fluids as solvents for chemical and materials processing. *Nature* 383:313.
73. Reverchon, E., and R. Adami. 2006. Nanomaterials and supercritical fluids. *J Supercrit Fluids* 37:1.
74. Liu, J. C., Y. Ikushima, and Z. Shervani. 2003. Environmentally benign preparation of metal nano-particles by using water-in-CO_2 microemulsions technology. *Curr Opin Solid State Mater* 7:255.
75. Eastoe, J., A. Dupont, and D. C. Steytler. 2003. Fluorinated surfactants in supercritical CO_2. *Curr Opin Coll Int Sci* 8:267.
76. Smalley, R. E., D. T. Colbert, A. Thess, P. Nikolaev, and T. Guo. 1995. Catalytic growth of single-walled nanotubes by laser vaporization. *Chem Phys Lett* 243:49–54.
77. Jacak, L., A. Wojs, and P. Hawrylak. 1998. *Quantum dots.* Berlin: Springer Verlag.
78. Green, M., and P. O'Brien. 1999. Recent advances in the preparation of semiconductors as isolated nanometric particles: New routes to quantum dots. *Chem Comm* 22:2235–41.
79. Castro Neto, A. H., and E. A. Kim. 2008. Graphene as an electronic membrane. *Europhys Lett* 84:57 007.
80. Kroto, H. W., R. E. Smalley, R. F. Curl, J. R. Heath, and S. C. O'Brien. 1985. C60: Buckminsterfullerene. *Nature* 318:162–63.
81. Iijima, S. 1991. Helical microtubules of graphitic carbon. *Nature* 354:56–58.

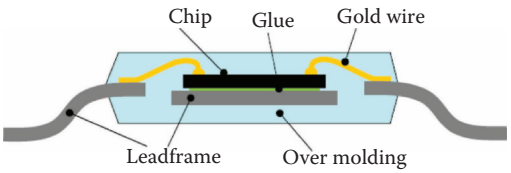
82. Novoselov, K. S., A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov. 2004. Electric field effect in atomically thin carbon films. *Science* 306:666–69.
83. Mintmire, J. W., B. I. Dunlap, and C. T. White. 1992. Are fullerene tubules metallic? *Phys Rev Lett* 68:631–34.
84. Overney, G., W. Zhong, and D. Tomanek. 1993. Structural rigidity and low frequency vibrational modes of long carbon tubules. *Zeitschrift für Physik D Atoms, Molecules and Clusters* 27:93–96.
85. Iijima, S., and T. Ichihashi. 1993. Single-shell carbon nanotubes of 1-nm diameter. *Nature* 363:603–05.
86. Rinzler, A. G., J. H. Hafner, P. Nikolaev, P. Nordlander, D. T. Colbert, R. E. Smalley, L. Lou, S. G. Kim, and D. Tománek. 1995. Unraveling nanotubes: Field emission from an atomic wire. *Science* 269:1550–53.
87. Dillon, A. C., K. M. Jones, T. A. Bekkedahl, C. H. Kiang, D. S. Bethune, and M. J. Heben. 1997. Storage of hydrogen in single-walled carbon nanotubes. *Nature* 386: 377–79.
88. Dekker, C., S. J. Tans, and A. R. M. Verschueren. 1998. Room-temperature transistor based on a single carbon nanotube. *Nature* 393:49–52.
89. Smith, B. W., M. Monthioux, and D. E. Luzzi. 1998. Encapsulated C60 in carbon nanotubes. *Nature* 396:323–24.
90. Smith, B. W., D. E. Luzzi, M. Monthioux, B. Burteaux, A. Claye, and J. E. Fischer. 1999. Abundance of encapsulated C60 in single-wall carbon nanotubes. *Chem Phys Lett* 310:21–24.
91. Tománek, D., S. Berber, and Y. Kwon. 2000. Unusually high thermal conductivity of carbon nanotubes. *Phys Rev Lett* 84:4613–16.
92. Collins, P. G., M. S. Arnold, and P. Avouris. 2001. Engineering carbon nanotubes and nanotube circuits using electrical breakdown. *Science* 292:706–09.
93. Kociak, M., A. Y. Kasumov, S. Guéron, B. Reulet, I. I. Khodos, Y. B. Gorbatov, V. T. Volkov, L. Vaccarini, and H. Bouchiat. 2001. Superconductivity in ropes of single-walled carbon nanotubes. *Phys Rev Lett* 86:2416–19.
94. Schedin, F., A. K. Geim, S. V. Morozov, E. W. Hill, P. Blake, M. I. Katsnelson, and K. S. Novoselov. 2007. Detection of individual gas molecules adsorbed on graphene. *Nat Mater* 6:183–91.
95. Saito, Y. 1994. Synthesis and characterization of carbon nanocapsules encaging metal and carbide crystallites. San Francisco: Fullerenes.
96. Kratschmer, W., L. D. Lamb, K. Fostiropoulos, and D. R. Huffman. 1990. Solid C_{60}: A new form of carbon. *Nature* 347:354.
97. Hummelen, J. C., M. Prato, and F. Wudl. 1995. There is a hole in my Bucky. *J Am Chem Soc* 117:7003.
98. Nemecek, S. 1995. A tight fit. *Sci Am* 34–36.
99. Akasaka, T., F. Wudl, and S. Nagase. 2010. *Chemistry of Nanocarbons*. Chichester: Wiley.
100. Holmes, B. 2004. Buckyballs cause brain damage in fish. *New Scientist*, March 29. http://www.newscientist.com/news/news.jsp?id=ns99994825.
101. Bacon, R. 1960. Growth, structure, and properties of graphite whiskers. *J Appl Phys* 31:283–90.
102. Bethune, D. S., C. H. Klang, M. S. D. Vries, G. Gorman, R. Savoy, J. Vazquez, and R. Beyers. 1993. Cobalt-catalysed growth of carbon nanotubes with single-atomic-layer walls. *Nature* 363:605–07.
103. Seraphin, S. 1995. Single-walled tubes and encapsulation of nanocrystals into carbon clusters. *Electrochem Soc* 142:290–97.
104. Endo, M., K. Takeuchi, S. Igarashi, K. Kobori, M. Shiraishi, and H. W. Kroto. 1993. The production and structure of pyrolytic carbon nanotubes (PCNTs). *J Phys Chem Solids* 54:1841–48.
105. Dai, H., Y. Li, D. Mann, M. Rolandi, W. Kim, A. Ural, S. Hung, A. Javey, J. Cao, D. Wang, E. Yenilmez, Q. Wang, J. F. Gibbons, and Y. Nishi. 2004. Preferential growth of semiconducting single-walled carbon nanotubes by a plasma enhanced CVD method. *Nano Lett* 4:317–21.
106. Dai, H., J. Kong, H. T. Soh, A. M. Cassell, and C. F. Quate. 1998. Synthesis of individual single-walled carbon nanotubes on patterned silicon wafers. *Nature* 395: 878–81.
107. Hoogenraad, M. S. 1995. Growth and utilization of carbon fibrils. PhD thesis, Utrecht University, Utrecht, the Netherlands.
108. Huang, S., M. Woodson, R. E. Smalley, and J. Liu. 2004. Growth mechanism of oriented long single walled carbon nanotubes using "fast-heating" chemical vapor deposition process. *Nano Lett* 4:1025–28.
109. Wang, X., Q. Li, J. Xie, Z. Jin, J. Wang, Y. Li, K. Jiang, and S. Fan. 2009. Fabrication of ultralong and electrically uniform single-walled carbon nanotubes on clean substrates. *Nano Lett* 9(9):3137–41.
110. Wagner, R. S. 1970. VLS mechanisms of crystal growth. In *Whisker technology*, ed. A. P. Levitt, 47–119. New York: Wiley.
111. Ando, Y., X. Zhao, T. Sugai, and M. Kumar. 2004. Growing carbon nanotubes. *Materials Today* 7:22–29.
112. McEuen, P. L., A. Zettl, J. Hone, M. Bockrath, A. G. Rinzler, and R. E. Smalley. 2000. Chemical doping of individual semiconducting carbon-nanotube ropes. *Phys Rev B* 61, R10606–08.
113. Ebbesen, T. W., and P. M. Ajayan. 1992. Large-scale synthesis of carbon nanotubes. *Nature* 358:220.
114. Endo, M., K. Takeuchi, S. Igarashi, K. Kobori, M. Shiraishi, and H. W. Kroto. 1993. The production and structure of pyrolytic carbon nanotubes (PCNTs). *J Phys Chem Solids* 54:1841–48.
115. Avouris, P., T. Hertel, R. Martel, T. Schmidt, and H. R. Shea. 1998. Single- and multi-wall carbon nanotube field-effect transistors. *Appl Phys Lett* 73:2447–49.
116. Lieber, C. M., P. Yang, M. Ouyang, and J. Hu. 1999. Controlled growth and electrical properties of heterojunctions of carbon nanotubes and silicon nanowires. *Nature* 399:48–51.
117. Dai, H., A. Ural, Y. Li, A. Javey, and Q. Wang. 2002. Carbon nanotube transistor arrays for multistage complementary logic and ring oscillators. *Nano Lett* 2:929–32.
118. Govyadinov, A., P. Mardilovich, and D. Routkevitch. 2000. Field emission cathode from aligned arrays of carbon nanotubes. *Electrochemical Society Abstracts* 553: 2000–28.
119. Hirsch, A. 1994. *The chemistry of fullerenes.* Stuttgart, Germany: George Thieme Verlag.
120. Dresselhaus, M. S., G. Dresselhaus, and P. Avouris. 2000. *Carbon nanotubes.* Heidelberg, Germany: Springer-Verlag.
121. Special section: Carbon nanotubes. 2000. *Phys World* 13:29–53.
122. Yakobson, B. I., C. J. Brabec, and J. Bernholc. 1996. Nanomechanics of carbon tubes: Instabilities beyond linear response, *Phys Rev Lett* 76:2511–14.
123. Peng, J., J. Wu, K. C. Hwang, J. Song, and Y. Huang. 2008. Can a single-wall carbon nanotube be modeled as a thin shell? *J Mech Phys Solids* 56(6):2213–24.

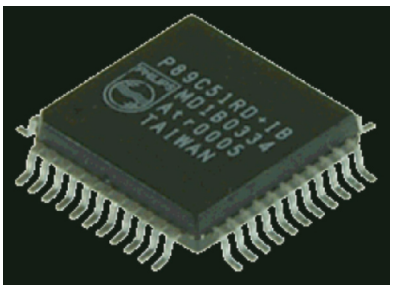
124. Zettl, A., P. G. Collins, K. Bradley, and M. Ishigami. 2000. Extreme oxygen sensitivity of electronic properties of carbon nanotubes. *Science* 287:1801–04.
125. Rao, A. M., N. Gothard, K. McGuire, S. Chopra, and A. Pham. 2003. Selective gas detection using a carbon nanotube sensor. *Appl Phys Lett* 83:2280–82.
126. Hara, M., G. Zhou, and J. Suehiro. 2003. Fabrication of a carbon nanotube-based gas sensor using dielectrophoresis and its application for ammonia detection by impedance spectroscopy. *J Phys D Appl Phys* 36, L109–14.
127. Kenny, J. M., L. Valentini, V. Bavastrello, E. Stura, I. Amentano, and C. Nicolini. 2004. Sensors for inorganic vapor detection based on carbon nanotubes and poly(o-anisidine) nanocomposite material. *Chem Phys Lett* 383:617–22.
128. Grimes, C. A., K. G. Ong, and K. Zeng. 2002. A wireless, passive carbon nanotube-based gas sensor. *Sensors J* 2:82–88.
129. Strano, M. S., D. A. Heller, S. Baik, and P. W. Barone. 2005. Near-infrared optical sensors based on single-walled carbon nanotubes. *Nat Mater* 4:86–92.
130. Compton, R. G., B. Ljuki, C. E. Banks, C. Salter, and A. Crossley. 2006. Electrochemically polymerised composites of multi-walled carbon nanotubes and poly(vinylferrocene) and their use as modified electrodes: Application to glucose sensing. *Analyst* 131:670–77.
131. Lison, D., J. Mullera, F. Huauxa, N. Moreaub, P. Missona, J. Heiliera, M. Delosc, M. Arrasa, A. Fonsecab, and J. B. Nagyb. 2005. Respiratory toxicity of multi-wall carbon nanotubes. *Toxicol Appl Pharmacol* 207:221–31.
132. Wallace, P. R. 1947. The band theory of graphite. *Phys Rev* 71:622–34.
133. Wagner, H., and N. D. Mermin. 1966. Absence of ferromagnetism or antiferromagnetism in one- or two-dimensional isotropic Heisenberg models. *Phys Rev Lett* 17: 1133–36.
134. Novoselov, K. S., A. K. Geim, D. Jiang, F. Schedin, T. J. Booth, V. V. Khotkevich, and S. V. Morozov. 2005. Two-dimensional atomic crystals. *Proc Natl Acad Sci USA* 105:10451–53.
135. http://nobelprize.org/nobel_prizes/physics/laureates/2010/press.html
136. Geim, A. K., and A. H. MacDonald. 2007. Graphene: Exploring carbon flatland. *Phys Today* 60:35–41.
137. Novoselov, K. S., A. K. Geim, J. C. Meyer, M. I. Katsnelson, T. J. Booth, and S. Roth. 2007. The structure of suspended graphene sheets. *Nature* 446:60–63.
138. Geim, A. K., and K. S. Novoselov. 2007. The rise of graphene. *Nat Mater* 6:183–91.
139. Novoselov, K. S., A. K. Geim, A. Das, S. Pisana, B. Chakraborty, S. Piscanec, S. K. Saha, U. V. Waghmare, H. R. Krishnamurthy, A. C. Ferrari, and A. K. Sood. 2008. Monitoring dopants by raman scattering in an electrochemically top-gated graphene transistor. *Nat Nanotechnol* 3:210–15.
140. Wagner, R. S., and W. C. Ellis. 1964. Vapor-liquid-solid mechanism of single crystal growth. *Appl Phys Lett* 4:89–90.
141. Saif Islam, M., S. Sharma, T. I. Kamins, and R. S. Williams. 2004. Ultrahigh-density silicon nanobridges formed between two vertical silicon surfaces. *Nanotechnology* 15:L5–L8.
142. Raible, I., M. Burghard, U. Schlecht, A. Yasuda, and T. Vossmeyer. 2005. V2O5 nanofibres: Novel gas sensors with extremely high sensitivity and selectivity to amines. *Sensors Actuators B Chem* 106:730–35.
143. Comini, E., G. Faglia, G. Sberveglieri, Z. Pan, and Z. L. Wang. 2002. Stable and highly sensitive gas sensors based on semiconducting oxide nanobelts. *Appl Phys Lett* 81:1869–71.
144. Li, C., D. Zhang, X. Liu, S. Han, T. Tang, J. Han, and C. Zhou. 2003. In2O3 nanowires as chemical sensors. *Appl Phys Lett* 82:1613–15.
145. Wan, Q., Q. H. Li, Y. J. Chen, T. H. Wang, X. L. He, J. P. Li, and C. L. Lin. 2004. Fabrication and ethanol sensing characteristics of ZnO nanowire gas sensors. *Appl Phys Lett* 84:3654–56.
146. Comini, E., L. Yubao, Y. Brando, and G. Sberveglieri. 2005. Gas sensing properties of MoO3 nanorods to CO and CH_3OH. *Chem Phys Lett* 407:368–71.
147. Wan, Q., Q. H. Li, Y. J. Chen, T. H. Wang, X. L. He, X. G. Gao, and J. P. Li. 2004. Positive temperature coefficient resistance and humidity sensing properties of Cd-doped ZnO nanowires. *Appl Phys Lett* 84:3085–87.
148. Yu, C., Q. Hao, S. Saha, L. Shi, X. Kong, and Z. L. Wang. 2005. Integration of metal oxide nanobelts with microsystems for nerve agent detection. *Appl Phys Lett* 86:063101.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

4

Packaging, Assembly, and Self-Assembly

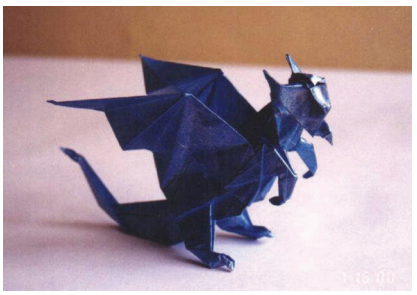


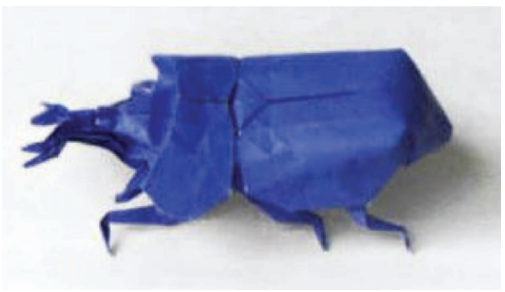


The quad flat pack (QFP) is a package for chips with 100 and more leads.

Outline

Introduction
Packaging
Level 1 Packaging (L1) and Level 2 Packaging (L2)
Electrical Connection Methods
Bonding of Layers in MEMS
Examples of MEMS Packages
Next-Generation Cap-on-Chip Cavity Sealing
Assembly and Self-Assembly
Self-Assembly Examples
Questions
References





Programmable self-assembly: constructing global shapes using biologically inspired local interactions and origami mathematics. (Courtesy of Radhika Nagpal, Massachusetts Institute of Technology.)

> Once the components were expensive and the wires cheap.
> Now the components are cheap and the wires expensive.
>
> **Len Schaper**
> *Arkansas University*

> The aboriginal people of Australia believed that the world began only when their Ancestor sang it into existence.
>
> **Bruce Chatwin**
> *The Songlines*

In the end the supremacy of organic biomaterials is tied in with the question of scale. Organic machinery can be made much smaller. Such clever

> things become possible as sockets, which can recognize, hold and manipulate other molecules. In any competition to do with molecular control the system with the smallest fingers will win.
>
> **Graham Cairns-Smith**
> *1985*

> What is the Heart, but a Spring; and the Nerves, but so many Springs; and the Joynts, but so many Wheeles, given motion to the whole Body, such as was intended by the Artificer.
>
> **Thomas Hobbes**

Introduction

This chapter addresses packaging, assembly, and self-assembly. Under packaging of micromachines, we consider issues that differ the most from integrated circuit (IC) packaging; with micromachines one faces the paradox of needing hermetic isolation from the environment that must be sensed. This has proven to be a challenge for mechanical sensors, such as for pressure and, to a lesser extent, for accelerometers, but it is the Achilles' heel of chemical and biological sensors. We distinguish between wafer-level (batch) packaging steps and serial packaging processes and learn that micromachining itself plays an important role in transforming costly serial packaging steps into less costly, wafer-level processes.

During packaging, miniature components are put together into a housing structure in a microassembly step. This task becomes more cumbersome with smaller building blocks and will increasingly necessitate the use of parallel self-assembly techniques. Self-assembly is imperative when components are nanoscale.

Packaging

Introduction

In electronic device and mechanical micromachine fabrication, processes can be divided into three major groups.

1. First, fabrication with additive and subtractive processes, as described in Volume II. These also are called *front-end processes*.
2. Packaging, involving processes such as bonding, wafer scribing, lead attachment, and encapsulation in a protective body.
3. Testing, including package leak tests, electrical integrity, and sensor functionality.

The last two process groups incorporate the most costly steps and are called *back-end processes*. In the case of a micromachine, which often interfaces with a hostile environment and where testing might involve chemical and/or mechanical parameters, packaging and testing are yet more difficult and expensive than in the case of ICs. More than 70% of the sensor cost may be determined by its package, and the physical size of a micromachined sensor often is dwarfed by the size of its package. Most conventional packaging approaches are space inefficient, with volumetric efficiencies, even in the case of ICs, often less than 1%.[1]

Contrasting the inefficiency of human packaging methods with nature's excellence at packaging is striking. As an example, consider how DNA is packed in the nucleus of a cell. Biomolecules in a solution coil up with a gyration radius r_g given by the de Gennes formula, $r = (Lb)^{1/2}$, where L is the molecule's contour length and b is determined by the degree to which the molecule resists thermal motion.[2] The total length of all DNA in a human cell is 2 m, which is a million times the cell nucleus diameter. For b of DNA, a value of 100 nm has been determined. The resulting radius of the 2-m DNA molecule is calculated at 0.45 mm, 1000 times the nucleus diameter. How does nature pack all this into the nucleus of a cell? The answer is efficient packaging. Like a thread on a spool, DNA is wound around a set of nuclear proteins—the histones. The molecule is wound twice around one "spool" and then passes onto the next, and so on. A necklace of histones is neatly packed into a chromosome. This packaging scenario is illustrated in Figure 4.1. This way, a 0.45-mm polymer coil is squeezed into a nucleus of less than a micrometer.[3]

Packaging issues of ICs and mechanical micromachines are somewhat similar, whereas packaging of devices that are used in fluids (e.g., level sensor) or that contain liquids (e.g., some electrochemical sensors) are totally different. The latter require very

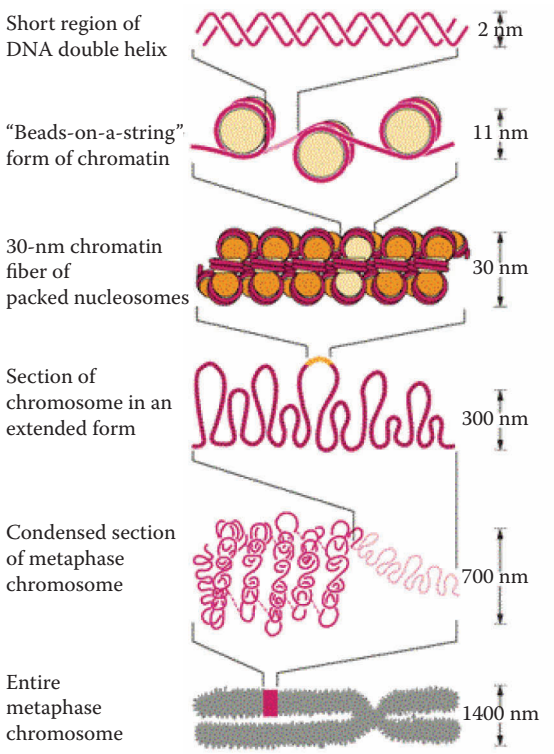


FIGURE 4.1 DNA is packaged very compactly in the cell. In eukaryotic cells, it is assembled into chromosomes, which have a hierarchical structure. The double-stranded DNA is wound twice around disc-like protein "beads" made from eight separate protein units called *histones*. The DNA-histone assembly is a nucleosome. The DNA chain, with nucleosomes all along its length, is then wound into a helix, which is itself packaged into a compact bundle in the chromosome.

different partitioning solutions and often must face biocompatibility issues as well.

Packaging in ICs and MEMS

Packaging is an essential bridge between components and their environment. In ICs, packaging provides at least four functions:

1. Signal redistribution
2. Mechanical support
3. Power distribution
4. Thermal management

In chemical and biological MEMS, the package might have several additional functions, for example:

1. Biocompatibility for in vivo devices
2. Fluidic manifold to guide liquids to different parts of the MEMS structure
3. Window/opening for analyte to interact with the chemical sensitive part of the MEMS structure

The signal redistribution is easy to understand: the electrical contacts of an IC are too closely spaced to accommodate the interconnection capacity of a traditional printed wiring board (PWB). Packaging redistributes contacts over a larger, more manageable surface—it fans out the electrical path. The mechanical support provides rigidity, stress release, and protection from the environment (e.g., against electromagnetic interference). Power distribution is similar to signal distribution except that power delivery systems are more robust than signal delivery systems, and the thermal management function is there to provide adequate thermal transport to sustain operation for the product lifetime.[1]

The various levels of packaging and their interconnections in the IC industry that we adapt here also for micromachines are summarized in Table 4.1[1] and illustrated in Figure 4.2a.[4,5] The lowest level in the IC electronics (L0 in the hierarchy) is single IC features interconnected on a single die with IC metallization lines into a level 1 IC or discrete component. Single-chip packages and multichip modules are at level 2. Levels 3 and 4 are the PWB (L3) and chassis or box (L4), respectively. Level 5 is the system itself. The hierarchy for packaging a micromachine such as a pressure sensor is illustrated in Figure 4.2b and 4.2c. For more information on advanced packaging schemes in ICs, we refer to the excellent reviews by Lyke,[1] Jensen,[4] and Lau.[6]

Packages for microelectronic systems have evolved over five decades to accommodate devices that largely have no need to interact with the environment. Packages for micromachines have all the same functions as listed for ICs but are considerably more complex, as they serve to protect from the environment while, somewhat in contradiction, enabling interaction with that environment so as to measure or affect the desired physical or chemical parameters. With the advent of integrated sensors, the protection problem became more complicated. Silicon circuitry is sensitive to temperature, moisture, magnetic field,

Table 4.1 Connection between Packaging Elements at Various Levels in the Hierarchy for ICs and Micromachines

Level	Element	Interconnected by
Level 0	Transistor within IC or resonator in MEMS	IC metallization
Level 1	ICs or discrete components such as a MEMS silicon/glass pressure sensor	Package leadframes (single-chip) or multichip module interconnection system
Level 2	Single- and multichip IC packages or a pressure sensor in a TO header in MEMS	Printed wiring boards
Level 3	Printed wiring boards	Connectors/backplanes (busses)
Level 4	Chassis or box	Connectors/cable harnesses
Level 5	System itself, e.g., a computer (IC-based) or a gas alarm system (MEMS-based)	

Source: Adapted from Lyke, J. C. 1995. In *Microengineering technologies for space systems*, ed. H. Helvajian, 131–80. El Segundo, CA: Aerospace Corporation.[1]

electromagnetic interference, and light, to name just a few items. The package must protect the onboard circuitry while simultaneously exposing the sensor to the effect it measures. For example, in an in vivo integrated sensor, a true hermetic seal is necessary to protect the circuitry from the effects of blood, and the housing needs to be biocompatible as well. The housing also might be equipped with a fluidic manifold, for example, to guide the analyte to the sensor surface. Sometimes the circuit itself can be used to reduce packaging concerns, for example, to transmit data about effects which cannot be screened out, as in the removal of a temperature effect from a pressure signal. But more frequently, integration of electronics with MEMS, especially when dealing with measuring chemical or biological signals,

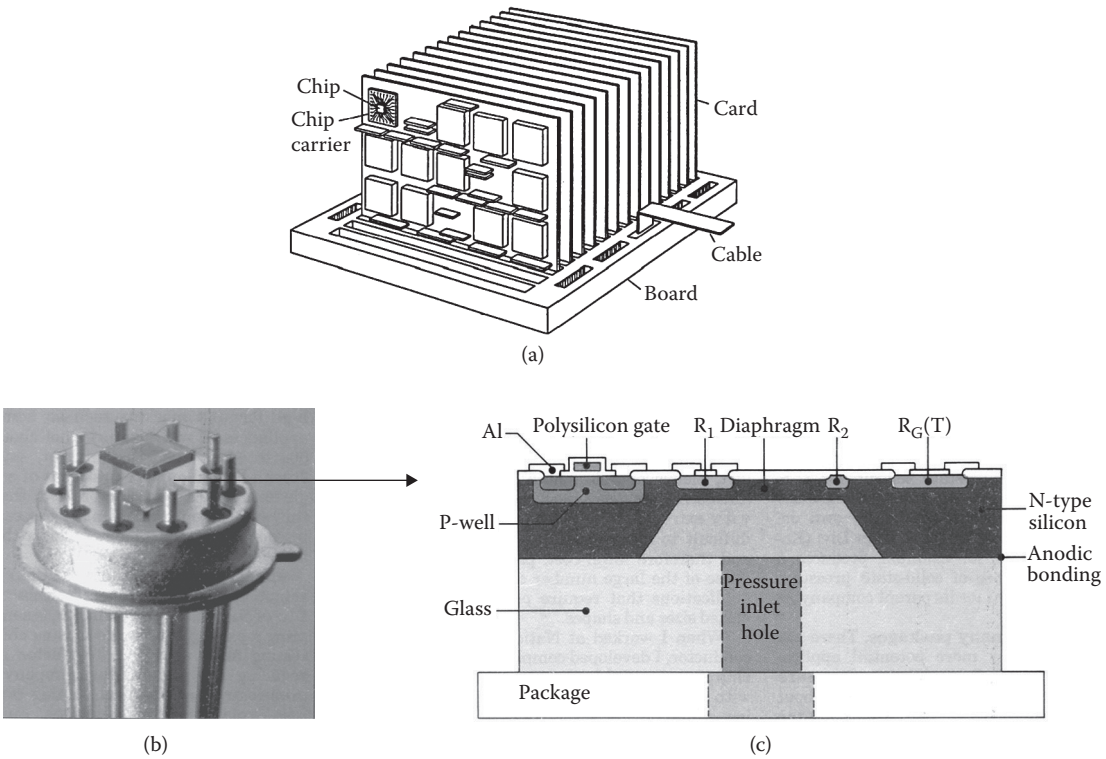


FIGURE 4.2 The packaging hierarchy in the integrated circuit (IC) world and in micromachining. (a) A large electronic unit with ICs. Transistors on the IC are level L0. The IC itself is level L1. The chip carrier is level L2, and the printed wiring board (card) is level L3. The chassis (or box) is level L4. Level 5 (L5), not shown, is the system itself, e.g., a computer. (From Jensen, J. R. 1989. In *Microelectronics processing,* ed. D. W. Hess and K. F. Jensen, 441–504. Washington, DC: American Chemical Society. With permission.[4]) (b) and (c) A micromachined unit: piezoresistive pressure sensor from NEC. L0, diaphragm on die; L1, Si/glass piece; L2, sensor mounted in TO header.

causes more problems than it solves. A package must also provide communication links, remove heat, and enable handling and testing. The material used must be one that will afford physical protection against normal process handling during and after assembly, testing, and prescribed mechanical shock. Chemical protection is also required during assembly and in practical usage. For example, during the process of installing dice on substrates, the package undergoes a series of cleanings and, under normal usage, the package may be exposed to oxygen, moisture, oil, gasoline, and salt water. The package must also be capable of providing an interior environment compatible with device performance and reliability; for example, a high-Q* resonator might need a good vacuum. Physical sensors usually have a physical barrier layer or fluid; for example, a pressure transmission fluid interposed between the sensing Si elements and the environment outside, making the encapsulation problem less severe.

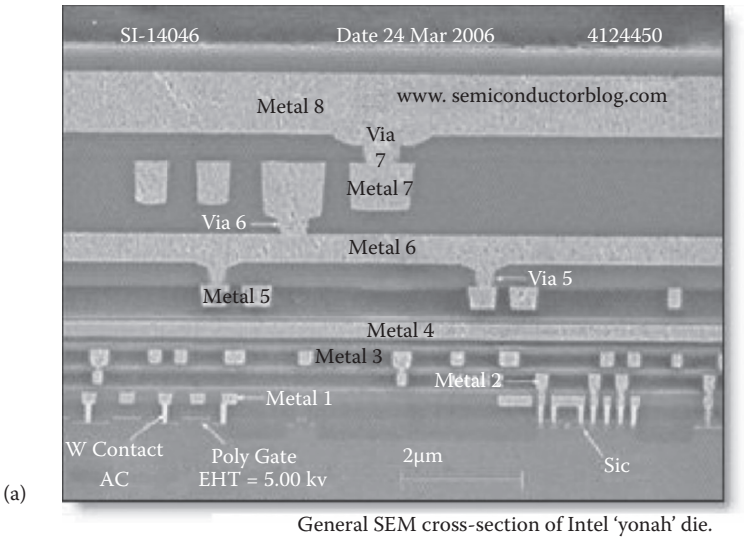
The sensor community is still searching for the perfect way to protect sensors from their environment while at the same time probing that environment. The packaging problem is the least severe for a physical sensor such as an accelerometer, which can be sealed and protected from all chemical environment effects while probing the inertial effects it measures. The problem is most severe for chemical and biological sensors, which must be exposed directly to an unfriendly world. Generalizing, chemical and biological sensors have been orders of magnitude more difficult to commercialize than physical sensors, in no small part because of packaging problems. In the latter case, problems are especially severe with electronic components and sensors integrated on the same side of a single Si substrate. For that reason, chemical and biological sensors are better made with hybrid MEMS. In hybrid MEMS, MEMS and electronics are fabricated in separate processes and put together afterward within a common package. Making a hybrid device or putting the electronics on the other side of the wafer often provides better protection of the electronics from the environment. The challenges in packaging chemical and biological MEMS are further compounded by the fact that these devices usually are less expensive than ICs or physical MEMS sensors and that they are often disposable.

* As we will learn in Chapter 8, this volume, there are two different common definitions of the Q factor of a resonator (see "Quality Factor" in Chapter 8). One is via the energy storage where the Q factor is given as 2π times the ratio of the stored energy to the energy dissipated per oscillation cycle, or equivalently the ratio of the stored energy to the energy dissipated per radian of the oscillation. The other definition is via the resonance bandwidth in which the Q factor is the ratio of the resonance frequency f_0 and the full width at half-maximum (FWHM) bandwidth Δf of the resonance. Both definitions are equivalent only in the limit of weakly damped oscillations, i.e., for high Q values.

Front-End and Back-End Processes

During the front-end processing phase, an IC or MEMS is created on a silicon substrate (referred to as a wafer). During the back-end processing phase, the IC chips created during the front-end processing phase are encapsulated into packages and thoroughly inspected before becoming completed products. The front-end and back-end processes are preceded by a design phase and by the mask creation in which the function of the IC is defined, the electric circuit is designed, and a mask for IC manufacturing is created on the basis of the design.

In this chapter we are primarily concerned with the more expensive, serial back-end processes. The back-end process used to start after the last of the IC fabrication steps, i.e., to make the interconnections within an underlying IC (from level 0 to level 1 in Table 4.1). Unfortunately, things are not that simple anymore; lately making internal IC connections has become so complicated that even within that process one talks about front end and back end of the line steps. As advanced ICs contain more than a 100 million transistors, the importance of internal IC connections is steadily increasing. As shown in Figure 4.3a, eight metal layers already are quite common. The old standard for making these metal lines was using insulating layers made of silicon dioxide, via formation by plasma etching of holes, covering the whole surface with tungsten (W), polishing the tungsten down to the silicon dioxide, leaving the vias filled with tungsten plugs, and finally sputtering an aluminum layer and selective etching it to leave the connecting Al lines. As we saw in Volume II, Chapter 8, the more advanced process for constructing interconnections in ICs uses the "dual damascene" technique (see Figure 4.3b). This process, as we glean



General SEM cross-section of Intel 'yonah' die.

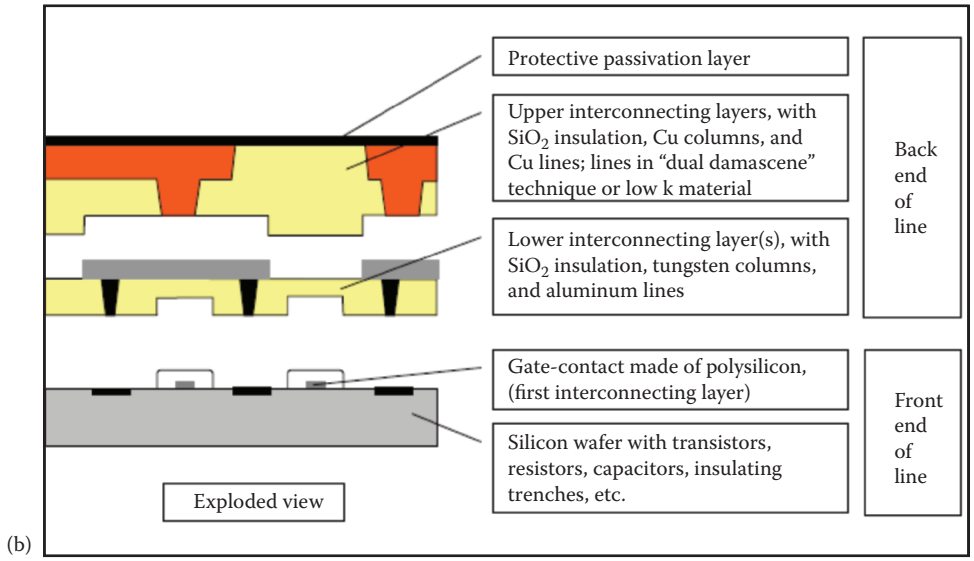


FIGURE 4.3 The back-end process usually takes of at the back end of the integrated circuit (IC) line, which makes the interconnections within an IC (from level L0 to L1). In the case of very complex interconnecting schemes such as the dual damascene, one also talks about back-end and front-end processes with the interconnecting scheme. (a) Eight-metal complementary metal-oxide-silicon chip. (b) Dual damascene technique is used to connect upper interconnecting layers with lower connecting layers (back end of line) and transistors on the wafer (front end of line).

from Figure 4.3b, is itself split up in a back-end and a front end of the line. In the dual damascene process, the last step in the front end of the line is a first interconnecting layer made up of poly-Si for the gate contacts. The back end of the IC interconnection process then starts with an insulating layer, either again made of silicon dioxide or from inorganic or organic materials with lower dielectric constants (so-called *low-k materials*). Next comes the creation of beds for lines and vias in two subsequent photoimaging process steps followed by plasma etching of the insulating layer. In a subsequent process step, the whole surface is covered with copper, and chemical mechanical polishing (CMP) removes the copper down to the top side of the insulating layer (see Figure 4.3b and Volume II, Figure 8.41).

CMP has developed into a very important step in the process chain of wafer fabrication (see Volume II, Chapter 1, Figure 1.66 for a CMP setup). The CMP slurry contains 1) chemistry to modify, e.g., oxidize, the top layer, and 2) abrasive particles to remove the modified material. CMP serves to planarize

insulating layers, to remove tungsten or copper, to define vias and lines, and to back grind the wafer to make thin flexible chips.

The resulting flat chip surface now features input/output (I/O) pads that are of rectangular shape and made from the metal of the top layer, i.e., from aluminum or copper (in GaAs chips, gold is used for contacts). These exposed metal surfaces are slightly alloyed to increase hardness and longtime stability. The chip surface, except the I/O pads and the dicing lines, are covered by a passivation made from silicon oxide and silicon nitride. Historically, this passivation was a crucial step in the development of plastic packages. Plastic packaged chips are not as well protected from humidity as chips in hermetically sealed housings.

Testing the outcome of a wafer process depends on hundreds of parameters that all have to be tightly controlled during production. All materials and parameter settings are recorded to allow the back tracing of deviations. Processed chips still on a wafer are difficult to test, because measurements at high frequencies with hundreds of contacts are difficult and expensive. Therefore, only a so-called *static test* at low frequencies is performed to do a first wafer screening, perhaps in a manual probe station with needle tips and a microscope, as shown in Figure 4.4.

The packaged chip—with its built-in (shielded) fan out of the contacts—is the best location in the fabrication chain to do a full testing of the device (L2; see below). But before we can mount the chip in its package we need to dice it.

FIGURE 4.4 Optical microscope with needle-tip manipulators.

Dicing

One of the final steps in front-end processing of an IC or three-dimensional (3D) microstructure fabrication and the first in back-end packaging is sawing the finished Si wafer into individual "dice." After the wafer is probed (see above; Figure 4.4), it is mounted onto sticky tape and put onto the dicing saw (Figure 4.5). A typical saw blade consists of a thick hub with a thin rim impregnated with diamond or carbide grit. Rotating at a speed of several thousand revolutions per minute, the blade encounters the Si wafer at a feed rate on the order of a centimeter per second, sawing partially or completely through the wafer. After dicing, the elastic plastic sheet can be stretched to distance the chips in case the saw kerf went all the way through the Si wafer.

Except when cutting a silicon-glass bonded wafer combination, standard IC cutting practices regarding surfactants, cleanliness, and blade width/depth ratio apply to MEMS wafers as well. Typical kerf widths are 50–200 µm, while typical roughness along the kerf edge is 10–50 µm. The restrictions of mechanical sawing include this roughness, which precludes the fabrication of smooth structures along the outside edge of the die. In addition, the vibration inherent in the sawing process makes it difficult to securely hold down dice during sawing if these dice are less than 0.5 mm on a side. For smaller dice, micromachining is necessary for edge definition and separation of individual devices. In the latter case, anisotropically etched V-grooves extending almost through the depth of the wafer separate the individual dice, which might be broken apart by applying a small mechanical force. Alternatively, the dice

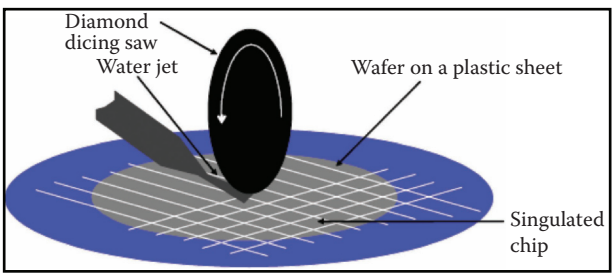


FIGURE 4.5 Dicing operation.

are etched completely free at the moment critical components (e.g., a pressure-sensitive membrane) have reached the specified thickness (see V-groove Thickness Control in Volume II, Figure 4.44).

Most mechanical sensors and actuators are equipped with a bonded cap or cover (Si, glass, etc.), protecting them during dicing. In the case of surface micromachined devices, a final sacrificial release etch (see Volume II, Chapter 7) is performed after the dicing is complete. This procedure ensures that there are no free mechanical structures during the dicing, but it implies also that the microstructures must be freed on individual dice, making for a more costly device. The TI DMD™ arrays described in Volume II, Example 7.2 are handled this way: they are diced first, and then the organic sacrificial layer is etched in an oxygen plasma. The cap approach is not applicable in the case of flow sensors, and chemical and biological sensors or other devices that require direct contact with their surrounding environments.

Level 1 Packaging (L1) and Level 2 Packaging (L2)

Introduction

Level 1 devices are packaged with leadframes (single-chip), transistor outline (TO)-headers, or multichip module interconnection systems, and the resulting devices are then packed onto printed wiring boards (see Table 4.1). Micromachining deviates most dramatically from ICs in the nonstandard nature of packages from level 2 up. In the case of some sensors, standard TO-5 and TO-8 headers still might be useful, but in most cases the package dictated by the specific application is of a much more complex nature.

Single-Device Packaging

A typical package for single chips with 100 and more leads is the quad flat pack (QFP) illustrated in Figure 4.6. The process steps necessary to build such a QFP are the stamping or etching of a leadframe, gluing the chip to the leadframe (also named die bonding), wire bonding the chip to the leadframe, overmolding everything except the outer leads and cutting and bending the leads to gull-wing shape. In the cross-section of the QFP, in Figure 4.6b, the different elements of a QFP are visible: the chip, the leadframe and the glue, the gold wires, the overmolding, and the final form of the leads. This package is now ready to be soldered to a printed wiring board.

Leadframes are stamped or etched strips of metal sheets (Figure 4.7). Chips can be mounted in one or several rows on the large central pads. The outside dimensions of the strips—width, length, and the holes at the edge—are standardized. The basic material is an iron-nickel-alloy. The inner leads, the lands for the gold wire, have to be gold or silver plated. The outer leads are pretinned for the later soldering to the printed wiring board. An auxiliary rectangular line circumventing the main package is needed to seal the overmolding form. This line is stamped out after overmolding.

The die is glued to the large central pad of the lead frame with a conductive epoxy resin. Precise mounting of the chips is required to facilitate the subsequent wire bonding. The glue has to be cured before wire bonding starts. Because standard glues cure in tens of minutes, rather than in seconds, the curing is quite a disturbing factor in a production line. Batches of leadframes have to be placed in a curing oven.

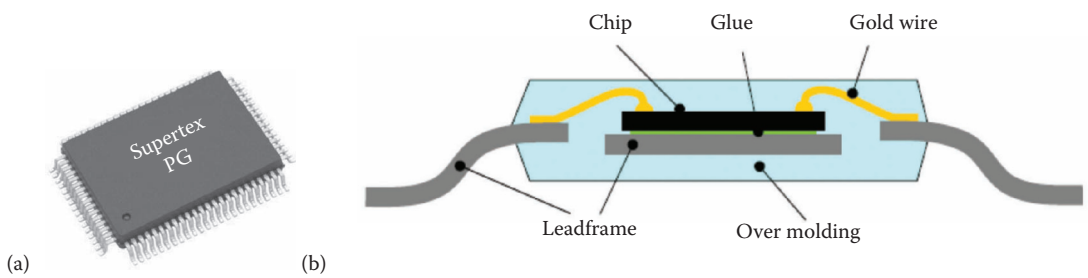


FIGURE 4.6 (a) An eighty-lead plastic quad flat pack (QFP) ready for soldering to a printed wiring board (PWB). (b) A QFP in cross-section.

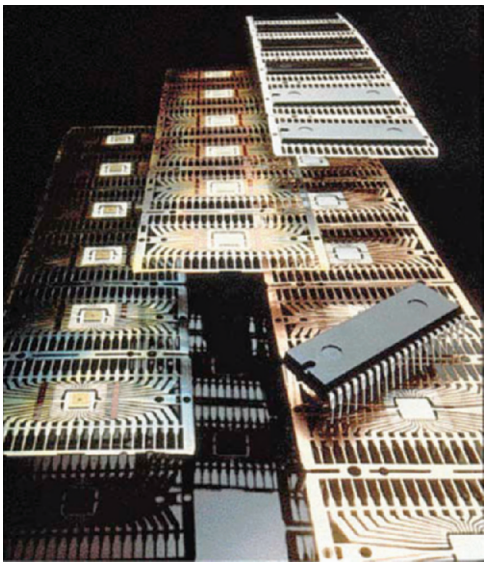


FIGURE 4.7 Leadframes are stamped or etched strips of metal sheets.

The QFP shown in Figure 4.6 is a relatively old packaging scheme introduced in the 1980s. The dual inline package (DIP), what most people envision when they think of an IC "chip," was developed even earlier (in the 1960s) (Figure 4.8a). In a DIP, the leads are perpendicular to the body of the package, making them ideally suited for insertion mounting onto printed wiring boards (PWBs) using plated through-holes.

A QFP is an example of surface mount technology (SMT). In SMT, the leads do not penetrate the PWB surface, and ICs are mounted, by reflow solder technology, on both sides of the PWB, resulting in higher-density electronics. In the industry, SMT has largely replaced the through-hole technology construction method of fitting components with wire leads into holes in the circuit board. An SMT component usually

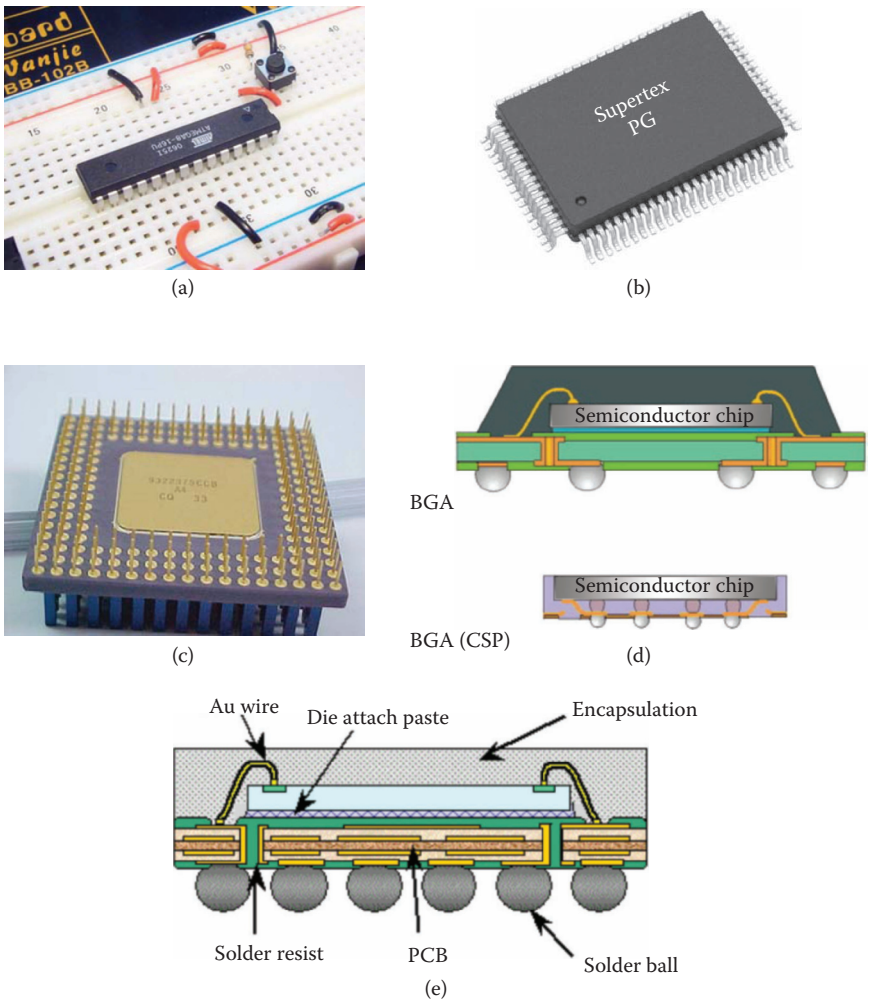


FIGURE 4.8 Packaging options. (a) The dual inline package (DIP), developed in the sixties. (b) The quad flat pack (QFP) was developed in the 1980s. It is a first example of surface mount technology. (c) Pin grid arrays (PGAs) were also developed in the 1980s. (d) Ball grid array (BGA), developed in the early 1990s as a replacement for QFPs, including a BGA implemented as chip scale packaging (CSP). (e) Detail of chip scale package (CSP).

is smaller than its through-hole counterpart because it has either smaller leads or no leads at all. The QFPs (see above and Figure 4.8b) and pin grid arrays (PGAs; Figure 4.8c) were developed in the 1980s. There are leadless or leaded ones and come in plastic or ceramic bodies. With contacts on all four sides, up to ~200 I/Os are possible with QFPs, and PGAs can have up to ~600 I/Os. Ball grid arrays (BGAs) (Figure 4.8d) were developed in the early 1990s as a replacement for QFPs. They feature solder bumps in a matrix on the bottom of the package. They have a higher density, smaller footprint, and shorter electrical paths (i.e., faster signal propagation) but are more expensive than QFPs. BGAs are the natural choice for multichip module (MCM) packaging (see below). Chip scale packaging (CSP) was developed in the late 1990s (see bottom drawing labeled "BGA (CSP)" in Figure 4.8d and Figure 4.8e). CSP is a general term for all semiconductor packaging schemes that aim to be almost the same size as the chip itself (typically about 1.2 times the bare chip). In CSP, the package is an extension of the bare IC and permits easier handling, testing, and assembly. In one version of CSP, the BGA packaging scheme is modified with electrodes that are directly installed on the bottom of the IC. [See bottom drawing labeled "BGA (CSP)" in Figure 4.8d and Figure 4.8e.]

A simple single die mounting method, often used in MEMS, is in a header, as shown in Figure 4.9, where we show the bulk silicon micromachined piezoresistive accelerometer manufactured by TRW NovaSensor (Fremont, CA) in a TO-5 header.

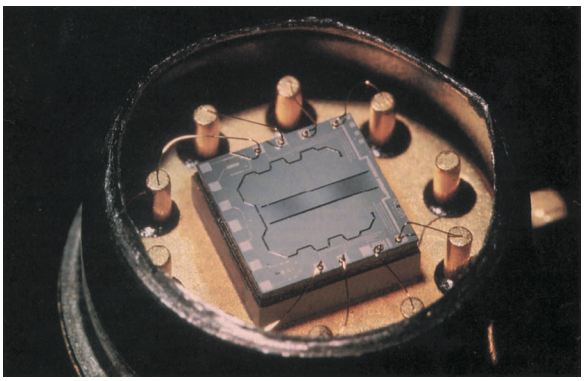


FIGURE 4.9 Commercially available Si accelerometer based on hybrid technology; bulk-micromachined 50-g accelerometer bonded into a TO-5 header. Signal conditioning electronics are on a separate chip. (Courtesy of TRW NovaSensor.)

After dicing a MEMS sensor die, the die attachment step to the header follows. In a typical header, gold-plated pins are hermetically sealed to the header base with glass eyelets, which provide a hermetic feed through with both high electrical and mechanical integrity. To protect the mounted sensor, a metal can is hermetically resistance-welded to the header. Since the thermal expansion coefficient of the package material to which the die is attached typically differs substantially from that of Si and glass, the die attachment must compensate for this. The three methods used the most for die attach are eutectic (Au-Si alloy), epoxy, and silicone rubber, in order of highest to lowest stress.[1,4] The next step in the process consists of the electrical connection to the Si sensor. Gold or aluminum wire is bonded by thermosonic, thermocompression, or wedge-wedge ultrasonic wire bonding (see "Wire Bonding", this chapter). The fine gold wires in the TO-5 package illustrated in Figure 4.9 are barely visible. Wire bonding occasionally damages MEMS components during packaging, as the applied ultrasonic energy, typically at a frequency between 50 and 100 kHz, may start the oscillation of suspended mechanical microstructures. Coincidentally, many micromachined structures have resonant frequencies in the 50–100 kHz range, increasing the risk of structural failure during wire bonding.

The TO-5 packaged accelerometer in Figure 4.9 is an example of a relatively simple packaging solution; typically MEMS packaging is a lot more challenging. Indeed, an accelerometer can be completely shielded from its environment, whereas many other MEMS sensors must be exposed directly to the environment they are sensing.

It is imperative that sensor elements are not subject to undesirable mechanical stresses originating from their packaging structure. Slow creep in the adhesive or epoxy that bonds the silicon die to the package housing, for example, may lead to long-term drift. Corrugated structural members, as used by Jerman,[7] decouple the sensor from its encapsulation, reducing the influence of temperature changes and packaging stress. Corrugated members have been implemented, for example, in single-crystal silicon, polysilicon, and polyimide. The technique was described in Volume II, Chapter 7, as one of the

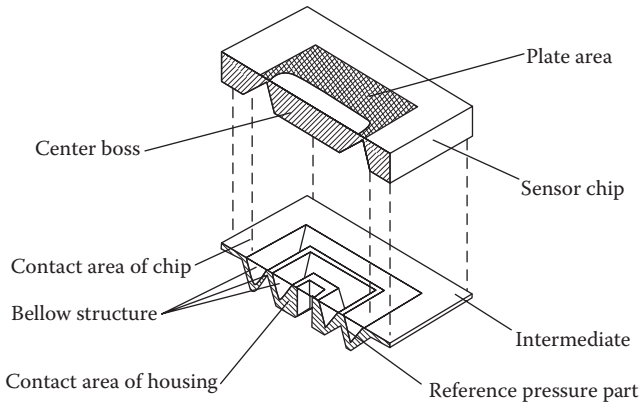


FIGURE 4.10 Stress release bellows. A corrugated intermediate Si piece decouples the sensor from its housing. (From Offereins, H. L., H. Sandmaier, B. Folkmer, U. Steger, and W. Lang. 1991. In *6th International Conference on Solid-State Sensors and Actuators [Transducers '91],* 986–89. San Francisco. With permission.[10])

methods to control stress in micromachined structures. The stress-decoupling, corrugated bellows either surround the sensor structure itself[8,9] or the sensor chip is mounted via an intermediate with a bellow structure to the housing.[10] The latter implementation is sketched in Figure 4.10.

Thermal management in MEMS involves cooling heat-dissipating devices, especially thermal actuators, but also includes understanding and controlling the sources of temperature fluctuations that may adversely affect the performance of the device. Thermal modeling must include both the die and its package. To facilitate heat flow from die to package, a die-attach scheme must be used that exhibits high thermal conductivity.

The last assembly step, after die attachment and electrical contacting, involves the protection of the silicon die and electrical leads. This is especially important for sensors and actuators that are to be used in intimate contact with harsh environments. Various methods are available, including the following:[1,4]

1. Vapor-deposited organics for mildly aggressive environments {e.g., 2–3-μm thick low-pressure chemical vapor-deposited [LPCVD] parylene [poly(*p*-xylene)]}. A negative aspect of this form of passivation is that the added layer often makes the die too stiff for accurate sensing or actuating functions.
2. Silicone oil over the die for medium-isolated pressure sensors.
3. Coating of the die surface with soft substances such as silicone gel. Hardening of the gel over time is a major problem. Silicones do not form a moisture barrier, and the exact mechanism by which they protect the die as a surface coat is not understood.[11]
4. SiC (e.g., deposited with plasma-enhanced chemical vapor deposition) for harsh environments.[12]
5. A plastic or ceramic cap for particle and handling protection.
6. A welded-on nickel cap with pressure port.

Multichip Packaging

A multichip module (MCM) is a chip package that contains two or more bare chips mounted close together on a common substrate. Shorter tracks between the individual chips increase performance and eliminate much of the noise that external tracks between individual chip packages pick up. There are four main categories of MCMs based on the substrate used:

1. Laminated (MCM-L): Constructed of plastic laminate-based dielectrics and copper conductors using PWB technology. These are the lowest cost MCMs.
2. Ceramic (MCM-C): Constructed on cofired ceramic or glass-ceramic substrates using thick-film (screen printing) technologies.
3. Deposited (MCM-D): Uses a dielectric layer over a ceramic, glass, or metal substrate, and thin film interconnects are created on the dielectric layer. These are the highest-performance MCMs.
4. Si-based (MCM-Si): Uses a silicon substrate with tracks created in the silicon like regular ICs. Transistors can also be formed in the substrate.

ICs and micromachined chips may be packed laterally or they may be stacked on top of one another. In the latter case, individual dice, blocks of dice, and entire wafers may be stacked with electrical interconnections either made by wires (see Figure 4.11)

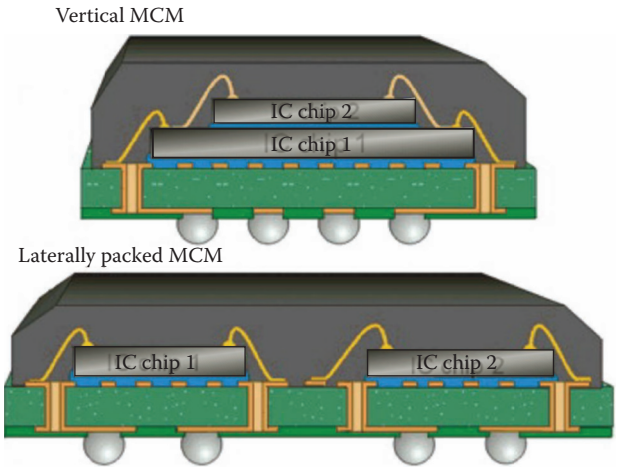


FIGURE 4.11 Schematic of a two-chip MCM.

or interconnects running vertically from plane to plane.[13] A commercial example of a four-chip MCM is shown in Figure 4.12.

MCMs increase the interconnect density on printed circuit boards and speed up the processing time. Multichip modules are an important supporting technology for MEMS, as most MEMS are better partitioned into separate sensing and electronic parts rather than being built monolithically in Si. MCM also allows complementary metal-oxide-silicon (CMOS) and bipolar technologies to be used in the same package.

Instead of the wire bonds shown in the MCMs in Figure 4.11, flip-chip bonding often is used to make contact to the chip. In flip-chip bonding, as its name implies, the die is bonded, facedown, on its substrate without the use of connecting wires (see "Electrical Connection Methods" and Figure 4.13). Electrical contacts are made directly between plated solder bumps on bond pads on the die and metal pads on the packaging substrate. The intimate spacing between the die and the substrate (<200 μm) presents two major advantages. With wire bonding, the pads on the die must always be positioned on the periphery to avoid crossing wires, but with flip-chip bonding, the bond pads may be placed over the entire die. This results in a significant increase in density of possible I/O connections (up to 700 simultaneously I/Os). Second, the effective inductance of each interconnect is minuscule because of the short height of the solder bump compared with that of a gold wire. These are the reasons why the integrated circuit industry has adopted flip chip for high-density, fast electronic circuits. Flip-chip mounting is equally attractive to the MEMS industry. Again, the ability to closely package a number of distinct dice on a single-package substrate with multiple levels of embedded electrical connections represents a major advantage. This approach may be used, for example, for bonding to electrically connect and package three dice, including an accelerometer die, a yaw-rate sensing die, and an electronic application-specific integrated circuit (ASIC) onto one ceramic substrate to build a fully self-contained navigation system. A similar unit could be built with wire bonding, but it would be significantly larger and its reliability would be questionable given the large number of "loose" gold wires within the package.[14] In the next section we explain how a typical flip-chip process works.

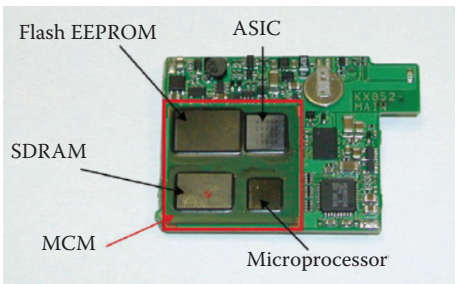


FIGURE 4.12 A 2-mm-thick MCM holding four ICs used in Casio's Exilim digital camera.

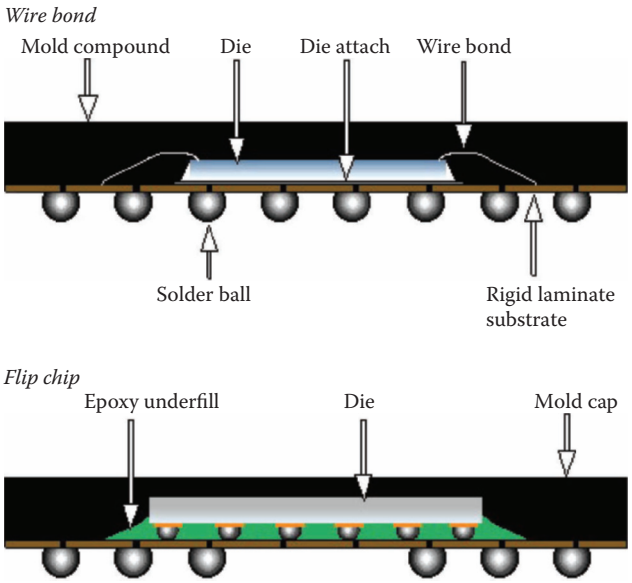


FIGURE 4.13 Flip-chip versus wiring bonding.

Electrical Connection Methods

Introduction

The three most used electrical connection methods between a chip and a substrate are illustrated in Figure 4.14. They are wire bonding, tape automated bonding (TAB), and flip-chip bonding.

Wire Bonding

The pads on a chip and the pads on a substrate (e.g., a leadframe, but also a printed wiring board or ceramic substrate) often are interconnected by wire bonding. A gold, copper, or aluminum wire is bonded between each bonding pad of the chip and the corresponding pad on the substrate. There are two major types of wire bonding: ball-wedge and wedge-wedge bonding. In gold ball-wedge bonding, a gold wire is fed through a ceramic capillary, as shown in Figure 4.15. The protruding wire end is melted by an electrical spark to form a ball, which is bonded to the first pad. This operation requires both ultrasonic excitation of the capillary and heat application to the substrate (200°C). The capillary guides the gold wire now to the second pad, where it forms a wedge and cuts the wire in a single step. The wire has to be clamped to cut it, and ultrasound action and heat application also are necessary to achieve good adhesion of the wedge to the second pad. Copper ball-wedge bonding is also possible when the copper wire is flushed by an inert gas during ball formation. Copper wires can be used instead of gold wires. They fit well to the copper pads of the most recent ICs, and their electrical conductivity is higher. Copper wires also are stiffer than gold wires and exhibit less wire sweep during overmolding. After the mounting and bonding, process steps to establish corrosion protection and mechanical protection of the chip are needed, either by hermetic sealing of the assembled board or locally by "glob top" sealing with a droplet of sealing material, e.g., epoxy. Wire bonding of bare chips is used in hybrids, most often with a ceramic substrate. Chip on board is also used with glass/epoxy laminates in some consumer applications (pocket calculators, watches, smart cards, etc.), and occasionally in professional applications.

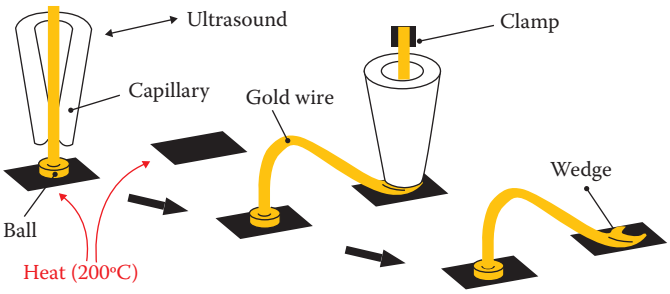


FIGURE 4.15 Gold ball-wedge bonding.

An aluminum ball cannot be formed as easily in air as it can with gold because aluminum oxidizes immediately in air (much faster even than copper). Another technique, wedge-wedge bonding, has to be used as illustrated in Figure 4.16. Aluminum is

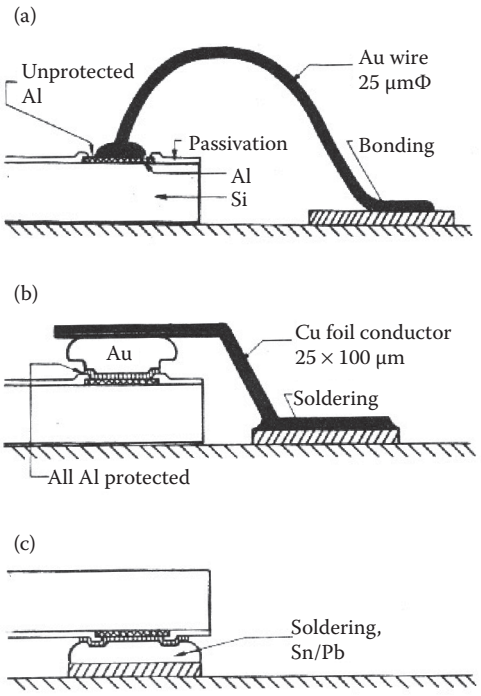


FIGURE 4.14 The three most used electrical connection methods between a chip and a substrate: (a) wire bonding; (b) tape automated bonding (TAB); (c) flip-chip bonding.

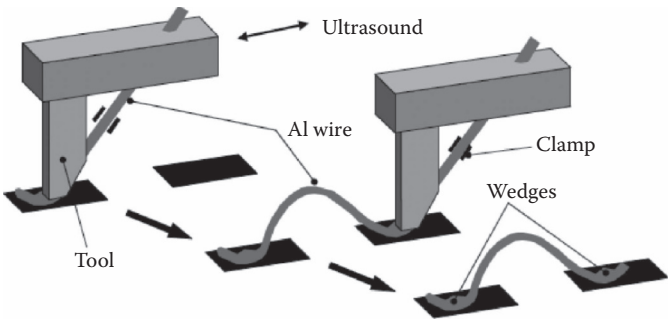


FIGURE 4.16 Wedge-wedge bonding of aluminum wire.

a soft metal, but its oxide at the surface is very hard and has to be broken by the ultrasound action of the wedging tool. Aluminum bonding can be performed with the substrate at room temperature. This is a huge advantage when bonding to a standard PWB substrate with a low glass transition, i.e., softening temperature (T_g) of around 150°C. The wedging tool used for aluminum bonding is directional, and it has to be rotated for different bonds in different directions. It is also larger than the capillary in ball-wedge bonding, so it needs more room on its back side and cannot reach into deep holes between components.

Tape automated bonding (TAB) is illustrated in Figures 4.14b and 4.17. It uses a leadframe made from a fine film of copper conductor on a dielectric film such as polyimide for interconnections between chip and substrate. The conductor patterns in the leadframe are imaged using a photolithography process from a continuous copper layer on the dielectric carrier film. The TAB leads connecting the die and the tape are the inner lead bonds, while those that connect the tape to the package or to external circuits are known as the outer lead bonds (see Figure 4.17). The mounting is done such that the bonding sites on the die, usually in the form of bumps or balls made of gold or solder, are connected to fine conductors on the tape, which provide the means of connecting the die to the package or directly to external circuits (Figure 4.17). To facilitate the connection of the die bumps or balls to their corresponding leads on the TAB circuit, holes are punched on the tape where the die bumps will be positioned. The conductor traces of the tape are then cantilevered over the punched holes to meet the bumps of the die. The leads are soldered to the substrate with a specialized tool called *thermode* ("impulse soldering"). Finally, the bare chip is encapsulated with epoxy or plastic (glob topped). The main advantage of TAB is the tight pitch of the circuit. Line pitches of 45 μm are achievable (22.5-μm lines/spaces). This allows for high-density circuits. TAB is in widespread use in liquid crystal displays, electronic watches, and other high-volume consumer products. Other TAB applications are VLSI circuits with a high lead count and other high-end needs.

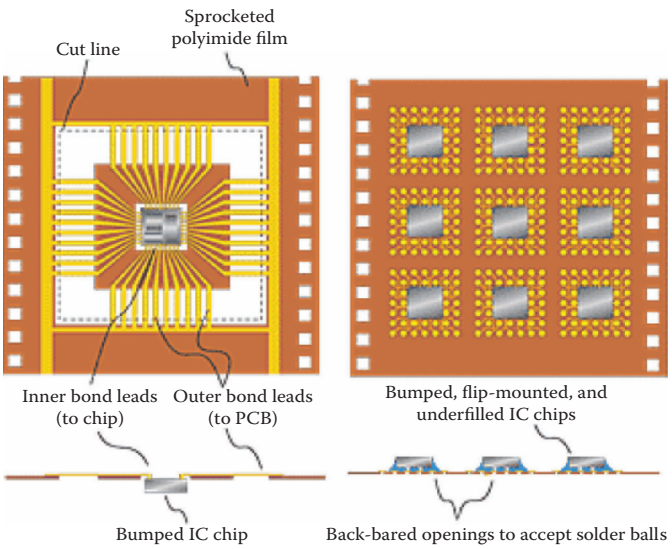


FIGURE 4.17 Tape automated bonding (TAB) takes advantage of flexible base films to make interconnections from the chip to a printed wiring board.

Flip-Chip Bonding

The concept of the flip-chip bonding process, with the semiconductor chip assembled facedown onto a substrate, is ideal for limited size applications because there is no extra area needed for contacting on the sides of the component (see Figure 4.13). Flip chip is carried out by mounting the chip upside down on the substrate, after having deposited solder bumps on all interconnection pads of the chip. The chip is then soldered directly onto the substrate, which requires close thermal matching of the chip and substrate. To make solder bumps on the die, a typical process involves sputter deposition of a thin titanium adhesion layer over the aluminum bond pad metal, followed by sputtering of a copper seed layer. The titanium and copper are patterned with photoresist to form a pedestal for the solder bump. Before the solder bump itself is formed, a thicker layer of copper is first electroplated on the patterned pedestal. The solder bump, typically a tin-lead alloy, is electroplated over the electroplated copper. On the substrate side, solder paste is screen-printed in patterns matching the landing sites of the solder bumps on the die. Pick-and-place robots position the die, facedown, and align the solder bumps to the solder-paste pattern on the substrate. Heating in an oven or under infrared radiation melts the solder, and the bond is created. Often an underfill step is required to fill the void space between the die and

the package substrate with epoxy. The underfill not only adds mechanical strength to the assembly but also protects the connections from environmental hazards. Underfill adhesive may be dispensed along one or more edges of the die and is drawn into the space under the die through capillary action. Heat-curing the underfill adhesive completes the assembly process.

Conductive adhesives are a viable alternative to tin-lead solders for flip chip joining. Conductive or nonconductive adhesives may attach a gold stud bumped chip to a substrate. The conductive adhesives used are either isotropic, conducting in all directions, or they are anisotropic, conducting in a preferred direction only. In Figure 4.18 we show a schematic drawing of flip chip bonding with isotropically conductive adhesives (ICAs) and anisotropically conductive adhesives (ACAs). In either case, the adhesives act as the electrical and mechanical joining agents.

Isotropic conductive adhesives consist of an adhesive binder filled with conductive particles that are normally in contact with each other and provide a low electrical resistance in all directions. Generally, the polymer resin is an epoxy and the conducting particles are silver. The adhesive may be dispensed by stencil printing onto the substrate bond pads, or the bumped die may be dipped into a thin layer of the adhesive, coating only the bumps. Stencil printing of the conductive adhesive limits minimum pad pitch to about 90 µm. Since dipping places adhesive only on the bump surface, the minimum bump spacing is smaller than for stenciling, to pad pitches of 60 µm or less. After the isotropic conductive adhesive is heat cured, a nonconducting underfill adhesive is applied to completely fill the under-chip space.

Anisotropically conductive adhesives are pastes or films that are filled with metal particles or metal coated polymer spheres to a concentration that ensures electrical insulation in all directions before bonding, but after bonding the adhesive becomes electrically conductive in the z-direction. Anisotropically conductive adhesives have the ability to connect finer-pitch devices.

Also, nonconductive adhesives can be used for flip chip bonding, in this case the joint surfaces are forced into intimate contact by the adhesive between the component and substrate.

In general, the pad pitch of the die is finer with adhesive flip-chip than possible with solder flip-chip. The cleaning step is not as rigorous as with solder-based systems; it eliminates process steps such as coating the interconnect pads with solder and eliminates the use of lead. On the downside, the electrical conductivity of ICAs and ACAs is inferior to that of tin-lead solders. ICAs have about 1

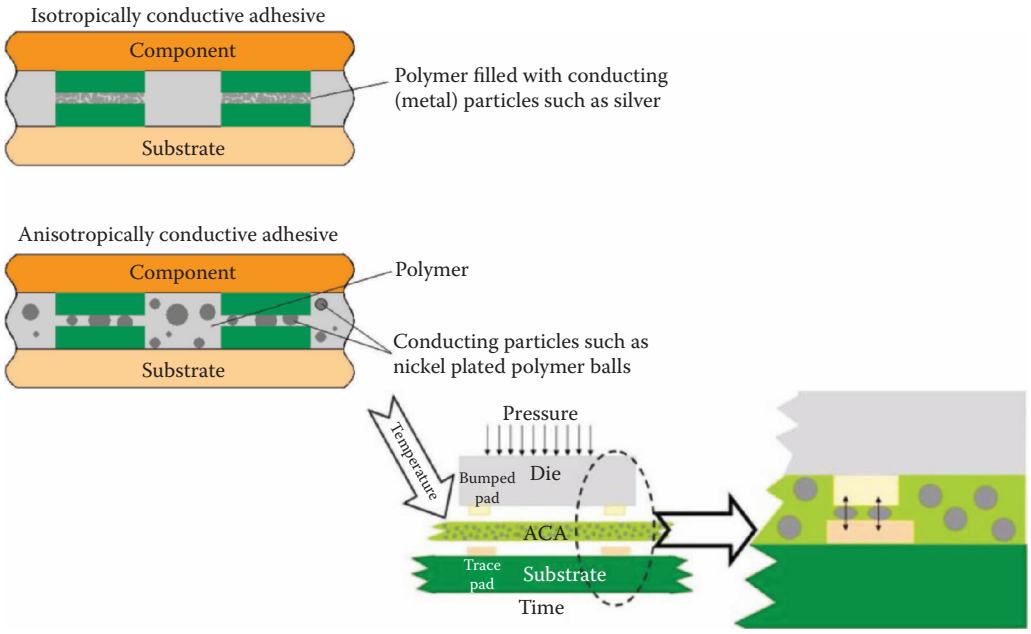


FIGURE 4.18 Schematic drawings of flip-chip joints made with conductive adhesives.

magnitude greater volume resistivity than eutectic tin-lead solder.

Flip-chip technology gives the highest packaging density of the different interconnection methods described here. It gives excellent electrical characteristics and high reliability when properly used, but it is a very demanding technology both to establish and operate. Initial costs are high, favoring high-volume or high-end applications. The performance in high-frequency applications is superior to other interconnection methods, because the length of the connection path is minimized. Also, reliability is better than with packaged components because of the decreased number of connections. Flip-chip technology is potentially cheaper than wire bonding because bonding of all connections takes place simultaneously, whereas with wire bonding one bond is made at a time. In practice, however, this price benefit is not always achieved because of immature processes, e.g., the cost of die bumping with current processes can be significant, especially in low volumes. Today flip-chips are widely used for watches, mobile phones, portable communicators, disc drives, hearing aids, LCD displays, automotive engine controllers, and mainframe computers.

Flip-chip mounting is not compatible with all MEMS devices. For example, there is a risk of damaging the thin diaphragm of a pressure sensor during a flip-chip process.

In what follows we review the different options for through-wafer electrical interconnects. Most methods work for Si only, but others may also be used to form vias in polymers, glass, and ceramics.

Making Vias through Silicon

At first glance, chemical etching is the simplest way to form holes (vias) through Si. The author applied this method for the fabrication of an electrochemical sensor array bonded to a bottom Si chip carrying the electronics (see Volume II, Figures 4.92 to 4.96). Unfortunately, the aspect ratio, length versus diameter, is low, less than 1, so integration is limited by the space between the holes.

With dry etching of vias, problems used to include a lack of sufficiently fast etch rates and sufficient selectivity over the mask. However, advances in high-density plasma sources now allow for etch rates up to 4 µm/min, a selectivity of Si versus SiO_2 of over 150 (70:1 for resists), and aspect ratios of 30:1.[15,16] Along this line, Chow et al. present a through-wafer electrical interconnect fabrication process compatible with standard CMOS processing.[17] In this pre-CMOS process step, a high-density SF_6 plasma (Bosch process; see Volume II, Chapter 3) is used to etch circular holes of 18–25 µm in a 400-µm-thick wafer. By etching from each side, the aspect ratio capability required of the etching process is halved.

Another via alternative is laser drilling. With laser drilling, an aspect ratio of 1:50 can be reached, and relatively high drilling speeds (e.g., 10 holes per second) are possible at a power density of 10^{11} W/cm^2.

Ultrasonic drilling makes for cleaner vias than laser machining, which, unless there is a reactive gas involved or a femtosecond laser, leaves debris on via rims. The ultrasonic technique is limited to "large" diameter holes (100 µm and up). Laser-machined holes are compared with ultrasonically drilled holes in Volume II, Figure 6.21. In all the techniques for making a through-wafer electrical interconnect thus far described, after the hole is created, a metal or highly doped polysilicon must be deposited for electrical connection. In the case of a conductive substrate, the walls of the hole must be passivated first, for example, by growing a thermal oxide. Most methods described to deposit the metal in the vias combine sputtering and electroless or electroplating.

Localized, very deep (throughout the thickness of a wafer) doping of Si is possible by temperature gradient zone melting (TGZM).[18] In this case, via formation and metal deposition become one and the same process. An example is the fabrication of Al interconnects through a Si wafer. At sufficiently high temperatures, aluminum will form an alloy with Si. If the Si substrate is subjected to a temperature gradient, the molten alloy zone will migrate to the hotter side of the wafer. In practice, Al thermomigration in Si starts with electron beam deposition of a thick layer of Al (>5 µm) and photolithographically defining the aluminum into the desired pattern on one side of the Si wafer. The wafer is then radiatively heated from the other side to temperatures considerably higher (1000–1200°C)

than the eutectic temperature to form the molten zone. The one-sided heating imposes a thermal gradient of up to 50°C/cm across the molten zone, resulting in a highly aluminum doped (2×10^{19} cm^{-3}) zone of single-crystal Si. Because of the speed (the zone moves through the silicon on the order of 10 μm/min), wafers can be doped through their thickness in minutes. With TGZM, the concomitant side diffusion is a few micrometers compared with millimeters for solid-state diffusion.

In principle, this technique offers very exciting opportunities for novel 3D structures, as connections can be made front to back on a Si wafer. However, the lines through the silicon, being active junctions, are light sensitive, and the Al on the exit side can turn very rough. Also, because the process is so extreme, it needs to precede the implementation of any other structure on the wafer.

Bonding of Layers in MEMS

Introduction

Wafer bonding is a technology that has found widespread use in IC as well as MEMS fabrication. It allows the permanent bonding of two wafers. Applications include packaging, fabricating 3D structures, and forming multilayer devices. In MEMS, more often than with ICs, one needs to bond a variety of materials together, including Si to a glass pedestal (e.g., in pressure sensors), Si to Si (in many multistack MEMS devices), and polymer to polymer (e.g., in fluidic devices).

Assisted Thermal Bonding or Anodic Bonding

Field-assisted thermal bonding, also known as anodic bonding, electrostatic bonding, or the Mallory process, is commonly used for joining glass to silicon. The main utility of the process stems from the relatively low process temperature (the temperature limit for IC processed Si substrates is about 450°C). Since the glass and Si remain rigid during anodic bonding, it is possible to attach glass to Si surfaces, preserving etched features in either the glass or the silicon. This method is mostly applicable to wafer scale die bonding (L1).

A bond can be established between a sodium-rich glass, say Corning #7740 (Pyrex), and virtually any metal.[19] Besides Pyrex, Corning #7070, soda lime #0080, potash soda lead #0120, and aluminosilicate #1720 are suitable as well.[20] In the case of Si, Pyrex is most commonly used. Bonding can be accomplished on a hot plate in atmosphere or vacuum at temperatures between 180°C and 500°C. Typical voltages, depending on the thickness of the glass and the temperature, range from 200 to 1000 V. The operating temperatures are near the glass-softening point but well below its melting point, as well as below the sintering temperature of standard AlSi metallization. At the most elevated temperatures, the wafers are bonded in 5–10 min, depending on voltage and bonded area.[20] Compared with Si fusion bonding (see below), anodic bonding has the advantage of being a lower-temperature process with a lower residual stress and less stringent requirements for the surface quality of the wafers. Figure 4.19 represents a schematic of an anodic bonding setup and an image of the actual operation. Generally, one places a glass plate on top of the Si wafer and makes a pinpoint contact to the uppermost surface of the glass piece, which is held at a constant negative bias with respect to the electrically grounded silicon. The bonding is easy to follow: looking through the glass, the bonded region moves from the contact cathode pinpoint outward and can be detected visually through the glass by the disappearance of the interference fringes. When the whole area displays a dark gray color, the bonding is complete.

A constant current, instead of constant voltage, could also be used but is avoided, since dielectric breakdown may occur after the bonding is complete and the interface becomes an insulator (see bonding mechanism below). The contacting surfaces need to be flat (surface roughness $R_a < 1$ μm and warp/bow < 5 μm) and dust-free for a good bond to form. The native or thermal oxide layer on the Si must be thinner than 200 nm. The height of metal lines crossing under the bonded area should be less than 600 Å, or 1000 Å for hermetic sealing. The thermal expansion coefficients of the bonded materials must match in the range of bonding. In Figure 4.20, we show the thermal expansion coefficient of Si, a borosilicate

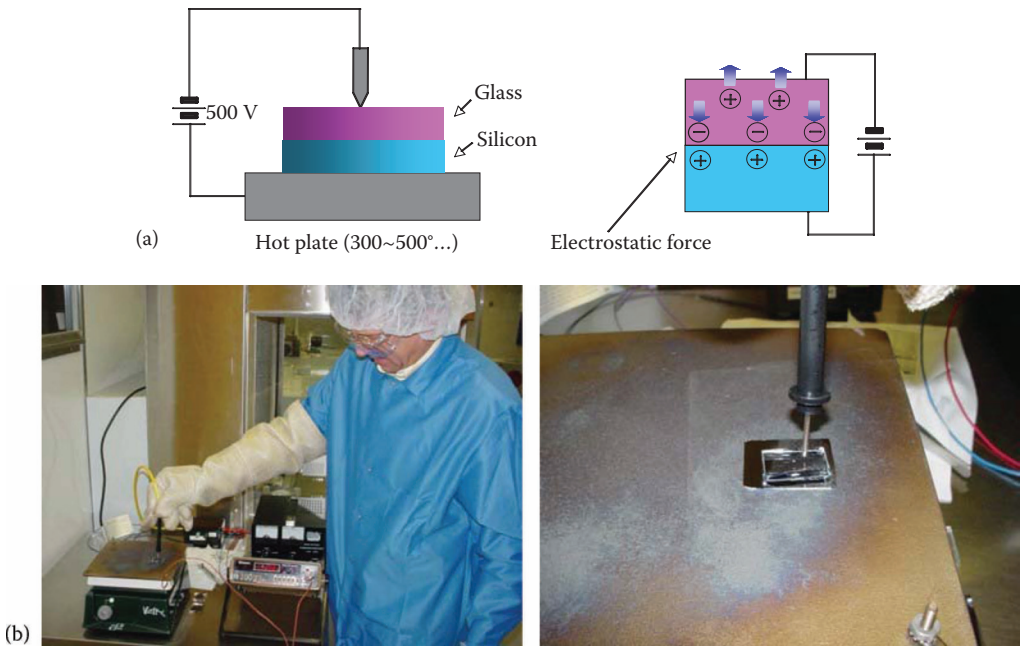


FIGURE 4.19 (a) Principle sketch of anodic glass-to-Si bonding. Control parameters are temperature (300–400°C), bias voltage (700–1200 V), time (~2 min), and materials (glasses, Si, SiO_2). (b) Actual anodic bonding process on a simple hot plate.

glass, and a Pyrex-like glass (HOYA's SD-2) as a function of temperature.[21] The curves of borosilicate glass and silicon single-crystal wafer cross at about 240°C. When anodic bonding is performed at 400°C, the difference of the expansions at high temperature creates residual stress in the Si chip during cooling to room temperature.

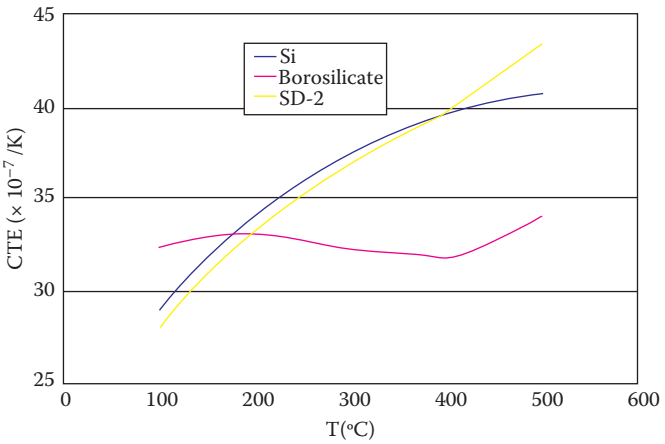


FIGURE 4.20 Thermal expansion coefficients of Si, borosilicate, and a Pyrex-like glass, HOYA's SD-2, a substrate engineered to minimize the distortion or bowing effect caused by the thermal mismatch between the two wafers. CTE, coefficient of thermal expansion. (From Peeters, E. 1994. *Process development for 3D silicon microstructures, with application to mechanical sensor design*. PhD thesis, Catholic University of Louvain, Belgium.[21])

Above 450°C, the thermal properties of the materials begin to deviate seriously; therefore, the process should be limited to 450°C. One also expects that Si would be under compression for seal temperatures below 280°C and under tension for temperatures in excess of 280°C.[21] Wafer curvature measurements indicate, however, that the transition from concave 7440 glass/Si sandwiches (Si under compression) to convex sandwiches (Si under tension) lies around a seal temperature of 315°C.[21,22] This indicates that other non-negligible, stress-inducing effects add an additional compressive component. As we learned before, for most applications tensile stress is preferred over compressive stress, and a considerable safety margin toward higher bonding temperatures must be respected to avoid buckled Si membranes and bridges.

The anodic bonding mechanism is not yet completely understood. Electrochemical, electrostatic, and thermal mechanisms and combinations thereof have been suggested to explain bond formation, but the dominant mechanism has not been clearly defined. It is suggested that at elevated temperatures, the glass becomes a conductive solid electrolyte, and the bonding results through the migration of sodium (Pyrex contains approximately 3–5% sodium) toward the cathode (see Figure 4.19). As

it moves, it leaves a space charge (bound negative charges) in the region of the glass/silicon interface. Most of the applied voltage drop occurs across this space charge region, and the high electrical field between the glass and Si results in an electrostatic force, which pulls the glass and Si into intimate contact. During this charging process, the electric field is high enough to allow a drift of oxygen to the positive electrode (Si) reacting with silicon and creating Si-O bonds. The elevated temperatures and high electric field thus result in covalent bonds forming between the surface atoms of the glass and the silicon. A good quantitative discussion on the many important effects in anodic bonding is by Anthony.[23]

Field-assisted bonding has also been applied to bond GaAs to glass. Corning #0211 is used at 360°C, and a bias of 800 V is applied for 30 min to complete the bonding process. It is well known that GaAs forms very poorly adhering oxides, leading to poor anodic bonding; prebaking of the glass at 400°C for 15 h in a reducing atmosphere (H_2 and N_2) is reported to lead to better bonding.[24] Von Arx et al. bonded glass capsules to a smooth polysilicon surface to form a hermetically sealed cavity large enough to contain hybrid circuitry of a biocompatible implant.[25]

The high electrical field and the migration of sodium make anodic bonding of glass plates to Si a rather difficult technology. The mismatch in thermal expansion coefficient between the glass and the Si causes both thermally induced and built-in mechanical stress. In addition, the viscous behavior of the glass results in degraded long-term stability of the components. As a result of these problems, several modifications of the basic technology have been introduced (see below). A typical commercial instrument for anodic bonding is from Electronic Visions Co.'s 500 Series-Wafer Bonding Systems (e.g., the EV560).

Field-Assisted Thermal Bonding Modifications

The pinpoint method for anodic bonding, as illustrated in Figure 4.19, requires a very high bias voltage and a long period of time to bond areas far removed from the cathode point, since the electrical field in the glass substrate diminishes fast as the distance from the pinpoint cathode increases. At NEC, a Ti mesh bias electrode is deposited over the whole glass wafer to accomplish faster bonding. Because of the mesh assistance, the whole wafer may be Si bonded at 400°C and 600 V in less than 5 min, compared with over an hour at the same temperature and voltage without the mesh.[26]

Another modification of anodic bonding by Sander[27] involves deposition of intermediate layers of Si dioxide and aluminum to screen the underlying Si from harm from the high electrical fields. First, Si dioxide is thermally grown on the Si surface. Then, a layer of aluminum is deposited on the oxide surface. Finally, a piece of glass is bonded to the aluminum. This technique produces a good hermetic seal, but the soft aluminum may be expected to creep after bonding, producing drift in the sensor output. In addition, the aluminum is not corrosion resistant for in vivo applications, so that the bond area may corrode rapidly. A similar modification of glass-to-silicon bonding also uses a sandwich structure, but a layer of polycrystalline silicon (polysilicon) is deposited on the oxide surface instead of aluminum, and a piece of glass is bonded to the polysilicon (see also Von Arx et al.[25]).

It is also possible to anodically bond two Si wafers with a thin intermediate sputtered[28,29] or evaporated[30] borosilicate glass layer (4–7 μm thick). Because two Si wafers are used, there are fewer problems associated with mismatch in thermal expansion coefficient. In addition, since the glass layer is very thin, the voltage necessary for the bond is usually no more than about 50–100 V. Using a thin intermediate film makes the devices akin to a monolithic structure. Residual stresses and thermal expansion coefficient mismatches only have minor effects on device performance. The surfaces of the Si to be bonded should be polished, and one of the two wafers should be coated with the thin glass layer. This intermediate layer may be sputtered from a Pyrex target, resulting in a rather slow process. Hanneborg et al., for example, report a growth rate of 100 nm/h at a pressure of 5 mTorr argon and sputtering power of 1.6 W/cm^2.[29] The films, as deposited, are highly stressed and must be annealed before bonding at the annealing point of 565°C in a wet oxygen atmosphere.[20] With a 4-μm-thick sputtered

glass layer, 50 V and a temperature of 450–550°C suffices to create a hermetic bond. With sputtered layers below 4 µm thick, a thermally grown SiO_2 must be included between the glass and the silicon to avoid electrical breakdown during anodic bonding. The bond strength with Pyrex as an intermediate layer, measured by pulling the wafers apart, ranges from 2 to 3 MPa. As the Pyrex thin films etch very slowly in KOH, they can be used as a mask material for anisotropic etching.[31] Esashi et al.[32] report room temperature anodic bonding using a 0.5–4-µm-thick film sputtered from a Corning glass #7570 target. This glass has a softening point of 440°C compared with 821°C for Pyrex, and bonding is completed within 10 min with 30–60 V applied. A pull test reveals a resulting bond strength in excess of 1.5 MPa. Application of pressure up to 160 kPa on the wafer sandwich apparently supports the bonding process; at a pressure of 160 kPa, the minimum voltage to achieve bonding drops by a factor of 2. The slow deposition rate of glass sputtering presents a major disadvantage for this technique, but evaporated glass layers can be deposited with much higher rates up to 4 µm/min and could provide a solution; unfortunately, in this case pinholes tend to reduce the breakdown voltage.[30]

Field-assisted bonding between two Si wafers, both provided with a thermally grown oxide film, is also reported to be successful. Both Si wafers must be covered with 1 µm of oxide. Bonding of a bare Si wafer to a second wafer with oxide failed because of oxide breakdown under very small applied bias. Temperatures range from 850°C to 950°C, and a voltage of 30 V must be applied for 45 min at the chosen bonding temperature.[33]

Silicon Fusion Bonding

The ability to bond two Si wafers directly without intermediate layers or applying an electric field simplifies the fabrication of many devices. This direct bonding of silicon to silicon, Si fusion bonding (SFB), is based on a chemical reaction between OH groups present at the surface of the native or grown oxides covering the wafers. The bonding relies on the tendency for both very smooth and flat surfaces to adhere, and heat annealing to increase bond strength. The method is of great interest for the fabrication of silicon-on-insulator (SOI) structures. The processes involved in making these bonds are simpler than other currently employed bonding techniques, the yields are higher, the costs are lower, and the mechanical sensors built according to this principle exhibit an improved performance. Finally, as demonstrated in Figure 4.21, the size of an SFB-type device can be almost 50% smaller than that of a conventional anodically bonded chip.

Wafer bonding may be achieved by placing the surfaces of two wafers in close contact and

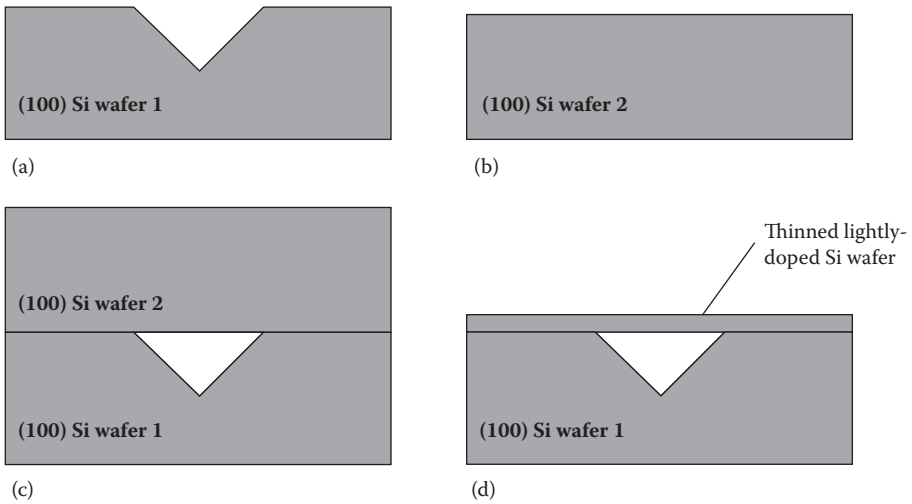


FIGURE 4.21 Silicon fusion bonding enables smaller-footprint microelectromechanical system devices to be fabricated. (a) Etch a silicon wafer to create a square pit (inverted pyramid). (b) A plain second silicon wafer is cleaned and ready for fusion bonding to the first wafer. (c) Bond the two wafer using fusion bonding. (d) The second wafer is then thinned down from the back side using either mechanical-chemical polishing or etching.

inserting them in an oxidizing ambient at temperatures greater than 800°C. It may be noted that the term *silicon fusion bonding* is somewhat misleading; the fusion point of Si at atmospheric pressure is 1410°C, well below the relevant process temperatures.[31] The quality of the bond depends critically on temperature and the roughness of the surface. Flatness requirements are much more stringent than for anodic bondings, with a microroughness R_a < 10 or 40 Å and bow < 5 µm versus 1 µm R_a and bow < 5 µm in anodic bonding. Because of the high temperature involved, active electronics cannot be incorporated before the bonding takes place (the temperature limit for bonding IC-processed standard Si substrates is about 420–450°C; see "IC Compatibility" in Volume II, Chapter 7). Higher temperature (above 1000°C) is usually required to get voidless and high-strength bonding.[34] Bond strength up to 20 MPa has been reported. The application of a small pressure during the bonding process further increases the final bond strength.[35] The bonding can be carried out successfully with one oxidized Si wafer to another bare Si wafer, two oxidized wafers, or two bare Si wafers, and even between one wafer with a thin layer of nitride (100–200 nm) and one bare wafer or two wafers with a thin layer of nitride. The same fusion-bonding technique can be applied for bonding quartz wafers, GaAs to Si, and Si to glass. Provided that the surfaces are mirror smooth and can be hydrated, the bonding process proceeds in an identical fashion as for Si-Si bonding (see Schmidt[36] and references therein).

According to most references, before fusion bonding, the oxidized Si surfaces must undergo hydration. Hydration is usually accomplished by soaking the wafers in a H_2O_2-H_2SO_4 mixture, diluted H_2SO_4, or boiling nitric acid. After this treatment, a hydrophilic top layer consisting of –OH bonds is formed on the oxide surface. An additional treatment in an oxygen plasma greatly enhances the number of –OH groups at the surface.[37] Then, the wafers are rinsed in deionized water and dried. Contacting the mirrored surfaces at room temperature in a clean-air environment forms self-bonding throughout the wafer surface, with considerable bonding forces.[38] In a transmission infrared microscope, a so-called *bonding wave* can be seen to propagate over the whole wafer in a matter of seconds. The bonded pair of wafers can be handled without danger of the sandwich falling apart during transportation. A subsequent high-temperature anneal increases the bond strength by more than an order of magnitude. The self-bonding is the same phenomenon described under stiction of surface micromachined features (see Volume II, Chapter 7).

At present, the fusion bonding mechanism is not completely clear. However, the polymerization of silanol bonds is believed to be the main bonding reaction:

$$\text{Si–OH} + \text{HO–Si} \rightarrow \text{H}_2\text{O} + \text{Si–O–Si} \qquad \text{Reaction 4.1}$$

Figure 4.22 shows the suggested bonding mechanism, including the transformation from silanol bonds to siloxane bonds. Silanol groups give rise to hydrogen bonding, which takes place spontaneously even at room temperature and without applying pressure, as long as the two wafers are extremely clean and smooth.[38]

Measurements of bond strength as a function of anneal temperature indicate three distinct bond strength regions, according to Schmidt.[36] For anneal temperatures below 300°C, the bond strength remains the same as the spontaneous bond strength measured before the anneal. For self-bonded

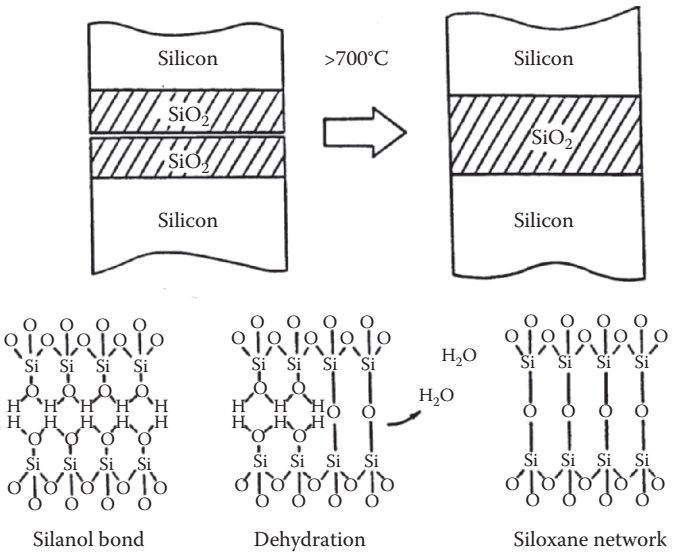


FIGURE 4.22 Schematic illustration of the silicon-to-silicon bonding by fusion and a proposed mechanism.

wafers with hydrogen bonding between the silanol groups of opposite surfaces, Tong et al.[39] measured an activation energy of 0.07 eV for temperatures <110°C—fairly close to the theoretical value of around 0.05 eV for hydrogen bonding. At about 300°C, according to Schmidt, the –OH groups form water molecules, and the voids that are sometimes observed at this temperature are believed to be due to water vapor formed in the process.[40] Stengl et al. report that the formation of voids can be avoided by first contacting the wafers at a temperature of 50°C.[41] The voids tend to disappear also at temperatures above 300°C, while the bond strength increases and then levels out. It is assumed that in this regime, Si-O-Si bonds (siloxane) start forming between the surfaces (Reaction 4.1). Tong et al.[39] found a pronounced increase of interface energy of room-temperature, self-bonded hydrophilic Si/Si, Si/SiO_2, and SiO_2/SiO_2 wafers after storage in air at room temperature and up to 150°C for periods between 10 and 400 h. The interfacial energy increase is ascribed to the generation of additional –OH groups due to a reaction with water and the strained oxide and/or Si below 100°C, as well as the formation of stronger siloxane bonds, which, this group claims, already forms from temperatures below 150°C. They find that the siloxane groups have a much higher bond energy with an activation energy of 1.8–2.1 eV.[39] After prolonged storage, interface bubbles were observed that seem to contain hydrogen and hydrocarbons. At the highest temperatures (>800°C), Schmidt finds the bond strength starts increasing again, and at 1000°C, the bond is reaching the fracture strength of single-crystalline silicon: 10–20 MPa. In this third region, it has been suggested that surfaces can deform more easily (oxide flow), bringing them into better contact.[36] At these temperatures, oxygen also diffuses into the Si lattice. Multiple Si wafers can be fusion bonded together. In some applications, up to a total of seven wafers have been bonded together.

Recall that for fusion bonding, surfaces have to be chemically activated to prepare for the bond. This is quite an undesirable step because the chemicals needed for surface preparation often are not compatible with other micromachining materials. Another way to activate the bonding surface and to make it hydrophilic is to use a plasma activation step. The plasma slightly roughens the surface and increases the number of –OH groups, in a manner similar to hydrophilic chemical activation. The plasma activation is a dry process that is much more compatible with many micromachining process flows. Also, because it is a dry process, it is compatible with released devices (no stiction problems). The bond still needs to be annealed, just as in regular fusion bonding, but the anneal temperatures and times required for a strong bond are much less (400°C vs. 1100°C), possibly because of the slightly roughened surface.

It should be noted that spontaneous bonding of Si wafers also has been claimed with hydrophobic surfaces. It was suggested that such low-temperature bonding is due to van der Waals forces, whereas hydrogen bridging is involved in hydrophilic surfaces. The bonding energy obtained with hydrophobic wafers was as low as 26 mJ/m^2, but after annealing at 600°C, the bonding energy reached a value of 2.5 J/m^2, only attainable at 900°C with hydrophilic wafers.[42] This is in contradiction with most other research done, which concludes that the wafers must be hydrophilic for SFB.

Early on, it was reported that Si_3N_4 was a non-bondable surface.[43] Based on Reaction 4.1, one would assume that the low-temperature direct bonding of Si_3N_4 could take place only if oxidation of the nitride introduced silanol groups. Indeed, an oxidized surface of nitride was reported to bond at high temperatures consistent with Si fusion bonding.[44] Unexpectedly, Bower et al.[45] demonstrated extremely strong bonds between Si wafers coated with a smooth, clean layer of LPCVD Si_3N_4 at temperatures ranging between 90°C and 300°C. A 500-Å-thick nitride was first deposited in the LPCVD reactor and then, without breaking the vacuum, activated in an ammonia stream for several minutes. This group speculates that a chemical reaction such as the following is responsible for bonding where the released hydrogen diffuses from the bonding interface:

$$\equiv Si{-}(N_xH_y) + (N_xH_y){-}Si \equiv \rightarrow \equiv Si{-}(2N_x){-}Si \equiv + yH_2$$

Reaction 4.2

Fracture occurs at about 2 MPa. In a further extension of this work, it was found that an NH_3 plasma

treatment of Si, SiO_2, and TiN for 5 min or longer also provided direct bondable surfaces. Using Si_3N_4 as an intermediate layer, Bower and coworkers bonded Si to Si, Si to SiO_2, and Si to GaAs.[46] This low-activation energy bonding allows bonding of optoelectronic components such as GaAs with completed electronic circuitry on Si with Al metallization in place.

Thermal Bonding with Intermediate Layers

When incorporating an intermediate layer between two substrates, many thermal bonding techniques are feasible. If the intermediate layers can be patterned by lithography, both L0 and L1 packaging becomes possible. Anodic bonding occurs uniformly, but the high fields are detrimental to field-effect devices. Some intermediate layer materials with a low melting point [e.g., phosphosilicate glass (PSG)] do not need electric field for bonding, but the thickness and uniformity are hard to control. An ideal method would require neither an electrical field nor high temperatures and would be very uniform. With films that become soft at low temperatures, all features on the wafer may be covered smoothly as long as they are not a significant fraction of the intermediate film thickness itself.

LPCVD PSG has been used for Si-to-Si wafer bonding. Fusion of two Si wafers coated with 1–2-µm-thick PSG layers is fast. Excellent bonding results, provided the wafers are clean and reasonably flat. Unfortunately, the bonding process occurs at a high temperature of 1100°C for 30 min.[47]

Several suppliers of glass materials offer low-temperature sealing glasses known as glass frits or glass solders. Corning Glass Works, for example, has introduced a series of glass frits (#75xx) with sealing temperatures ranging from 415°C to 650°C.[48] These glasses are applied by a variety of methods, including spraying, screen-printing, extrusion, and sedimentation. After the glass is deposited, it needs to be preglazed to remove the organic residues produced by vehicle and binder decomposition (see "Types of Inks/Pastes and Applications" in Volume II, Chapter 9). The substrate-glass-substrate sandwich is then heated to the sealing temperature while a slight pressure is applied (>1 psi). Devitrifying glasses and thermoplastic, vitreous glasses are offered. The devitrifying glasses are thermosetting, and they crystallize during the sealing procedure, thereby changing their mechanical properties. The vitreous glasses are thermoplastic, melt at a lower temperature than devitrifying glasses, and flow during sealing, but their material properties do not change after sealing is completed. An important problem today remains the uncertainty regarding the mechanical and chemical behavior of these glasses.[27] Moreover, almost all of the low-melting-point formulations of glass frits contain some lead, which is a major drawback because the European Union and Japan are forbidding the use of lead in all electronics manufacturing in the coming years.

Ko et al.[20] used RF sputtering from a target made from a Corning 7593 glass frit to obtain a thinner (8000 Å) and more uniform intermediate glass film for bonding. The sputtered glass did not need glazing. Annealing at 650°C in an oxygen atmosphere resulted in an excellent bond strength of two Si wafers coated with these films.

Legtenberg et al.[49] used a thin film of atmospheric pressure chemical vapor-deposited (APCVD) boron oxide with a softening temperature of 450°C. However, boron oxide is hygroscopic and does not present a viable solution. Field et al.[50] used boron-doped Si dioxide, which also becomes soft at 450°C. The doping is performed in a solid source drive-in furnace. The bond seal is hermetic, and a crack propagates right through the material rather than along the bonded surface. A problem with this method is the sensitivity of the film to phosphorous contamination.

Spin-on-glass (SOG) was reported by Yamada et al.[51] and Quenzer and Benecke[52] In Yamada et al.'s effort, SOG [$Si(OH)_x$ (with $2 < x < 4$)] was coated on the wafer surface to be bonded. After the 50-nm-thick film was baked at a temperature of 250°C for 10 min, the wafers were contacted and pressurized in vacuum at 250°C for 1 h. The bonding was already strong after this step, but wafers were further annealed for 1 h at 1150°C in air to sinter the SOG layer and to improve the breakdown voltage of the layer. Quenzer et al. used sodium silicate as an intermediate layer. The wafers were made hydrophilic in an RCA1 cleaning step, and the glass was spun on from a dilute sodium silicate solution. The film thickness was between 3 and 100 nm, and

annealing at 200°C produced good bonds between bare Si wafers, Si wafers with native oxide, thermal oxide, and Si nitride. No need for further annealing is mentioned in this work.

To bond the unpolished back side of a Si die to another Si part, an aqua-gel based on a hydrophilic pyrogenic silica powder (grain diameter smaller than 7 nm) and polyvinyl alcohol (PVA) as a binder may be used. In this scheme, the adhesion depends on hydrogen bonding between the −OH groups of the PVA, the SiO_2 grains, and the Si surfaces. No special surface treatment is required, and the bonding can be achieved in 15 min at room temperature. Bond strength is in excess of 1 MPa. This new technique works for bonding Si to Si, Si to SiO_2, SiO_2 to SiO_2, and glass to Si or SiO_2. Surprisingly, this technique was also shown to work for bonding III-V materials. This type of bonding is proposed for the fashioning of multichip-on-silicon packaging (L2 level).[53]

Silicon microstructures can be sealed together by eutectic bonding. Eutectic films are alloys of two materials and act as the "glue" or solder to bond two wafers. Solder is a form of eutectic alloy that is commonly used in building printed circuit boards. Solder has been used since antiquity to bond metal surfaces to one another. It is an alloy with a low melting point, typically containing tin or indium, used to join electrical components to a circuit board or to join metal objects together. In Au-Si eutectic bonding, one applies a contact force and increases the temperature beyond the Au-Si eutectic temperature (363°C). This results in diffusion of silicon into the gold (not diffusion of gold into silicon) until the Au-Si eutectic composition is reached (19 atomic percent Si). The Au-Si eutectic bonding takes place at a temperature of 363°C, which is well below the critical temperature for Al metallized components, and Au is a quite commonly used thin-film material. These are good reasons to select the Au-Si eutectic, but other material combinations are possible. It is possible to bond bare Si against Au-covered Si, or Au-covered Si against Au-covered Si. Another alternative is to use Au-Si preform with a composition close to the eutectic concentration. Eutectic bonding in MEMS packaging is demonstrated below, in Figure 4.37, where we illustrate a wafer-to-wafer transfer of an encapulsation package. Tiensuu et al. have demonstrated a mean fracture strength of the Au-Si bond of 148 MPa.[54] This compares well with fusion bonding with a typical bond strength of 5–15 MPa.

There are some considerable disadvantages associated with Au-Si eutectic bonding. It is difficult to obtain complete bonding over large areas, and even native oxides prevent the bonding to take place and eutectic preform (e.g., with Au-Sn) bonding is reported to introduce substantial mounting stress in piezoresistive sensors, causing long-term drift due to relaxation of the built-in stress.[20]

Bonding with Organic Layers

Two silicon wafers also can be bonded using an intermediate polymer film that forms the "glue" between the two. Polymers such as polyimide, or "epoxy" can be used; if the glue can be deposited in a uniform film and has good adhesion to Si, it can be used to bond the two wafers together. One of the issues with polymers is that as they are baked to harden them, they produce vapors of solvents within them, and these vapors create voids, which tend to push the two wafers apart. Therefore, polymers that tend to outgas very little are desired. One of these materials is benzo-cyclobutene (BCB), which is used as an insulator in building integrated circuits. BCB can be spin deposited and then patterned like photoresist. Thus, it is possible to create fine features on one wafer that can then be used to bond to a companion wafer. BCB patterns retain their shape after bonding, and the bond strength is very high (one needs to destroy the BCB to break the bonded wafers). It is also possible to use another polymer, such as Parylene, which is simply a plastic. When heated, the plastic joins the two wafers together. Parylene is a very inert, tough, and easy to deposit polymer. It is used extensively in implantable biomedical systems. It can be evaporated at room temperature (0.1–10 μm range) and when deposited it is conformal, uniform, and pinhole-free.

A most interesting option for packaging is through lithographic patterning of thick resist layers. Prominent new candidates for packaging micromachines are thick ultraviolet photoresists, such as AZ-4000 and SU-8, reviewed in Volume II, Chapter 1, as well as LIGA resists such as PMMA.

Using polymers, very low bonding temperatures are possible. The bond strength can be quite high, no metal ions are present, and the elastic properties can reduce stress. One can envision that with thick enough polymer layers, packaging as part of the front-end processing will become possible (up to level 2). Disadvantages include impossibility of hermetic seals (see Figure 4.29), high vapor pressure, and in cases where a high Young's modulus and low hysteresis is required poor mechanical properties.

An example of photopatterned bonding is provided by the flexible polysiloxane interconnection between two substrates demonstrated by Arquint et al.[55] UV-sensitive cross-linking of polysiloxane layers with a thickness of several hundred micrometers is used to pattern the polymer onto the first substrate, and a condensation reaction is used to form a chemical bond to the second surface. The procedure is illustrated in Figure 4.23.[55]

Good bonding results with photopatternable resists also have been obtained by den Besten et al.,[56]

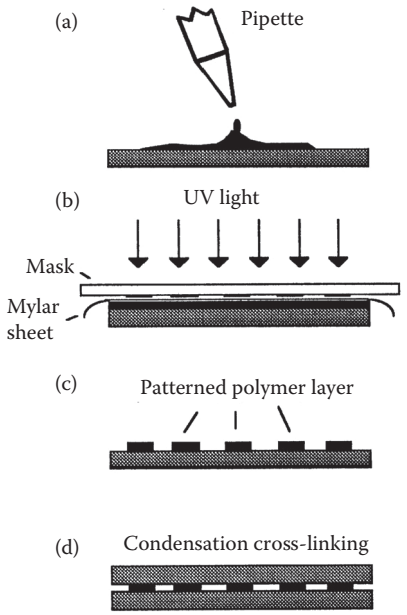


FIGURE 4.23 Procedure for silicon/polymer/silicon bonding: (a) Monomer solution is deposited either by spin coating or casting; (b) exposure to ultraviolet light for photopolymerization; (c) polymeric pattern is developed in xylene; (d) second wafer is pressed onto the polymeric pattern and left for humidity-induced polymerization in ambient air (10 h). (From Arquint, P., P. D. van der Wal, B. H. van der Schoot, and N. F. de Rooij. 1995. In *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*, 263–64. Stockholm, Sweden. With permission.[55])

who used negative photoresist at a bonding temperature of 130°C. In an important potential application, UV-curable encapsulant resins have been employed for wafer-level packaging of ion-sensitive field effect transistors (ISFETs).[57]

Bonding of Plastic to Plastic

A major concern of polymer-based MEMS, especially in microfluidic devices, is how to bond plastic parts together. This bonding or application of plastic lids to plastic microparts can be achieved using adhesives or tapes, plastic welding (e.g., hot plate and ultrasonic welding), and selected organic solvents (i.e., partially dissolving the bonding surfaces and evaporating the solvent). In our laboratory, several of these techniques have been tried, including vacuum-assisted heat bonding, bonding with double-sided adhesive tape, and water-soluble polymer-assisted bonding. Figure 4.24 shows a cross-section of microchannels bonded by these three methods. For large features (several hundred micrometers to millimeters), conventional plastic welding techniques such as hot plate, ultrasonic, and infrared (IR) welding can be used. Applying vacuum before welding may reduce voids at the welded interface, and consequently, lower pressure and temperature can be used in welding. This minimizes the molten polymer from flowing into the bonded channels (see Figure 4.24a). Ideally, small overflow traps should be designed along the edge of the channels (as shown in Figure 4.25) to completely prevent the channels from being blocked by the molten polymer during welding. For smaller channel sizes (less than 100 μm), the design and fabrication of overflow traps become difficult. Double-sided adhesive tape can be a useful way for microbonding, as shown in Figure 4.24b. The main disadvantage is that the presence of an adhesive tape tends to change the channel size and channel wall properties. This is the technique we often used in the assembly of the compact disc (CD) fluidic platform (see Chapter 5, Figure 5.28). For very small channels (less than 1 μm), prefilling the channels with a water-soluble polymer (serving as a sacrificial material), drying the polymer, sealing the plate with channels to a cover lid by either welding or adhesive bonding, and then washing out the water-soluble polymer turns out to be a viable way to

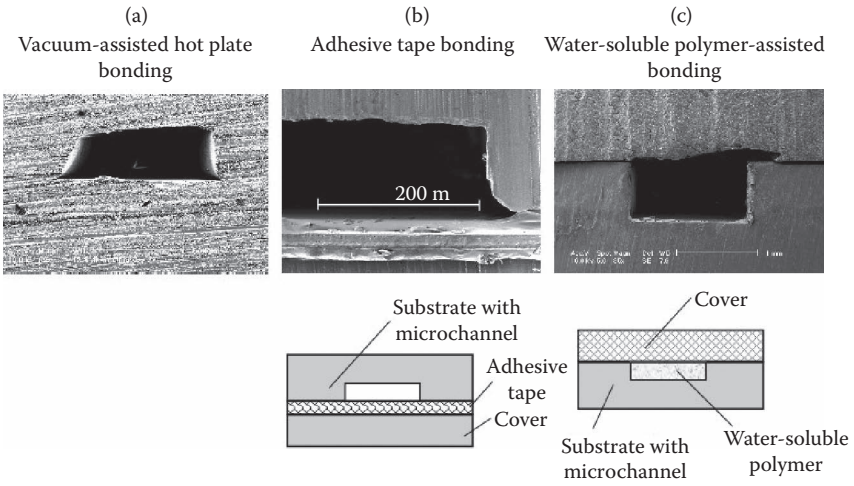


FIGURE 4.24 Scanning electron microscope cross-section of bonded microchannels by three different methods (see text). (Courtesy of Dr. Marc Madou and Dr. James Lee, Ohio State University.)

prevent channel blockage during bonding (see Figure 4.24c). This method, however, is more tedious and may not always be applicable. Development of low-cost, high-speed, and reliable bonding techniques for microfluidic devices remains an important and challenging issue in bio-MEMS fabrication.

Polydimethylsiloxane (PDMS), introduced in Volume II, Chapter 2, is the most common example of a group of polymers known as silicones. Silicones are used as a component in silicone grease and other silicone-based lubricants, as well as in defoaming agents, damping fluids, heat transfer fluids, cosmetics, hair conditioner, and other applications; they are also the primary component in Silly Putty. PDMS is used extensively in microfluidics research to create channels, valves, pumps, and other components that make up microfluidic systems. As we saw in Volume II, Chapter 2, the molecular structure of PDMS is made up of a backbone of siloxane with two methyl groups attached to each silicon atom. The backbone is terminated with either a methyl group or hydroxyl group. PDMS can be reversibly and irreversibly bonded to a variety of substrates. Without pretreatment, PDMS bonds reversibly to glass structures with a bond strength that can survive fluidic pressures of ~4 psi, enough for many microfluidics applications. Activating the PDMS surfaces with a plasma treatment before bonding can create an irreversible bond. The plasma treatment temporarily replaces one of the methyl groups of the PDMS with a silanol group. This activation typically lasts for tens of minutes. If the surface of an activated PDMS layer comes in contact with another activated surface, the silanol groups combine to create a siloxane bond and a water molecule, similar to the hydrophilic fusion bonding process described above. There is no need to anneal this bond because the high permeability of the surrounding polymer allows the water molecule to diffuse away easily (although heat speeds the process up). Irreversible PDMS bonds can stand pressures of up to 70 psi.

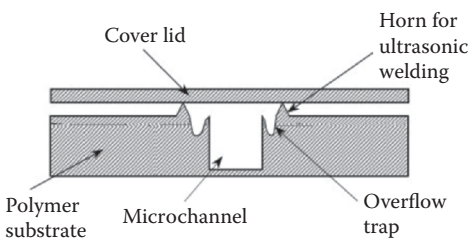


FIGURE 4.25 Schematic of ultrasonic welding with overflow trap. (Courtesy of Dr. Marc Madou and Dr. James Lee, Ohio State University.)

The bonding processes detailed here are summarized in Table 4.2. A good review of polymer micromolding, including polymer bonding, is that of Becker and Gärtner.[58]

Alignment during Bonding

Obtaining good alignment between the device wafer and support substrate (e.g., glass or another Si wafer) during bonding poses some extra challenges, because a transfer from aligner to annealing furnace

Table 4.2 Comparison of Bonding Techniques

	Anodic	Fusion	Glass Frit	Eutectic	Solder	Parylene	PDMS
Bond strength	Very strong	Excellent	Strong	Strong	Strong	OK	Weak
Hermeticity	Good	Excellent	Good	Good	Good	Poor/OK	Poor
Temperature							
Formation	250–400°C	200–1000°C	400–500°C	>363°C	57–400°C	230°C	RT–90°C
Service	>400°C	>1000°C	<400°C	<363°C	<<formation	<<230°C	<90°C
TCE mismatch	OK	Good	OK	OK	Poor	OK	OK
CMOS-compatible	OK	Poor	OK	OK	Good	Good	Good
Process complexity	Low	Some	Some	Some	Some	Some	Some
Planarization capability	Poor	Very poor	Good	Good	Good	Good	Good
Maturity	Decades	Several years	Several years	Several years	Years	Newer	Years
Cost	$	$$$	$$	$$$	$$	$$	$

generally is required after alignment. One technique used for bonding alignment is to generate holes in both the device wafer and the substrate, followed by putting them in a specially designed fixture to perform the bonding. The accuracy achievable with this option is only around 50 µm, though. Higher alignment accuracy (~2.5 µm) can be achieved by using a bonding machine equipped with an in situ optical alignment setup. Aligned wafer bonding has been worked on extensively by Bower et al.[59] They used an infrared aligner modified to hold two imprinted wafers face to face while projecting an infrared image of the surfaces to a viewing screen (see Figure 4.26).[59] An array of V-grooves etched into the surface of the Si wafers was then precisely aligned, and the wafers were brought in contact for initial bonding. The

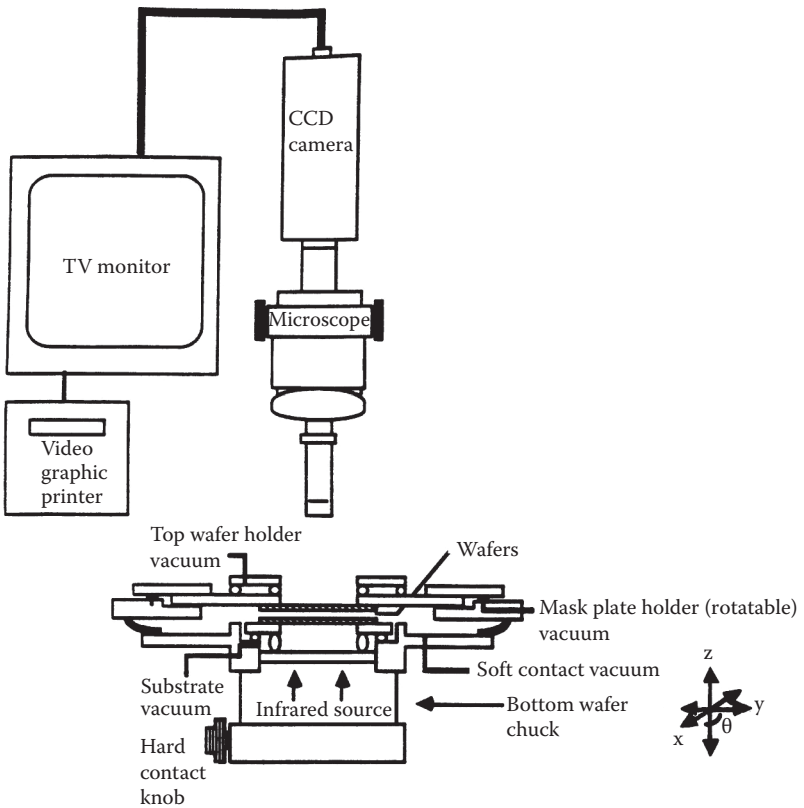


FIGURE 4.26 Alignment fixture for bonding. The schematic diagram shows the infrared aligner system used for wafer bonding. Wafer holders are designed with a large opening at the center to allow maximum infrared energy to pass through to the camera. (Courtesy of Dr. R. Bower, University of California, Davis.)

(111) planes that define the walls of the V-groove create a shadow image without the necessity of metal features commonly used for two-sided backside alignment. The initial bonding is nothing more than the normal soft contact provided by the aligner itself. Hydrogen bonds are formed between the two surfaces in intimate contact. These relatively weak bonds of about 0.05 eV are sufficient to hold the two pieces firmly together without relative displacement during the transport from the aligner to the annealing furnace. Subsequent high-temperature annealing (950°C for 30 min) was used to strengthen and complete the chemical bonding. The aligned-wafer technique also can be used when dealing with dissimilar substrates.

Shoaf and Feinerman[60] developed an alignment technique to assist in precise Au-Si eutectic bonding of Si structures. A (100)Si wafer is anisotropically etched to create V-grooves around the periphery of the structure to be bonded. Gold is deposited onto one of the wafers prior to dicing into individual dice. Optical fibers are placed into the orthogonal V-grooves, as demonstrated in Figure 4.27, and used as precision location keys for assembly prior to bonding. The entire structure is placed on a hot chuck at 400°C for bonding, and the fibers are removed after bonding. Results have shown a maximum misalignment of 5 μm for a 1 × 1 cm die. This technique allows a sensor die to be precisely bonded to an electronics die without the aid of a microscope or micropositioners.

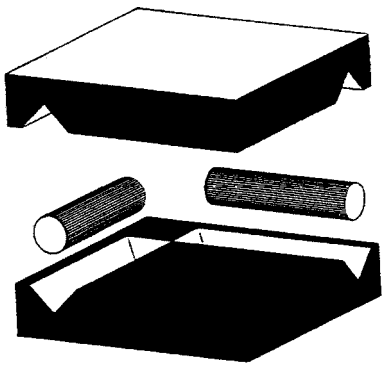


FIGURE 4.27 Schematic representation of the anisotropically etched V-groove/optical fiber alignment technique. Optical fibers are placed into the V-grooves of the bottom Si die. A second die with etched V-grooves is then placed on top of the first die. The fibers act as precision locating keys and align the two Si dice. (From Shoaf, S. E., and A. D. Feinerman. 1994. Aligned Au-Si eutectic bonding of silicon structures. *J Vac Sci Technol* A12:19–22.[60])

Imaging and Bond Strength and Package Hermeticity Tests

Methods for imaging a bonded pair of Si wafers include infrared transmission (voids larger than 20–30 μm can be seen this way), ultrasound (qualitative information about the bond quality), and x-ray topography. The most common mechanical techniques to characterize bond strength are illustrated in Figure 4.28.[36] Both the burst test (Figure 4.28a) and the tensile/shear test (Figure 4.28b) yield important engineering insights for sensor construction but do not yield information about the detailed nature of the bond because of the complicated loading of the interface.[36] In the Maszara method,[61] a thin blade is inserted between the bonded wafers and a crack is introduced (Figure 4.28c). The length of the crack, measured by infrared imaging, gives a value for the surface energy inferred through knowledge of the sample and blade thickness and the elastic properties of the wafer. This method has the advantage of creating a well-defined loading on the bonded interface. Unfortunately, the surface energy is a fourth

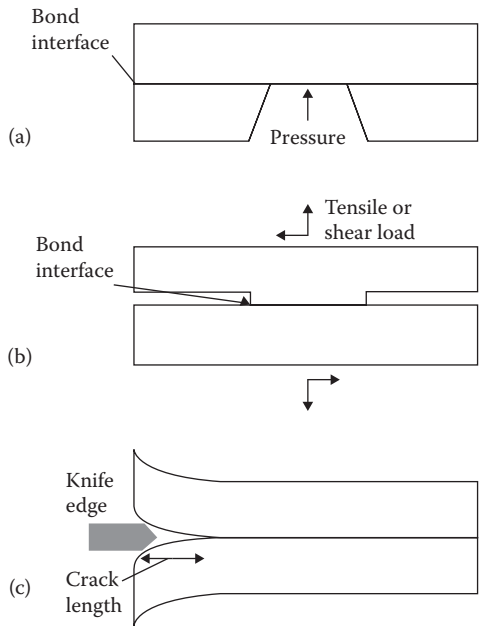


FIGURE 4.28 Three methods for bond strength measurements. (a) Burst test. (b) Tensile and shear test. (c) Maszera test. (From Schmidt, M. A. 1994. In *Technical Digest: 1994 Solid State Sensor and Actuator Workshop,* 127–30. Hilton Head Island, SC. With permission.[36])

power of the crack length and uncertainties in measuring that length make for large errors.

The hermeticity of sealed cavities in packaging is important for physical protection purposes and, in some cases, for the performance of the sensor inside. The word *hermetic* means airtight, so a good hermetic seal does not let gasses, vapors, or liquids pass across it. Some materials are better than others at creating hermetic seals because their molecules (or atoms) are more closely spaced together. Hermeticity is an important parameter for micropackaging because certain MEMS devices need vacuum to operate and ICs need a dry environment to maintain their reliability. For example, the quality factor, Q, of a resonator, the vacuum reference of an absolute pressure sensor, and the cavity of a pneumatic infrared sensor (i.e., a Golay cell) all are critically dependent on a good hermetic package. Testing the hermeticity of electronic devices usually is carried out by helium leak detection. While this technique, with a minimum detectable leak rate of 5×10^{-11} to 5×10^{-10} Torr L/s, is appropriate for testing of relatively large packages, it is unacceptable for nearly every Si sensor application (it is not valid for package sizes under 1000 nL). Nese et al. have introduced Fourier transform infrared (FTIR) spectroscopy for measuring the gas concentration inside a sealed Si cavity using N_2O as a tracer gas.[62] Although the sensitivity for leakage is about the same as with a helium leak detector, the method is accumulative, so by prolonging the exposure times, the sensitivity can be increased, which is not possible with the dynamic helium leakage testing method.

Although anodic bonding may be carried out in a vacuum chamber, the pressure inside a Si cavity differs from the chamber pressure because of gas generation, probably oxygen, during bonding. To control cavity pressure for critical damping of packaged micromechanical devices, Minami et al.[63] used nonevaporable getters. Getters are agents that counteract harmful contaminants within a sealed package; they may include solids, liquids, gases, and combinations thereof. The solid nonevaporable getter used by Minami et al.[62] is a Ni/Cr ribbon covered with a mixture of porous Ti and Zr-V-Fe alloy that absorbs gases after activation at 400°C and is built into the microdevice. These authors monitored the effectiveness of their method to control cavity pressure by measuring deflection of a thin membrane covering the cavity.

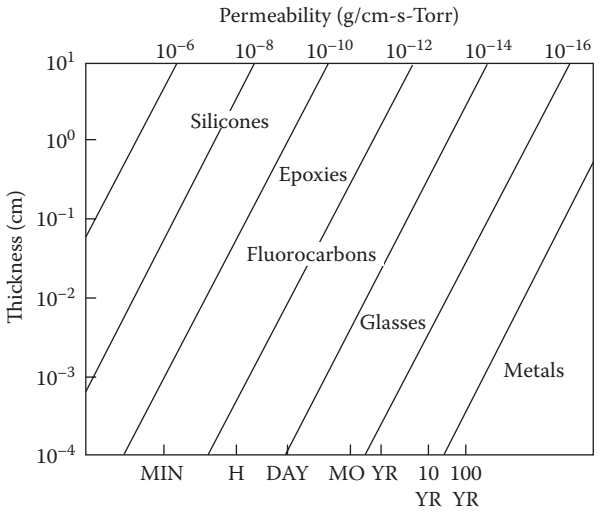


FIGURE 4.29 The calculated time for moisture to permeate various sealant materials (to 50% of the exterior humidity) in one defined geometry. Organics are orders of magnitude more permeable than materials typically used for hermetic seals. (From Striny, K. M. 1988. In *VLSI Technology,* ed. S. M. Sze, 566–611. New York: McGraw-Hill. With permission.[64])

Figure 4.29 shows the relative capabilities of several materials to exclude moisture from the encapsulated components over long periods. Organic materials are not good candidates for hermetic packages; for almost all high-reliability applications, the hermetic seal is made of glass or metal. Plastic packaging for integrated circuits (ICs) is governed by several industry standards that have evolved over the years; plastic packaging for MEMS, on the other hand, is not yet governed by any standard. For the simplest mechanical sensors, standard or slightly modified IC plastic packages are used. New plastic packaging technologies and their standardization for MEMS will begin only once the volume of MEMS products approaches a larger fraction of that of ICs.

For measuring moisture penetration in a package, temperature-accelerated soak tests may be performed. Moisture penetration can be followed, for example, with an integrated on-chip dew-point sensor.[25] Striny[64] points out that the cost or difficulty of obtaining a truly hermetic seal for a particular package often is prohibitive. In that case, in addition to polymer lid sealing, surface die coats, for example

with silicones, are applied. From Figure 4.29,[64] we can conclude that silicones do not act as a moisture barrier; the exact mechanism by which they protect the die when applied as a surface coat is not yet well understood.[65]

Examples of MEMS Packages

Introduction

Pressure sensors and accelerometers typically fit in standard packages, but few other MEMS devices do. Examples of nonstandard packages that we have already discussed include the catheter-based electrochemical sensor shown in Volume II, Figures 4.92 to 4.96, where a dual-lumen catheter forms the basis of the packaging scheme, and the thin-film magnetic head illustrated in Chapter 1, Figure 1.1d, this volume. The very fragmented nature of micromachining applications and the nonstandardness of the package necessitate a design approach starting from the package.

Cap-on-Carrier and Cap-on-Chip

In Figure 4.30, two MEMS packaging schemes are compared. The cap-on-carrier packaging method (L1) (Figure 4.30a) is suitable for accelerometers, gyroscopes, and other motion detectors that can be hermetically sealed. It is the prevalent method in industry today. The cap-on chip is an on-chip packaging method (Figure 4.30b), also called *chip scale packaging* (CSP), that is the preferred approach, as the packaging can be made part of the front-end batch manufacturing process and may be much smaller. The two approaches are contrasted in Figure 4.31.

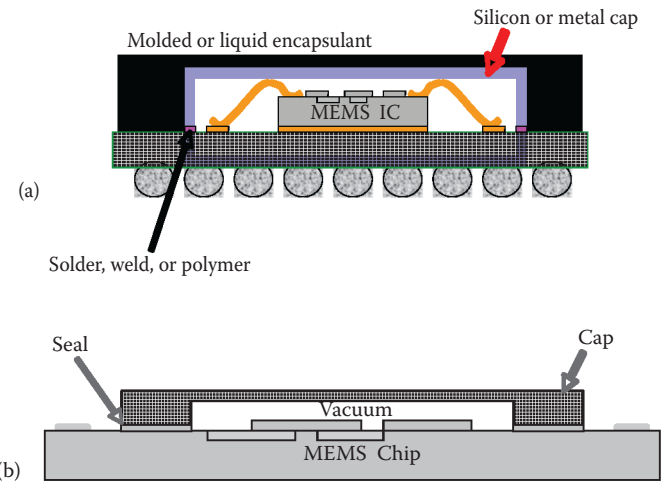


FIGURE 4.30 Cap-on-carrier (a) and cap-on-chip (b) compared.

One reoccurring problem in MEMS packaging is how to protect sensitive movable mechanical structures during dicing. The dicing process is very harsh, and any loose structure might break off. Fortunately, in most MEMS process sequences, the step where movable structures are released is the final so-called *release step* (typically, a sacrificial material is etched away to free the movable part; see Volume II, Chapter 7). By carrying this release step out after dicing, breakage of sensitive MEMS structures is avoided.

But individual die-level release is expensive and slow. Released dice, as shown in Figure 4.31, are packaged in ceramic or metal cavity packages. These packages often are large and expensive. Alternately, one does the MEMS release at the wafer level and then wafer-level packages the devices before dicing. Wafer-level packaging must follow the wafer-level release, again to avoid damaging the MEMS. The process is called *CSP* with caps-on-chip. These caps may be made of glass or another Si piece and typically are still quite bulky (millimeters rather than tens or hundreds of micrometers). These bulk wafer caps are

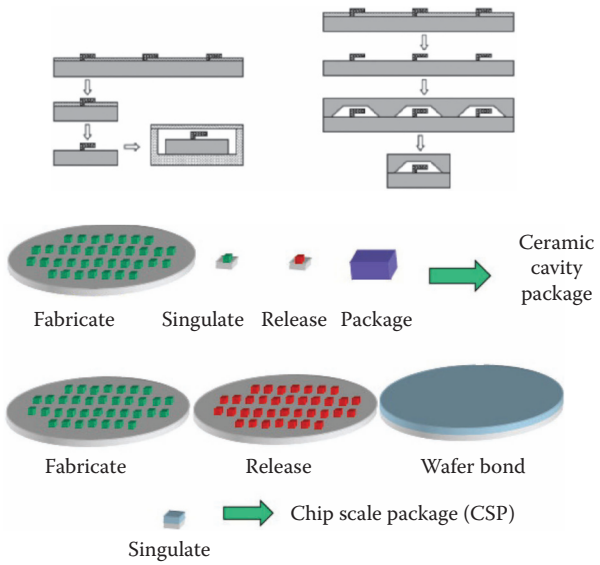


FIGURE 4.31 Microelectromechanical systems (MEMS) are either diced, then released and finally packaged in ceramic or metal cavity packages. Or one does the MEMS release at the wafer level and then wafer-level package the devices before dicing. The latter process is called *chip scale packaging* (CSP).

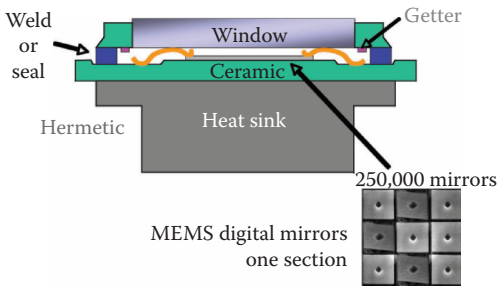


FIGURE 4.32 Packaged Digital Mirror Device™ from Texas Instruments utilizes a vacuum hermetic package with a transparent optical window. The package also includes a getter to adsorb any residual moisture.

the current industry standard, used, for example, in Motorola accelerometers, Bosch gyroscopes, Clarisay SAW filters, and Radant MEMS switches. The bulk wafer caps are robust, hermetic, and implemented on a wafer-level, but they require a large on-chip area for the seal ring. Below, we review a number of alternative wafer-level cavity sealings, i.e., cap-on-chip techniques such as the microassembled HEXSIL caps, techniques that are today mostly in the research and development phase. They all have a considerable smaller footprint than the bulk caps discussed here.

The Digital Mirror Device™ from Texas Instruments, discussed in Volume II, Example 7.2, utilizes a vacuum hermetic package with a transparent optical window as shown in Figure 4.32. The package also includes a getter to adsorb any residual moisture. This package is an example of a special type of cap-on-carrier package.

Next-Generation Cap-on-Chip Cavity Sealing

Introduction

Many microdevices have been fabricated making use of some type of cavity sealing or cap-on-chip technique. Cavity sealing can serve as a batch-compatible zero- and first-level packaging technique by encapsulating a die feature (L0) or a whole die at a time (L1). In bulk micromachining (Volume II, Chapter 4) and Si fusion-bonded (SFB) surface micromachining (Volume II, Chapter 7), cavities are fabricated by bonding, respectively, a glass plate (anodic bonding) or Si wafer (fusion bonding) over etched cavities in a bottom Si wafer. The Si and glass layers bonded typically are rather thick and do not lend themselves well to die feature level packaging. In polysilicon and selective epitaxy surface micromachining (see Volume II, Chapter 7), on the other hand, sealing cavities often is a more integral part of the overall fabrication process, and both die and die features might be packaged this way. These and other lithography-defined packages, such as those involving ultraviolet patternable polymers, lend themselves to inexpensive batch solutions and represent a new technology area where micromachining could provide solutions for the MEMS and IC industries.

Sealing of Polysilicon and Silicon Nitride Cavities

In Figure 4.33a, we illustrate how a polysilicon vacuum shell encapsulates a polysilicon resonator element. The micromachined surface package (microshell) illustrated is much smaller than typical bulk-micromachined packages. Microshells can be made by defining thin gaps (~100 nm) between the substrate and the perimeter of the structural elements by etching away a sacrificial layer sandwiched

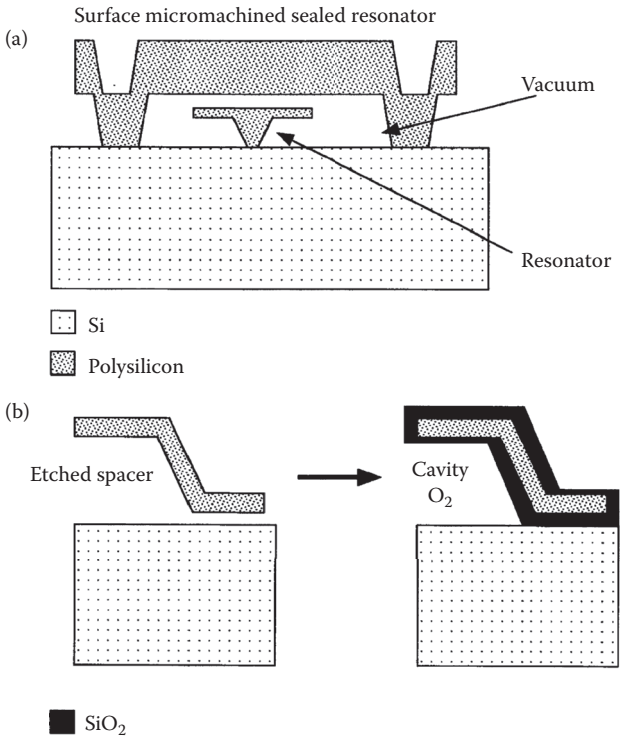


FIGURE 4.33 Sealed cavities in surface micromachining. (a) Typical sealed cavity with a resonator structure inside. (b) Reactive sealing.

between the two and then sealing the resulting gaps. In so-called *reactive sealing*, demonstrated in the schematic fabrication sequence in Figure 4.33b, thermal oxidation of the polysilicon and Si substrate at 1000°C seals the narrow openings left after removal of the spacer phosphosilicate glass (PSG).[66–68] Reactive sealing is possible even in vacuum because of the reaction of oxygen, trapped inside the cavity.

Alternatively, sealant films, such as oxides and nitrides, can be deposited over small etchant holes,[67,69] as illustrated in the fabrication sequence in Figure 4.34A. In the latter case, the cavity is sealed by low-pressure chemical vapor-deposited (LPCVD) silicon nitride. The excellent coverage of this method ensures that the nitride closes the etch channel quickly before too much deposition in the chamber itself can take place. Typical deposition conditions during the sealing step are 850°C and 250 mTorr, so the residual pressure in the microshell, at room temperature (assuming ideal gases) should be about 67 mTorr. Some researchers have reported residual pressures of 200–300 mTorr.[70] Eaton and Smith[71]

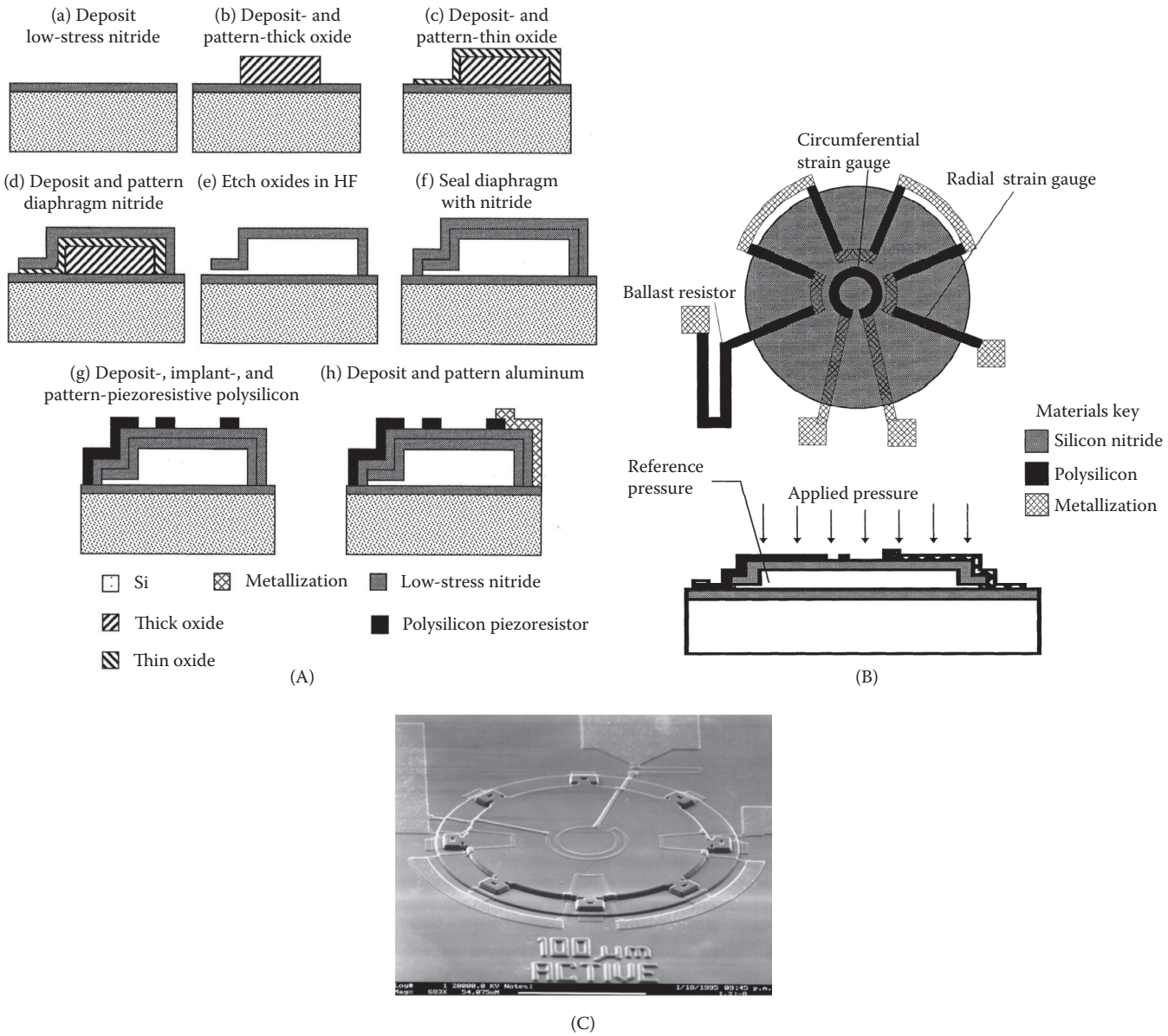


FIGURE 4.34 Absolute pressure sensor.[70] (A) Schematic fabrication sequence of vacuum shell with low-pressure chemical vapor-deposited Si nitride sealing. The vacuum shell constitutes the reference chamber of an absolute pressure sensor. (B) Schematic of absolute pressure sensor with Si nitride membrane. (C) Scanning electron microscope microphotograph of a finished device. (Courtesy of Dr. J. H. Smith, Sandia National Laboratories.)

established that the resulting cavity pressure is stable for a given membrane, but the residual pressure is variable and nonrepeatable across a substrate and even less from substrate to substrate.

The fabrication time of packaging shells often makes for a long process. The sacrificial material inside the shell is removed very slowly given the poor access via narrow etch holes to it by the etchant. Lebouitz et al. at the University of California, Berkeley, introduced an interesting means to speed up the fabrication of microshells.[72] The process is outlined in Figure 4.35. Permeable polysilicon windows (see Volume II, Chapter 7) are used as an etch access for removing the underlying sacrificial PSG. Using concentrated HF, shells 3 μm high and as wide as 1 mm have been cleared of PSG in less than 120 s. Subsequent low-pressure hermetic sealing using low-stress Si nitride leads to deposition of less than 100 Å inside the package.

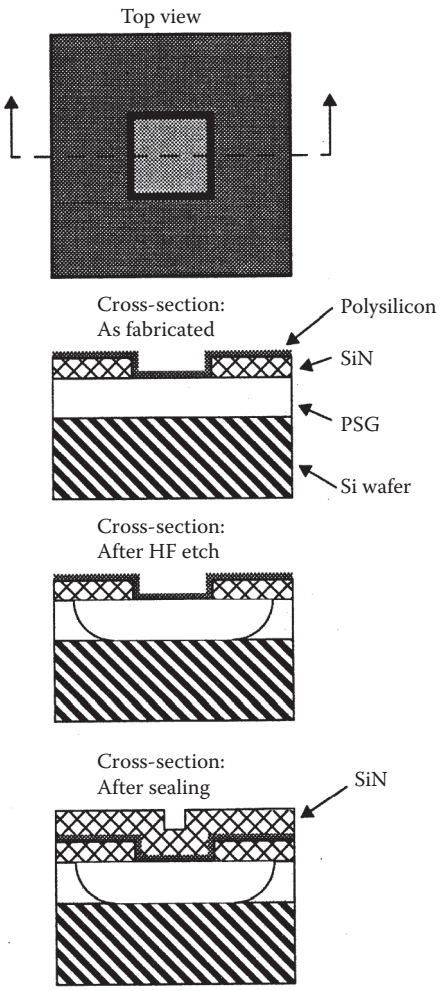


FIGURE 4.35 Schematic of etch access window design, operation, and sealing. [From Lebouitz, K. S., R. T. Howe, and A. P. Pisano. 1995. In *8th international conference on solid-state sensors and actuators (Transducers '95)*, 224–27. Stockholm, Sweden. With permission.[72]]

Sealed cavities, as shown here, lend themselves well to pressure sensing applications. Depending on the atmosphere to which the chip is exposed to during the sealing process, gauge, vacuum, or absolute pressure sensors can be created.[73,74] The process outline in Figure 4.34A actually illustrates the case of an absolute pressure sensor with a Si_3N_4 membrane.[71] The sensor consists of a circular Si nitride diaphragm, which forms the top of a sealed vacuum cavity, providing the pressure reference. Polysilicon strain gauges are fabricated on top of the diaphragm and the measured resistance changes are, to first order, directly proportional to the applied pressure. In Figure 4.34B, we show a schematic of the sensor, and in Figure 4.34C, we feature an scanning electron microscope (SEM) microphotograph of a finished device.

Of reactive sealing and sealant films, the reactive sealing process is by far the most elegant and highest in performance. The first commercial absolute polysilicon pressure sensor, incorporating a reactively sealed vacuum shell, was introduced for automotive applications in 1995 by SSI Technologies, but it is not a product offering anymore.[75] Another application of the above type of sealed surface shells is the vacuum packaging of lateral surface resonators. Most resonator applications share a need for resonance quality factors from 100 to 10,000. However, the operation of comb-drive microstructures, as shown in Volume II, Figure 7.76, in ambient atmosphere results in low-quality factors of less than 100 due to air damping above and below the moving microstructure.[76] Vacuum encapsulation is thus essential for high-Q applications. It also has been shown that lateral comb-drive microresonators,[77] besides their use in commercial accelerometers, have potential applications in areas such as mechanical filters for signal processing,[78] noncontact electrostatic voltage sensing,[79] and rotation-rate sensing.[80]

Epitaxial Deposition Cavity Sealing

Deposition of a set of epitaxial Si layers with varying doping has shown to afford the formation of a hermetically sealed surface cavity.[81] The cavity, in this

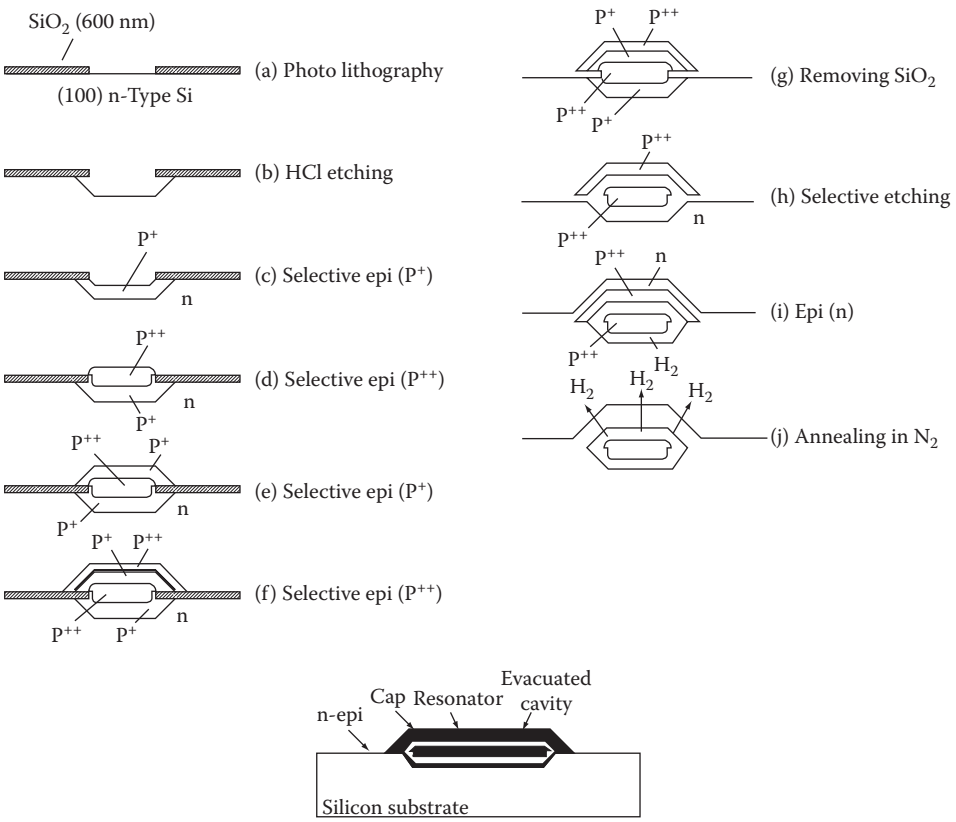


FIGURE 4.36 Epilayer-based fabrication process of a vacuum-encapsulated resonant beam. Process from (b) to (f) is carried out in one batch epitaxial process. For process step details see text. (From Ikeda, K., H. Kuwayama, T. Kobayashi, T. Watanabe, T. Nishikawa, T. Yoshida, and K. Harada. 1990. Silicon pressure sensor integrates resonant strain gauge on diaphragm. *Sensors Actuators A* A21:146–50. With permission.[82])

example of selective epitaxy surface micromachining, is formed by selective etching of p^+ epitaxial Si over more heavily doped Si p^{++} layers. The process is illustrated in Figure 4.36. The fabrication starts with a HCl dry etch at 1050°C in an epireactor through a hole in an oxide mask (Figure 4.36, steps a and b). Then follows a selective epitaxial growth of silicon in the following sequence: doped p^+, p^{++}, and p^+, p^{++} as indicated in Figure 4.36, steps c, d, e, and f. These steps are all carried out in the same epireactor, simply by changing the concentration of the B_2H_6 dopant. The next step consists of stripping of the oxide in HF (Figure 4.36, step g), followed by selective electrochemical etching in hydrazine (step h). In this electrochemical etching step, the n-substrate is passivated against the etching by the imposed potential and the p^{++} structures by the boron etch-stop mechanism. At this juncture, one ends up with a microbridge covered by a cap, both made out of single-crystalline, heavily boron-doped Si (p^{++}). The cap finally is sealed by growth of an n-type epitaxial layer (Figure 4.36, step i). The residual hydrogen in the cavity, after sealing, is diffused through the epitaxial layer by high-temperature annealing in a nitrogen ambient, resulting in a residual pressure inside the cavity of 1 mTorr (step j).[81,82] To make a pressure sensor that incorporates such an encapsulated resonator in a suspended membrane, one has to fabricate a membrane by etching from the back side of the wafer; a rather trivial step compared with the process just described. In the particular device shown here, the resonance of the encapsulated beam is activated by a Lorentz force and detected by measuring the resulting inductance.

Microassembled Caps or HEXSIL Cavity Sealing

Researchers at the University of California, Berkeley (UCB) developed another vacuum encapsulation method suitable for L0 and L1 level

packaging involving the wafer-to-wafer transfer of micromachined caps as demonstrated in Figure 4.37a.[83] Reactive sealing requires a temperature of 1000°C, and thick SiN sealing requires 850°C. Moreover, in the latter technique some sealant gas deposits on the encapsulated microdevices.

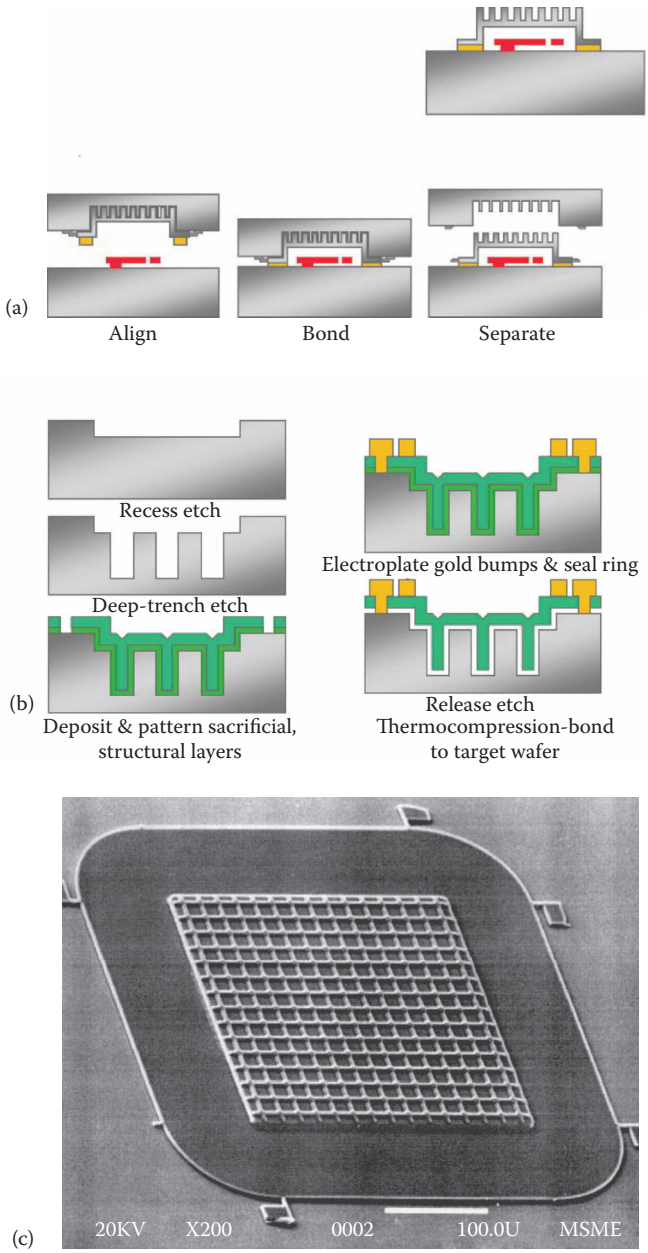


FIGURE 4.37 Wafer-to-wafer transfer of encapsulation structures. (From Hok, B., C. Dubon, and C. Ovren. 1983. Anodic bonding of galium arsenide to glass. *Appl Phys Lett* 43:267–69. With permission.[24]) (a) Principle of the wafer-to-wafer transfer method of encapsulation caps. (b) Fabrication of the tethered caps in a HEXSIL process. (c) Scanning electron microscope micrograph of a transferred cap.[82] (Courtesy of Dr. M. Cohen, University of California, Berkeley.)

In the alternative UCB packaging method, tethered cap structures are sealed down to the substrate employing a low-temperature Au-Si eutectic bond (the gold-silicon eutectic at 363°C is safely below that for aluminum-silicon). The encapsulation caps are made in the HEXSIL* process, as demonstrated in Figure 4.37b (see Volume II, Chapter 7, for details on the HEXSIL process). The caps have the general shape of a top hat and are suspended by breakaway polysilicon flexures in the handle wafer. The brim of the hat is coated with 0.7 µm of gold. The HEXSIL wafer with the embedded caps is positioned ($\Delta x <$ 10 µm) over the Si wafer with the active devices to be encapsulated. In situ annealing in vacuum by infrared heating forms the eutectic bond between the Si of the bottom wafer and the gold rim of the cap. The transfer occurs at a chamber pressure of 10^{-5} Torr, with 10 psi of mechanical pressure applied to the wafer sandwich. When the handle wafer is withdrawn, the polysilicon flexures break, leaving the caps sealed to the substrate wafer. In some limited test runs in a laboratory setting, a transfer yield of 100% was obtained for arrays of thirty caps. The HEXSIL mold wafer is reusable. An example of a transferred encapsulation cap is shown in the SEM micrograph in Figure 4.37c. To summarize, in the microassembled caps process, microcaps are fabricated on a donor wafer and transferred to a target wafer by wafer bonding and separation. The thin seal ring on the cap requires very little real estate (~1% of a bulk cap). The process potentially might be much less expensive than wafer-bonded caps as well.

Many of the newfound ways of hermetically sealing and bonding different layers and making contacts between them increasingly interest IC manufacturers. An important opportunity for micromachinists is to transfer the developed 3D machining technologies to the newest generations of 3D ICs.

Assembly and Self-Assembly

Introduction

Assembly is used to build complex products from relatively simple parts and to integrate incompatible

* HEXSIL is the acronym for hexagonal honeycomb poly-Si.

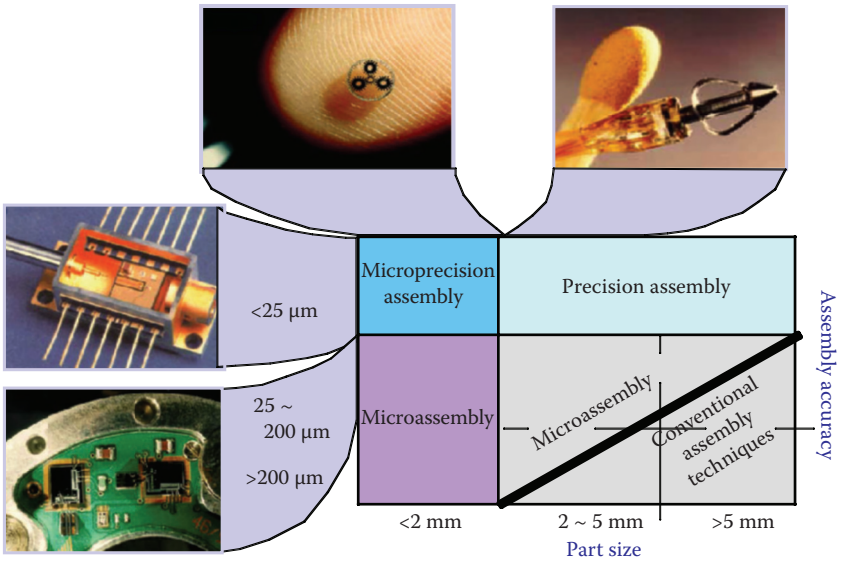


FIGURE 4.38 Assembly processes: part size assembled and assembly accuracy. (Courtesy of Kory Ehmann.)

manufacturing processes. Today, IC and MEMS assembly is mostly carried out by humans with tweezers and microscopes or with pick-and-place robots.[84] Micro- and mesoscale assembly processes handle objects from several micrometers to several millimeters in size. In Figure 4.38, we consider part size and assembly accuracy.

The current challenge is to assemble smaller and smaller components from different technologies with micro- and nanoprecision. The application field of micro- and nanoprecision assembly lies well beyond conventional or microscale assembly, with part dimensions larger than 2 mm (see Figure 4.38). The various micro- and nanoassembly approaches, divided into serial and parallel processes, are summarized in Table 4.3. In serial micro- and nanoassembly, parts are picked up and placed together on a substrate or workpiece one by one. Table 4.3 covers many methods enabling serial manipulation of very small parts/particles, including individual atoms for nanoassembly. Unless one can manipulate many parts/particles in parallel, most of these techniques will remain research tools rather than manufacturing options. In parallel micro- and nanoassembly, where parts are assembled simultaneously, we distinguish between deterministic and stochastic assembly.[84] In the former, the relationship between the part and its destination on a substrate is known in advance. In the latter, this relationship is unknown or random; parts "self-assemble" during a stochastic process driven by some motive force.

Stochastic self-assembly is an attractive option to integrate dissimilar process technologies. To achieve self-assembly, bonding forces must be present, the bonding must occur selectively, and the assembling parts must be able to move so that they can come together. Stochastic assembly often encompasses vibration of the parts in combination with electrostatic, fluidic, and other forces that operate on singulated parts in various media (fluids, air, or vacuum). After reviewing scaling issues of the assembly process, several examples of micro- and nanoassembly are presented. The smaller the building blocks, the higher the need for stochastic self-assembly, as handling the individual components becomes more and more of a mechanical challenge, and the process would be too slow. Nature, building on the nanoscale, does rely heavily on self-assembly (Chapter 2), and DNA-assisted self-assembly is an early attempt at co-opting nature for human manufacturing endeavors.

The best way to grab microparts is to use a microgripper, where grasping forces and jaw size match the requirements. Many kinds of microgrippers have been built. In Volume II, Figure 7.88, we show an example of a microgripper made of high-aspect-ratio molded polycrystalline silicon. The white bar at the bottom of the picture represents 100 μm. The actuator is an electrically heated

Table 4.3 Micro- and Nanoassembly Approaches

Serial Microassembly		Parallel Microassembly			
Type	**Refs.**	**Type**	**Refs.**	**Type**	**Refs.**
Deterministic		Deterministic		Stochastic	
Examples		**Examples**		**Examples**	
Manual assembly with tweezers and microscopes		Flip-chip wafer-to-wafer see, e.g., Figure 4.13	87, 88	Fluidic agitation and mating part shapes	89–91
Visually based and teleoperated microassembly	85, 86	Microgripper arrays for massively parallel pick-and-place (see Volume II, Figure 7.88 for an individual gripper)	94	Vibratory agitation and electrostatic force fields	95–99
High-precision macroscopic robots with submicron resolution	92, 93			Vibratory agitation and mating part shapes	103
Microtweezers	94, 100–102			Mating patterns of self-assembling monolayers	105
Optical trapping, see, e.g., Figure 2.33 this volume	104			Magnetic field-mediated separation	107
Microcapillary/pipette-based positioning, see, e.g., Figure 4.43	106			Dielectrophoretic trapping	109
AFM/STM atom manipulation, see, e.g., Volume II, Figures 2.43 to 2.45	108			Ultrasonic levitation and transportation in standing acoustic waves	110
				Hydrodynamic/electro-osmotic object manipulation in microchannels	111, 112

Source: Based on Böhringer, K. F., R. S. Fearing, and K. Y. Goldberg. 1998. In *Handbook of industrial robotics,* ed. S. Nof. New York: Wiley.[84]

thermal expansion beam, which causes the compound lever linkage to move the tips. Such grippers have been used to seize various microscopic objects, including 2.7-µm-diameter polystyrene spheres, dried red blood cells, and various protozoa.

Scaling of the Assembly Process

Scaling effects in microassembly represent a daunting task. The newest pick-and-place robots with submicrometer resolution must be used, as conventional robots that pick, transport, and place small parts have a control error of 100 µm at best, which would translate in relative errors of 100% on micrometer-sized parts. It might be difficult to pick up large objects with a robot gripper, but small parts (<1 mm and masses less than 10^{-6} g) are difficult to release, as adhesion forces between gripper and object are significant compared with gravitational forces at that scale.[84] In Chapter 7, this volume, on scaling, we will indeed learn that in the microdomain, surface forces such as adhesion dominate over volume forces such as gravity or inertia. Adhesive forces involved in pick-and-place include surface tension due to moisture on the part and gripper, van der Waals forces (significant for gaps below 100 nm), and electrostatic attractions (e.g., induced by triboelectrification).

In Figure 4.39, we compare the magnitude of attractive surface forces (electrostatic, van der Waals, and surface tension) with the gravitational force as a function of part size, assuming a sphere (part)-plane (gripper) interaction. It is observed that in the microworld, surface forces dominate over gravitational influence. At 1 µm the gravitational

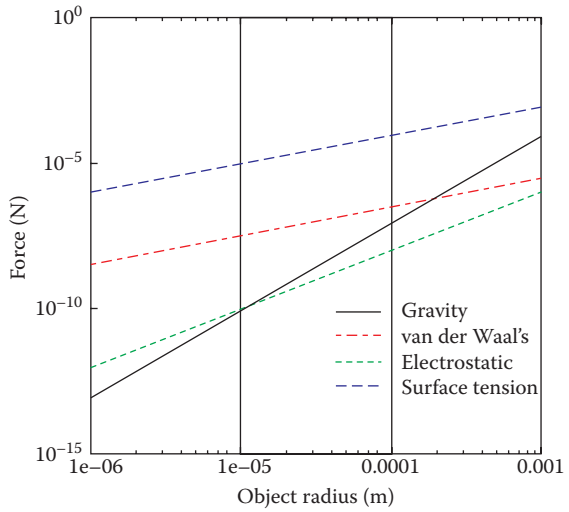


FIGURE 4.39 Comparison of the magnitude of attractive surface forces, i.e., electrostatic, van der Waals, and surface tension, with the gravitational force as a function of part size, assuming an interaction between sphere (part) and plane (gripper).

force is the least important force of the four, and at 10 µm the gravitational force equals the electrostatic force.

The electrostatic force is caused by charge attraction. Between a charged spherical object and a conductive plane it is approximated as follows:

$$F_{el} = \frac{\pi}{4\varepsilon_0}\frac{\varepsilon-\varepsilon_0}{\varepsilon+\varepsilon_0}d^2\sigma^2 \qquad (4.1)$$

where

d = the diameter of the particle
σ = the charge density on the surface
ε_0 and ε = the air's and the plane's dielectric-constants, respectively

The distance *z* between the sphere and the plane is assumed to be very small (Figure 4.40).

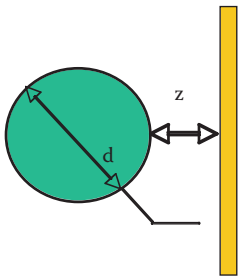


FIGURE 4.40 The electrostatic force between a spherical particle and a wall.

It is difficult to ensure that parts and grippers are electrically neutral because of the friction force and the difference in contact potentials between part and gripper. The gripper is typically coated with conductive coatings and grounded (the electrostatic force between two good insulators at 1 µm is 10^6 Pa!).

We already know that van der Waals forces may cause sticking together of MEMS parts and that this stiction may become permanent (see Volume II, Chapter 7, on stiction in surface micromaching). The van der Waals force is an atomic force caused by momentary movements of electrons, i.e., dipoles (see Volume I, Chapter 7). The sphere-plane van der Waals force is calculated as follows (see also Volume I, Table 7.8):

$$F_{vdw} = \frac{Hd}{12\,z^2} \qquad (4.2)$$

with the Hamaker constant *H* (see Volume I, Chapter 7, Table 7.6 for some typical values), the object diameter *d*, and the object-plane distance *z* (it is again assumed that $z \ll d$). This force becomes significant at less than 100 nm and increases with contact area (remember the popular study of Gecko Setae, with a force as high as 60 µN on a single seta; see Volume I, Chapter 7). Rough surfaces increase the effective distance *z* and therefore reduce the van der Waals force. Making the gripper surface rough might compromise ultimate positioning precision though. Also, gripper coatings can reduce the van der Waals force.

The presence of water on surfaces increases the normal force between the contacting surfaces as a result of capillary effects. The source of the liquid film between a gripper and part can be either a preexisting film of liquid and/or capillary condensates of water vapor from the environment. The surface tension force can be attractive or repulsive depending upon the configuration and the materials involved. The degree of hydrophobicity/hydrophilicity affects the contact angle (θ) and thus the surface tension force. The capillary tension force between a ball and a plate, with reference to Figure 4.41, is given by the following:

$$F_{tens} = \pi R_2^2 \gamma\left(\frac{1}{R_1}+\frac{1}{R_2}\right) \qquad (4.3)$$

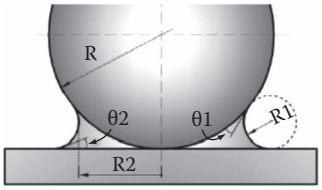


FIGURE 4.41 The capillary tension force between a ball and a plate.

where $\gamma = 72 \times 10^{-3}$ N/m is the water interfacial energy, and R_1 and R_2 are the axial and the tangential radius of the meniscus. To a good approximation, $2R_1$ corresponds to the thickness of the water film between the ball and the plane at the level of the meniscus, and both radii R_1 and R_2 can be expressed as function of ball radius R.

For small values of $2R_1$ the capillary force between the ball and the plane is given by the following equation:

$$F_{tens} = 2\pi R\gamma\left(2 + \sqrt{\frac{R_1}{R}}\right) \qquad (4.4)$$

For $R_1 << R$, the capillary force F_{tens} does not depend on the meniscus radius, and this relation simplifies to the following:

$$F_{tens} = 4\pi R\gamma = 2\pi d\gamma \qquad (4.5)$$

where R is the ball radius and d its diameter. When humidity is high, or with hydrophilic surfaces, a liquid film between a spherical object and a planar surface contributes a large capillary force. This effect can be reduced, for example, by working in a controlled environment or in a vacuum.

Surface forces make the release of objects after picking them up very difficult. To overcome these adhesion problems, some simple mechanisms for small-scale manufacture have been developed. They include electrostatic probes and vacuum probes. Also, the smallest components often are manipulated in a liquid medium (e.g., with laser trapping or dielectrophoresis). Microassembly obviously requires new handling techniques and electrostatic contact-less handling appears to be quite promising. Using a desktop factory (DTF) (see Volume II, Chapter 6), Nakao et al. used an electrostatic probe tip to assemble the microhouse shown in Figure 4.42.[113] The electrically charged tool can pick up and place

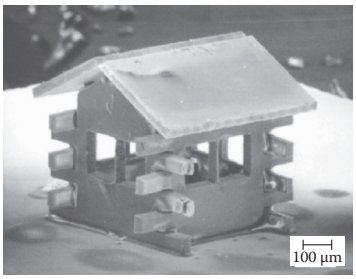


FIGURE 4.42 Scanning electron microscope of a microhouse. Using a desktop factory and an electrostatic probe tip, a microhouse was assembled.[113]

down micro-objects, and a semiconductor laser bonds both metal and nonmetal materials.

Vacuum pipettes, commonly used in microbiology for handling living cells, are another microassembly option. A vacuum gripper consists of a glass pipette and a vacuum control unit, and the actuating force is generated by the pressure difference on both sides of microcomponents. They are easy to use and are capable of picking, holding, and placing, and to overcome adhesion during placement they may use stripping off, pushing, or blowing away (see Figure 4.43).

On a large scale, vibration is used extensively in industrial part feeders that singulate, position, and orient parts before they are transferred to an assembly station. Sony's Advanced Parts Orientation System (APOS) part feeder, for example, uses an array of nests (silhouette traps) cut into a vibrating plate.[114] The traps and the vibratory motion are designed so that parts can remain in the nest only in one particular orientation. The traps in the vibrating plate are filled by tilting the plate and letting the parts flow across it. The APOS system with its controlled vibration to provide stochastic motion, combined with gravity as a motive force for nonprehensile manipulation, is a good model for microassembly. Again, scaling effects must be taken into account; vibration relies on the inertia of the parts and becomes less efficient with decreasing part size, so that higher vibration amplitudes or frequencies are required for smaller parts.

In the microscale, objects can be transported, for example, using acoustic traveling waves generated on a surface (see Chapter 8, this volume, Figure 8.28). Two types of waves are used, Lamb and Rayleigh waves. Lamb waves cause bending of the plate over its whole thickness (the substrate thickness is small

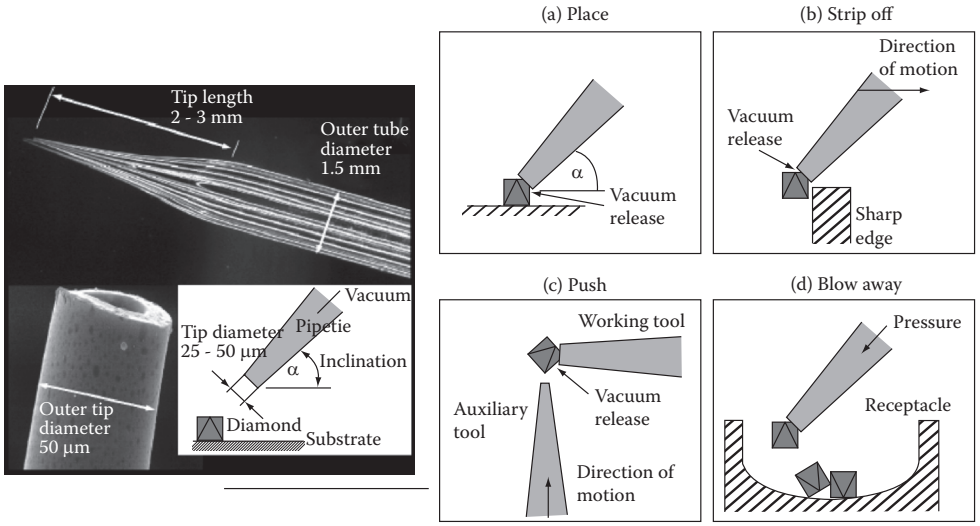


FIGURE 4.43 A glass pipette and a vacuum control unit are capable of picking, holding, and placing, and to overcome adhesion during placement they may use stripping off, pushing, or blowing away.

compared with the wavelength), while Rayleigh waves propagate only in the surface layer of the plate. Objects placed on the surface move in the opposite direction of the traveling wave (see Figure 8.28). Separation of particles may be effected by the competition between the acoustic force and the electrostatic force. The displacement of particles from pressure nodes in acoustic standing waves depends on the effective charge, radius, and stiffness of the particles. Yasuda has used lead zirconate titanate (PZT) transducers in a quartz cell to continuously concentrate, separate, and mix cells (e.g., blood cells).[115]

Another option is the use of distributed micromotion systems using arrays of tiny simple actuators that cooperate to move objects over relatively large distances and possibly in different directions and orientations. Several MEMS research groups have built such actuator arrays for micromanipulation; these arrays typically consist of a regular grid of "motion pixels" (mechanical members such as cilia).[116–123] These actuators are called *cilia* by analogy to their biological counterparts (see Chapter 8, this volume, Figure 8.11, and Chapter 5, this volume, Figure 5.13). Cilia arrays can move small objects placed on them by synchronous vibrations of the tiny hairs. To move objects in the x direction as well as the y direction, four cilia, one for each direction and sense, are grouped in motion pixels. The motion system consists of an array of these motion pixels. Appropriate control allows motion in any direction.

Table 4.4 summarizes some of the assembly issues one faces when manipulating parts in the mesoscale to the microscale and then into the nanoscale.

Serial Microassembly

In Volume II, Chapter 6, we saw how a 280-μm-high Japanese temple (Gojyunoto) on a polyimide platform was erected with the help of a microrobot manipulator. Many of the serial manipulation methods listed in Table 4.3 are used for manipulation of even smaller components, smaller than the typical MEMS device, such as biological cells, subcellular organelles, and even atoms.[125]

A small-enough object that is relatively transparent to laser light of a particular frequency refracts an incident laser beam, bending the light. As a result of this refraction, momentum is transferred from the light to the particle (see laser tweezers in Figure 2.33 this volume and in Volume I, Figure 5.34). When the geometry of light beams and particle is correct, the imparted momentum pulls the particle in the direction of the incident beam. The beam (or multiple beams) can then hold the particle in place, or by moving the beam(s), the laser operator can move the particle from place to place. In other words, light has momentum* and can be used to catch and manipulate objects in a size range from nanometers to micrometers in so-called *optical tweezers*.

* As far back as 1619, Johannes Kepler explained that the tails of comets are always directed away from the sun by the pressure of sunlight.

Table 4.4 Assembly Issues as a Function of Part Size[124]

Assembly Attribute	Assembly Scale		
	Mesoscale	Microscale	Nanoscale
Positioning	Easy	Difficult	Very difficult
Velocity	cm/s or m/s not unusual	Slow (μm/s or mm/s), vibration suppression	Very slow (nm/s or μm/s)
Force sensing and control	Easy, necessary to avoid part damage and improve manipulability	Difficult; range of forces could be as low as μN	Difficult; AFM used to measure force
Dominant forces	Gravity, friction	Surface forces (stiction, friction, electrostatic, Van der Waals)	Molecular/atomic forces
Throughput	Serial assembly provides adequate throughput	Serial assembly usually not sufficient; parallel manipulation methods preferred	Parallel manipulation methods or self-assembly necessary
Gripper	Mechanical	Micromechanical, gripper-free manipulation preferred	Other methods needed: optical, proximity forces, etc.
Fixturing	Mechanical	Micromechanical fixturing must be used	Chemical
Compliance	Gripper compliance not necessary if force measured	Gripper compliance usually necessary	Mechanical compliance not applicable
Vision	Easy	Difficult (expensive optics)	Impossible with visible wavelengths; SEM and TEM are used

The laser wavelengths in optical tweezers typically are between 0.7 and 1.06 μm, and 25 to 500 mW of power is generated in a focal spot between 0.5 and 1.0 mm in diameter. The first work with optical tweezers was performed by Ashkin et al. at AT&T Bell Laboratories. In pioneering works in 1985 and 1986, they successfully demonstrated that small particles could be trapped by focused low-power laser light (<1 W).[126,127] The first demonstration on trapping of living biological objects was also by Ashkin et al. in 1987.[104] The authors not only trapped and manipulated single cells, they demonstrated that the trap was harmless to living biological species. For example, reproduction of yeast cells was achieved in a trap capable of manipulating them at velocities of 100 m/s. Combined with other techniques (such as a laser scalpel), a number of important studies within microbiology have been carried out and many more can be visualized in the future. The difference between tweezers and scalpels is that the latter employ short pulses of high irradiance, whereas tweezers use continuous, low-irradiance beams. The laser scalpel can be used for cutting biological objects inside cells (e.g., DNA segments). Steven Chu,* a 1997 Nobelist in Physics (see Volume I, Figure 1.28) showed while at Stanford University that laser tweezers can be used to stretch molecules such as DNA: with colloidal spheres attached to the DNA strands, the spheres were trapped and the DNA was stretched by moving in fluid flows[128] (see also Chapter 2, Figure 2.33). The most common form of optical tweezers in biology is the "single-beam gradient trap," immobilizing transparent dielectric particles near the focal point of a tightly focused continuous-wave laser with a force proportional to the electric field gradient:[125,129,130]

$$F_{grad} = -n\left(\frac{1}{2}\alpha\nabla E^2\right) \tag{4.6}$$

where

n = the index of refraction of the solvent/medium

α = the polarizability of the particle

Such a beam generates forces in the piconewton range, more than sufficient for trapping cells and moving organelles inside and outside the cells. Usually, the technique cannot be used to lift objects much larger than the wavelength of the light. Links to pages related to optical tweezers can be found at http://www.phys.umu.se/exphys/OpticalTweezers.

Two different serial manipulation approaches can be used together. For example, cages for biological

* Dr. Steven Chu is currently the US Secretary of Energy.

cells have been made by combining optical trapping and dielectrophoresis.[131] Such a combination enables manipulation of two bioparticles in close proximity, which would be impossible using two laser beams or two electrical fields, as those fields would interfere. Dielectrophoresis is an AC electrokinetic effect (electro-osmosis and electrophoresis are direct current electrokinetic effects) discussed in detail in Volume I, Chapter 7. It is a powerful tool for separating particles of the size of biological cells but is more difficult to apply for particles below 1 µm.

In mechanosynthesis, an STM is used to manipulate individual atoms, as demonstrated in Volume II, Figure 2.43, where Pt atoms have been arranged on a Pt(111) surface. The Kanji character for *atom* in Figure 6.26 was also made this way. As pointed out in Chapter 1, serial manufacturing processes that use building blocks that are too small are not viable manufacturing candidates unless an army of them can work together on the same construction project.

Parallel Microassembly

The concept of self-assembly is inspired by biological processes in which complementary pairs such as antigens bind to antibodies on the basis of shape recognition. In the nonbiological world, objects can be similarly assembled on the basis of geometry. The advantage of self-assembly is that it is a batch process, in which hundreds or thousands of objects or even more are assembled at a time. When the interface is well designed, the objects are also self-aligned, and different types of objects can be assembled simultaneously by using different interfaces.

An example of deterministic parallel assembly by Cohn et al. for wafer-to-wafer transfer of MEMS and other microstructures was shown already, in Figure 4.37.[83,87] In this HEXSIL flip-chip process, the finished microstructures are suspended on breakaway tethers on the substrate. The target wafer with indium solder bumps is precisely aligned and pressed against the substrate, such that the microstructures are cold welded onto the target wafer. In another example of deterministic parallel microassembly, Keller et al. demonstrated microgrippers (see Volume II, Figure 7.88) and proposed gripper arrays for parallel transfer of palletized microparts.[94,102]

Many stochastic approaches for microassembly are under investigation. Yando, for example, employs an array of magnetic sites to assemble magnetically coated semiconductor parts in a square array. The components are vibrated at a gradually attenuated amplitude to "place" the array elements.[132] Similarly, Yeh et al.[90,91] trap semiconductor light-emitting diodes (LEDs), suspended in a liquid, in micromachined wells on a wafer by solvent-surface forces. As illustrated in Figure 4.44, carrier fluid containing the GaAs parts is dispensed over the host Si wafer with etched holes to assemble the GaAs blocks. Because of the trapezoidal design, the blocks fit preferentially into the holes in the design orientation. Random mechanical vibration of the microstructures enables large numbers of microstructures to be positioned into precise registration with the sites. More than 90% of the holes etched in Si were correctly filled by the blocks before the carrier fluid evaporated. However, during evaporation, surface tension pulls some objects out of the holes, reducing the yield to 30–70% locally. This problem can be addressed by using liquids with lower surface tension or by using supercritical drying methods.

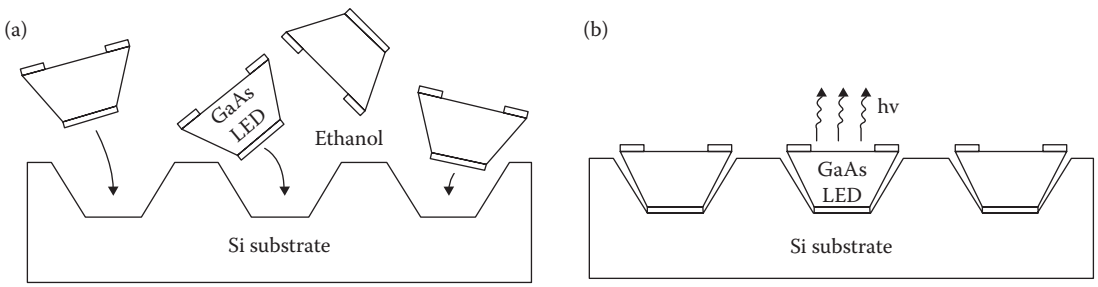


FIGURE 4.44 Fluid self-assembly of GaAs light-emitting diodes (LEDs) on Si micromachined substrate. (a) Solution containing the GaAs blocks is dispensed over the patterned silicon substrate. (b) Silicon substrate with GaAs LEDs integrated by fluidic self-assembly. [From Yeh, H.-J., and J. S. Smith. 1994. In *Proceedings of the IEEE International Workshop on Micro Electro Mechanical Systems (MEMS '94)*, 279–84. Oiso, Japan. With permission.[90]]

In 1991, Cohn et al.[96] first reported on stochastic assembly experiments that use vibration and gravitational forces to assemble periodic lattices of up to 1000 silicon chiplets.[133] Later, the same group[96] demonstrated an interesting variation on the work of Yando[132] and Yeh et al.[90,91] by using alignment capabilities over electrostatic traps. The process is illustrated in Figure 4.45. A critical problem in attempting to vibrate microstructures into specific sites from a liquid is that when structures adhere to the substrate or to each other, their progress toward optimal, precise registration is halted. This is caused by the same stiction phenomena we described for surface micromachined structures in Volume II, Chapter 7. Cohn et al. addressed this problem by levitating the microstructures a short distance above the target electrostatic trap site before letting them settle. As a result, random sticking is prevented, and critical positioning and orientation of the microstructure with respect to the trap become possible with a relatively low applied field (10 V/μm). The levitation over the trap site is accomplished by creating a short-range repulsive force between the microstructure and the target site. Microstructures with a relatively high permittivity (ε = 10) are suspended in a low-permittivity solvent (hexane), resulting in a net attraction to the charged electrodes making up the sites. However, a small amount of a more polar solvent (acetone) is added, which segregates into the high-field region between the two attracted parts. This prevents contact between the site and the suspended microstructure. Critical positioning and orientation become possible, in the manner of a compass needle resting on a low-friction bearing.[96] Once the sites are populated, the polar solvent is titrated out, gradually lowering the trapped devices to the target sites and pinning them onto the surface. A subsequent sintering step could effect permanent attachment—say, by eutectic bonding. Böhringer et al. subsequently demonstrated the ability to break surface forces using ultralow-amplitude vibration in vacuum, which holds promise as an extremely sensitive technique for positioning parts as well as discriminating part orientation, shape, and other physical properties.[97] Hosokawa et al.[134] modeled the kinetics of self-assembly and, using planar parts of various simple geometries floating on water,

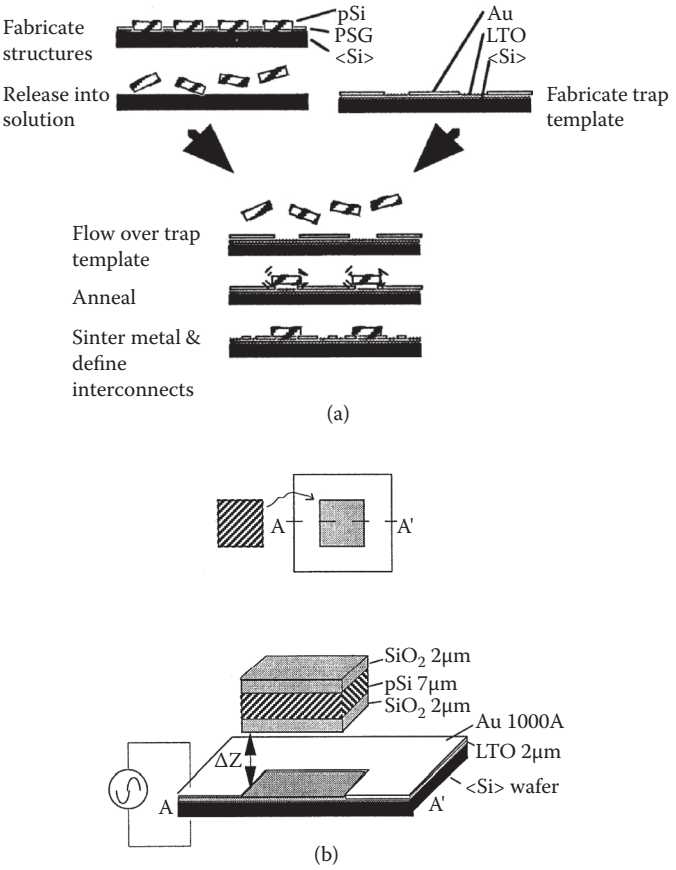


FIGURE 4.45 Self-assembly using electrostatic levitation. (a) Batch assembly of microstructures into binding sites on a substrate. Structures may include microelectromechanical system, complementary metal-oxide-silicon, optoelectronics, magnets, etc. (b) Detail of an electrostatic binding site. The section view shows position of a levitated microstructure. The levitation height (Δz) is in the range of 0–100 mm. The microstructure, a SiO_2-pSi-SiO_2 sandwich, has an average dielectric constant (ε) of 10, between that for hexane and acetone. The aperture in the gold film may range from ~2 to 100 mm in diameter, depending on the size of the microstructure to be trapped. (Courtesy of Dr. M. Cohen, University of California, Berkeley.)

demonstrated two-dimensional (2D) self-assembly. The units, consisting of a polyimide layer on a polysilicon layer (assembly units were 400 μm in size), selectively bonded to each other because of surface tension (a dominant force in the microscale; see above). The floating parts that were at equal height attracted each other. Sharp corners induced yet larger selective attractive forces, and parts located at different heights were mutually repulsed. To disturb the self-assembly system, magnetic or fluid forces were used. The floating units tended to align, with their sharpest features pointing toward each other.

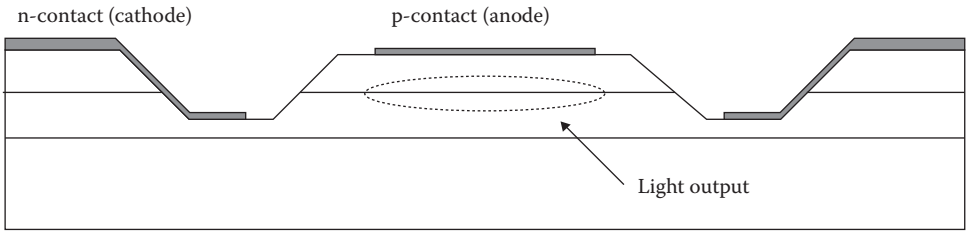


FIGURE 4.46 Schematic InGaAs microlight-emitting diodes have contacts on one side of the device as shown in the cross-section. (Courtesy of Dr. M. Heller, Nanogen.)

At Nanogen, Inc., electric fields are used to transport and control the placement of proteins, RNA, and DNA, as well as semiconductor components on a substrate equipped with an array of electrodes in a low-conductivity solution.[99,135] The electrophoretic nano assembly of DNA on the Nanogen platform is reviewed in Volume II, Chapter 8, Example 8.2. Here we illustrate the use of electric fields for precision placement of microlight-emitting diodes (microLEDs) (20 μm in diameter) onto a Si substrate using electro-osmosis (for details on electrophoresis and electro-osmosis, see Volume I, Chapter 7). The InGaAs microLEDs used all have contacts on one side of the device, as shown in the cross-section schematic in Figure 4.46. The corresponding substrate site has two sets of electrodes designed for the mounting process. As seen in Figure 4.47, a first pair of larger horseshoe electrodes brackets an inner set of electrodes. The inner set of electrodes is designed to mate with and bond to an LED and is made of standard tin/lead, electroplated to a height 50 nm above the surrounding surface. The linear structures seen in the pictures connect the exposed electrodes and are insulated to avoid shunting through the overlaying aqueous solution. Activation of the outer electrodes with respect to a counterelectrode (not shown in the figure) causes an electro-osmotic flow (see Volume I, Chapter 6) along the surface of the chip and moves the LED placed in the solution. Once the LED is on the electrode, the electro-osmotic flow is briefly reversed, and the reversal of the fluid flow drives the LED onto the awaiting bonding points

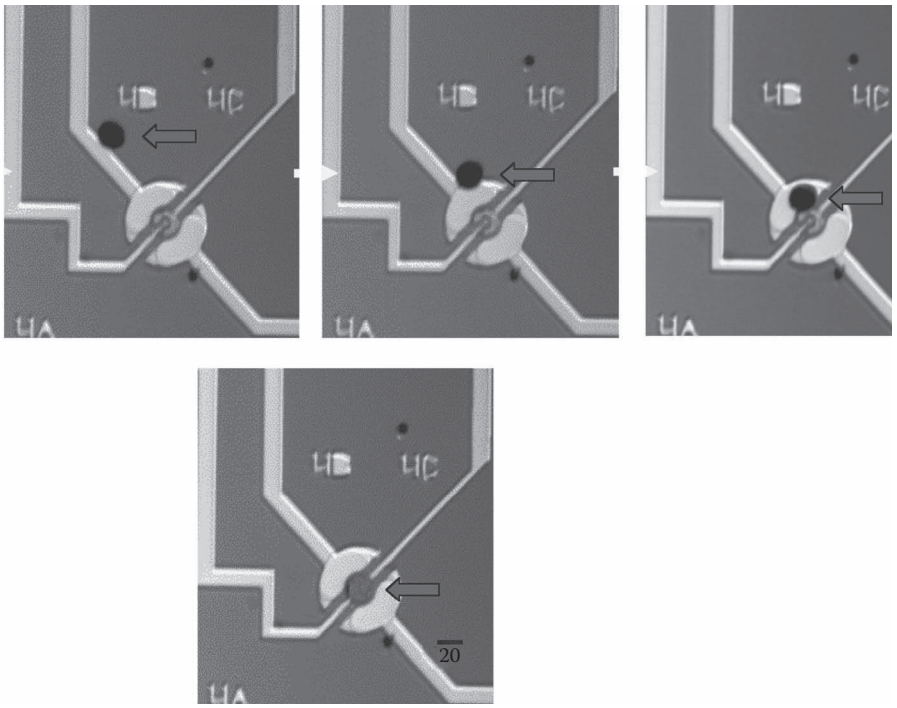


FIGURE 4.47 Electric field directed placement of light-emitting diodes (LEDs). The arrow points toward the moving LED. (Courtesy of Dr. M. Heller, Nanogen.)

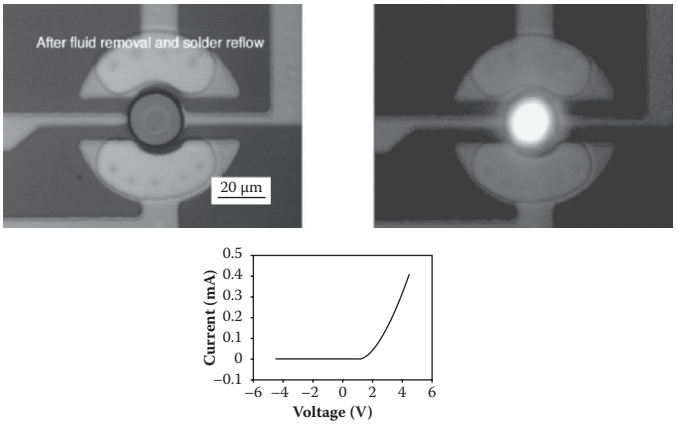


FIGURE 4.48 Circuit integrity and light-emitting diode (LED) activation demonstrated by forward biasing the LED. (Courtesy of Dr. M. Heller, Nanogen.)

with microscale precision. After the fluid is removed, the solder is reflowed under forming gas at 250°C. Circuit integrity and LED activation are demonstrated by forward-biasing the LED using the silicon circuit's contacts, as demonstrated in Figure 4.48.

The Nanogen authors point out that the same process could be used to assemble a large variety of components simultaneously and selectively on a circuit by the parallel activation of sets of electrodes while maintaining a number of components present in the low-conductivity solution. They speculate that the process might be aided by the use of a self-sorting mechanism by attaching short sections of complementary DNA on components and target sites on the circuit (see next section).[135]

Self-Assembly Examples

DNA-Mediated Assembly

While work continues on self-assembly of MEMS components with top-down methods, many researchers have recognized that the exquisite molecular recognition of various complementary biological molecules may be used to assist in the assembly of micro- or nanocomponents into complex networks. In this approach, DNA nanostructures and DNA mediated assembly of nano- and microstructures have received most attention in the literature. In the DNA nanostructures approach, pioneered by N. C. Seeman, the 2-nm-wide DNA molecules themselves are used to form 2D and 3D structures.[139–143] As shown in Figure 4.49a, DNA four-armed branched junctions are made by properly choosing the sequence of complementary strands. These DNA scaffolds can be linked together to make periodic crystals as shown in Figure 4.49b.

To use DNA for assembling nano- or microparts, one first needs to attach DNA molecules to these parts. The most widely used method is the covalent bond between sulfur and gold, that is, a metal thiolate (~44 kcal/mol). DNA molecules can be functionalized with thiol (S-H) or disulfide (S-S) groups, and upon immersion of microparts with clean gold surfaces in solutions of such thiol or disulfide derivatized oligonucleotides, the sulfur adsorbs on the gold surfaces, forming a single layer of molecules. Using this methodology, Mirkin et al. described a method of assembling colloidal gold

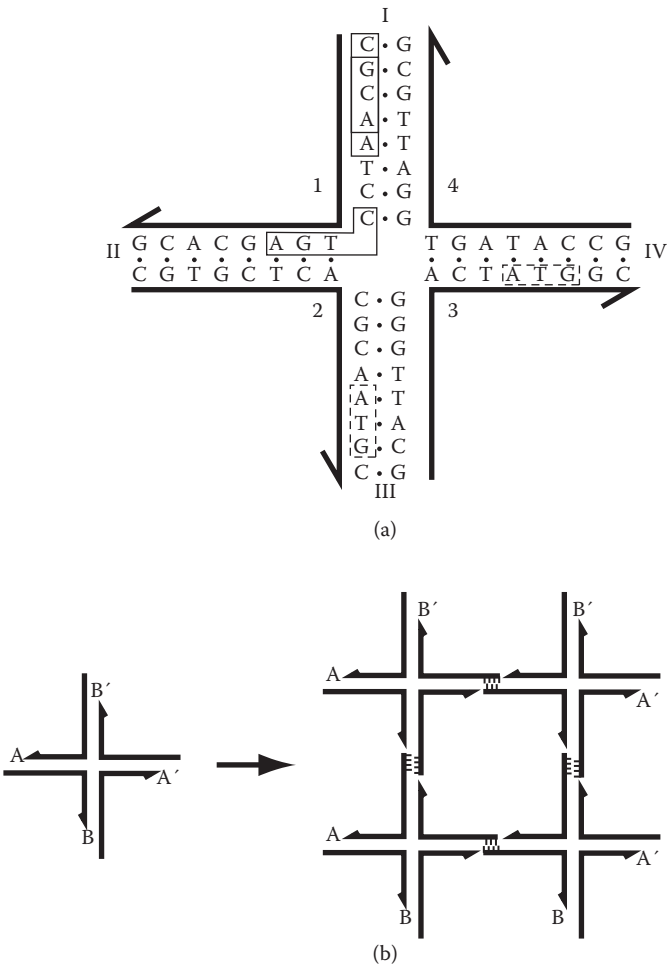


FIGURE 4.49 (a) Four-armed stable branched junction made from DNA molecules. (b) Use of branched junctions to form periodic crystals. (Based on Seeman, N. C. 1991. *Nanotechnology* 149.[140])

nanoparticles into macroscopic aggregates using DNA as linking element.[144] Noncomplementary DNA oligonucleotides were attached to the surfaces of two batches of 13-nm gold particles. An oligonucleotide duplex with ends complementary to the grafted sequences was then introduced into the particle mixture, and the nanoparticles self-assembled into aggregates. The process is illustrated in Figure 4.50. It can be reversed by increasing the temperature, which melts the double-stranded DNA linkers. Closely packed assemblies of aggregates with uniform particle spacings of about 60 Å were demonstrated this way.

The concept of DNA-mediated self-assembly has been extended to metallic nanowires/rods.[145] The idea behind this is to fabricate gold and/or platinum wires, functionalize them with single-stranded DNA (ssDNA), and assemble a wiring network on a substrate that has the complementary ssDNA molecules attached at specific sites. Thus, self-assembly of interconnects and wires is rendered possible. DNA has also been used as a template for the fabrication

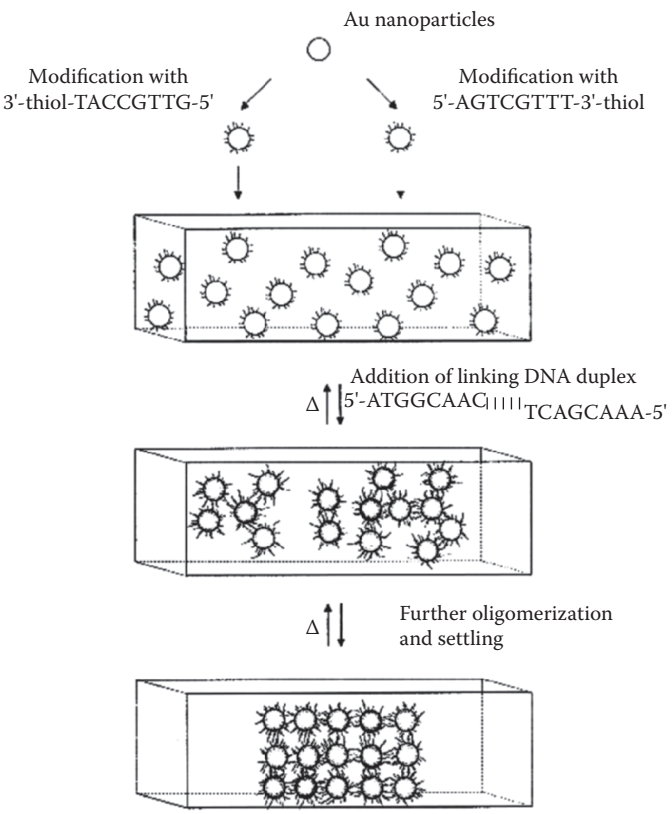


FIGURE 4.50 Fabrication process for the aggregated assembly of DNA conjugated gold nanoparticles. (Based on Mirkin, C. A., R. L. Letsinger, R. C. Mucic, and J. J. Storhoff. 1996. *Nature* 382:607.[144])

of nanowires. Braun et al.[136] formed a DNA bridge between two gold electrodes, again using thiol attachment. Once a DNA bridge is formed between the 12–16-μm spacing of the electrodes, a chemical deposition process is used to deposit silver ions along the DNA through Ag^+/Na^+ ion exchange and reduction of the silver complexes with hydroquinone. The result is a silver nanowire that is formed using DNA as a template or skeleton. The most ambitious goal is the DNA-assisted assembly (molecular "Velcro") of semiconductor active devices, such as transistors, laser diodes, etc., proposed by Edman et al. at Nanogen.[135] For a review of this area, check out Bashir[137] and Lee et al.[138]

Micro-Origami

Often MEMS includes hinged parts that must be set in place before operation. This can lead to challenging and time consuming manual manipulation of components at small length scales. Researchers from the ATR Adaptive Communications Research Laboratories in Japan (http://www.atr.co.jp) have developed a technique they call micro-origami to fabricate MEMS devices that automatically move parts into position.[146] Micro-origami makes use of the energy stored as strain between lattice-mismatched semiconductor epitaxial layers to create self-positioning, three-dimensional structures with simple and robust hinges. The key to the micro-origami technique is to manufacture hinges out of a pair of material layers with slightly different atomic spacings. The lattice mismatch causes a stress that in turn bends a hinge, as shown generically in Figure 4.51. In the case shown, a distributed Bragg reflector* (DBR) or micromirror is moved out of the plane of the substrate. Micro-origami structures are manufactured by using epitaxial growth techniques to construct multilayered films designed with folding hinges etched into their surfaces in the manner of

* A distributed Bragg reflector is a structure formed from multiple layers of alternating materials with various refractive indexes. Each layer boundary causes a partial reflection of an optical wave. For waves whose wavelength is close to four times the optical thickness of the layers, the many reflections combine with constructive interference, and the layers act as a high-quality reflector. The range of wavelengths that are reflected is called the *photonic stopband*. Within this range of wavelengths, light is "forbidden" to propagate in the structure.

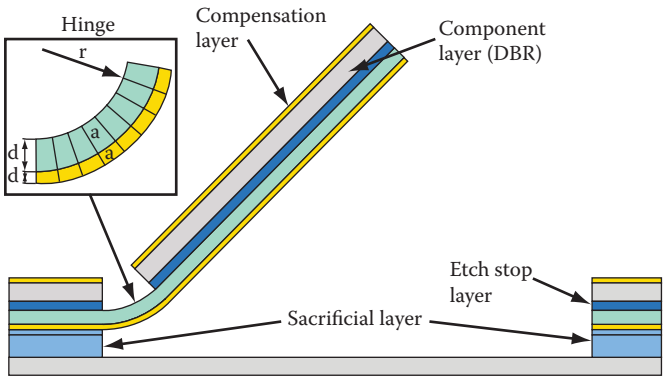


FIGURE 4.51 Schematic of the origami process. (From http://microfanatic.com/origami.html.)

the creased folds of paper. Adjusting the thickness and the length of the hinge that connects the plates to each other controls the angle between different plates. When the film is removed from its substrate by etching away a sacrificial layer, the microstructure takes on its designed shape because of the released stress in the hinge areas. In this way, this fabrication method reflects the design principles of origami, the traditional Japanese art of sculpting objects from folded paper. Because it is possible to fabricate three-dimensional structures on a semiconductor substrate, these structures can be integrated with lasers and photodetectors.

This fabrication method of simple and robust hinges for movable parts can be applied to any pair of lattice-mismatched epitaxial layers in semiconductors or metals. The first work by the ATR team involved a multilayer structure, including AlGaAs/GaAs component layers and an InGaAs strained layer, which was grown by molecular beam epitaxy on a GaAs substrate. After definition of the hinge and mirror's shape by photolithography, the micromirrors were released from the substrate by selective etching. They moved to their final position powered by the strain release in the InGaAs layer. The curvature radius of the hinges is determined by the thickness ratio of the bilayer and the indium composition in the strained layer. Depending on the growth structure, the hinge can fold up or down, making the fabrication of complex 3D semiconductor structures possible. For example, in Figure 4.52, we show GaAs corner plates assembled by the strain in two hinge regions, and in Figure 4.53 we illustrate a microstage fabricated on a GaAs substrate using

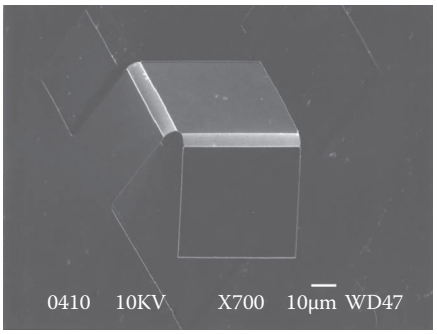


FIGURE 4.52 GaAs corner plates assembled by the strain in two hinge regions. (From http://microfanatic.com/origami.html.)

hinges with different bending directions called *tani-ori* (valley fold) and *yama-ori* (mountain fold) in origami. In the latter case, a multilayered structure including InGaAs strained layers was grown by molecular-beam epitaxy on a GaAs substrate. After the multilayered structure from the substrate is released by the selective etching of a sacrificial layer, the microstage moved into its final position powered by the strain release in the InGaAs layers, and the stage was kept parallel to the substrate owing to the combination of the two types of hinges.[147]

The ATR team also fabricated SiGe/Si free-standing micro-objects on an SOI substrate using SiO_2 as sacrificial layer material. SiGe/Si strained films were grown by molecular beam epitaxy on silicon-on-insulator (SOI) substrates and were released from the substrate by selective etching of the SiO_2 layer. The released films rolled up because of the elastic strain in the SiGe layer. Microtubes and microspirals with a diameter of about 40 μm were obtained as shown in Figure 4.54.[148]

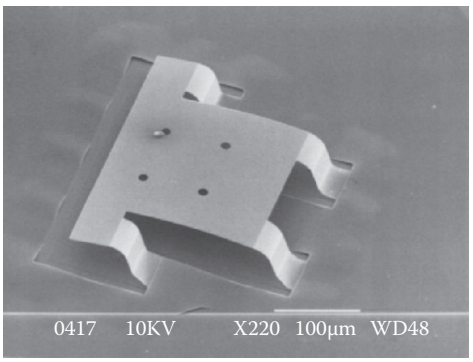


FIGURE 4.53 A self-assembled microstage fabricated using hinges with different bending directions. (From http://microfanatic.com/origami.html.)

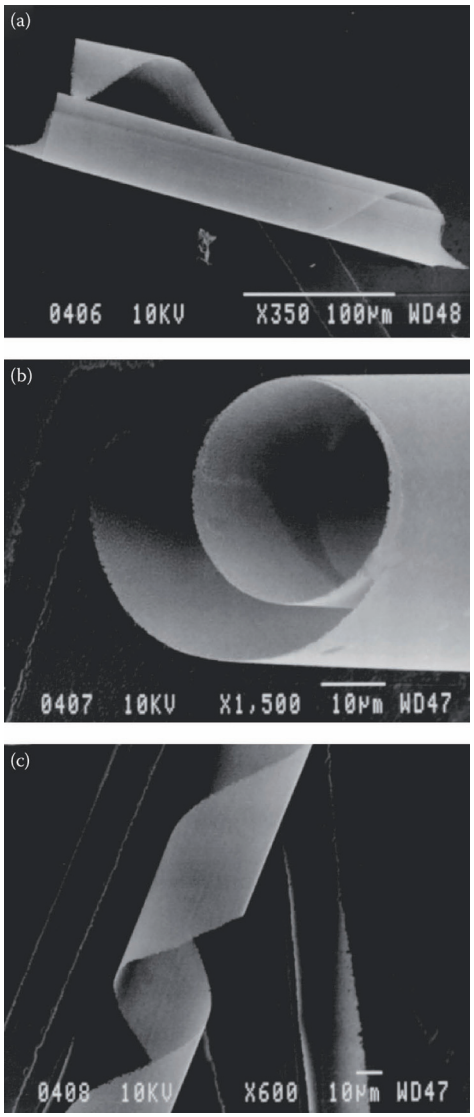


FIGURE 4.54 Free standing micro-objects on SOI substrate. (a) Top view of released SiGe/Si film scrolled in a microtube. (b) Rolled film at large magnification. (c) SiGe/Si microspiral. Curvature radius is about 20 µm.[148]

On a somewhat larger scale, researchers at the École Supérieure de Physique et de Chimie Industrielles in Paris, together with a team from the Paris Institute of Technology, have shown that water droplets can be used to make flat sheets of plastic (40–80 µm thick), a couple of millimeters across fold into 3D objects such as tiny pyramids, boxes, and spheres.[149] A drop of water, large enough to reach all corners, is placed on top of a cut-out shape. As the liquid evaporates, its surface tension tugs the membrane closer around the water's decreasing volume, until the remaining liquid is completely encapsulated inside a sphere, tetrahedron,

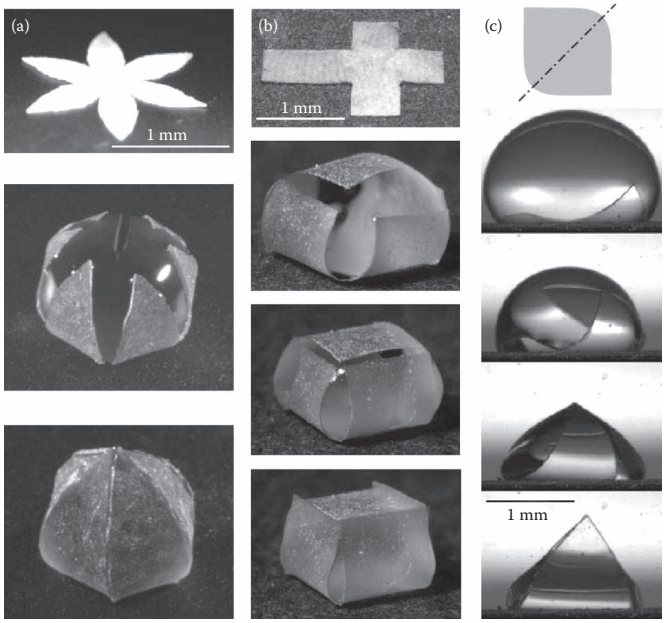


FIGURE 4.55 It is the interaction between elasticity and capillarity that is used to produce three-dimensional structures through the wrapping of a liquid droplet by a planar sheet. Tuning of the initial flat shape to obtain (a) a spherical encapsulation, (b) a cubic encapsulation, or (c) a triangular mode-2 fold.[149]

or some other shape. Tailoring the initial geometry of the thin membrane, as demonstrated with some experimental results in Figure 4.55, controls the final encapsulated 3D shape. For example, a flower-like pattern produces a sphere; a triangle becomes a tetrahedron, and a flat cross shape folds into a cube. Varying the thickness of the sheet controls how much a structure folds up, and as expected, the effect of surface tension becomes stronger as the size is decreased.

In Chapter 3, this volume, we saw several examples of DNA origami (see Figures 3.5 and 3.6).

The intersection between the subjects of origami and mathematics is rich with interesting results and applications. For details, consult http://kahuna.merrimack.edu/~thull/origamimath.html and http://www.paperfolding.com/math.

Questions

4.1: Explain front-end and back-end processes in IC or MEMS manufacturing. To which hierarchical level of manufacturing is each one related?

4.2: What are some of the advantages of a silicon-silicon bonding process over anodic bonding? Are there any disadvantages?

4.3: What is the role of packaging in ICs and in chemical and biological MEMS?

4.4: Which is easier to package, an absolute pressure sensor or an accelerometer? A pressure sensor or a pH sensor?

4.5: Describe and compare different wafer bonding techniques.

4.6: Describe the process of silicon-to-glass anodic bonding. Make a sketch of the setup and describe typical conditions required. What is the process that takes place on the molecular level?

4.7: Suggest a method for preventing an accelerometer sensor element from being subjected to undesirable mechanical stresses (from temperature changes or packing stress).

4.8: One of the modifications of anodic bonding is to deposit intermediate layers of silicon dioxide and aluminium. What are the advantages and disadvantages of this method?

4.9: Surface forces make the release of objects after picking them up in microassembly very difficult. Suggest several methods to make this process easier.

4.10: What are the three major electrical connection methods in chip-mounting?

4.11: Describe at least four techniques one could use to make a via through an Si wafer.

4.12: Describe as many methods that you know to vacuum seal a MEMS device on-chip.

4.13: Describe the difference between serial microassembly and parallel microassembly.

4.14: How do gravity, van der Waal's, electrostatic, and surface tension scale as a function of part size in microassembly, assuming a sphere (part) and a plane (gripper) interaction?

4.15: What is likely to be a major concern with DNA-assisted assembly? How could you improve the situation?

References

1. Lyke, J. C. 1995. Packaging technologies for space-based microsystems and their elements. *Microengineering technologies for space systems*. Ed. H. Helvajian, 131–80. El Segundo, CA: Aerospace Corporation.
2. de Gennes, P. G. 1979. *Scaling concepts in polymer physics*. Ithaca, NY: Cornell University Press.
3. Frank-Kamenetskii, M. D. 1997. *Unraveling DNA: The most important molecule of life*. Reading, MA: Perseus Books.
4. Jensen, J. R. 1989. *Microelectronics processing*. Eds. D. W. Hess and K. F. Jensen, 441–504. Washington, DC: American Chemical Society.
5. Allen, R. 1984. Sensors in silicon. *High Technology* September:43–81.
6. Lau, J. H. 1998. *Electronic packaging: Design, materials, process, and reliability*. New York: McGraw-Hill.
7. Jerman, J. H. 1990. The fabrication and use of micromachined corrugated silicon diaphragms. *Sensors Actuators A* A23:988–92.
8. Spiering, V. L., S. Bouwstra, J. Burger, and M. Elwenspoek. 1993. Membranes fabricated with a deep singel corrugation for package stress reduction and residual stress relief. *Fourth European Workshop on Micromechanics (MME '93)*. Neuchatel, Switzerland.
9. Spiering, V. L., S. Bouwstra, R. M. E. J. Spiering, and M. Elwenspoek. 1991. On-chip decoupling zone for package-stress reduction. *Sixth International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco.
10. Offereins, H. L., H. Sandmaier, B. Folkmer, U. Steger, and W. Lang. 1991. Stress free assembly technique for a silicon based pressure sensor. *Sixth International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco.
11. Tanger, R. K. 1976. Hermeticity of polymeric lid sealants. *26th Electronic Components Conference*. San Francisco.
12. Flannery, A. F., N. J. Mourlas, C. W. Storment, S. Tsai, S. H. Tan, J. Heck, D. Monk, T. Kim, B. Gogoi, and G. T. Kovacs. 1998. PECVD silicon carbide as a chemically resistant material for micromachined transducers. *Sensors Actuators A* 70:48–55.
13. Linder, S., H. Baltes, F. Gnaedinger, and E. Doering. 1994. Fabrication technology for wafer through-hole interconnections and three-dimensional stacks of chips and wafers. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '94)*. Oiso, Japan.
14. Maluf, N. 2000. *An introduction to microelectromechanical systems engineering*. Boston: Artech House.
15. Bhardwaj, J. K., and H. Ashraf. 1995. Advanced Si etching using high density plasmas. *Micromachining and microfabrication process technology (proceedings of the SPIE)*. Austin, TX.
16. Craven, D., K. Yu, and T. Pandhumsoporn. 1995. Etching technology and applications for "through-the-wafer" silicon etching. *Micromachining and microfabrication process technology (proceedings of the SPIE)*. Austin, TX.
17. Chow, E. M., A. Partridge, C. F. Quate, and T. W. Kenny. 2000. Through-wafer electrical interconnects compatible with Stanford semiconductor processing. *Solid-State Sensor and Actuator Workshop*. Hilton Head Island, SC.
18. Lischner, D. J., H. Basseches, and F. A. D'Altroy. 1985. Observations of the temperature gradient zone melting process for isolating small devices. *J Electrochem Soc* 132:2991–96.
19. Wallis, G., and D. I. Pomerantz. 1969. Field assisted glass-metal sealing. *J Appl Phys* 40:3946–49.
20. Ko, W. H., J. T. Suminto, and G. J. Yeh. 1985. Bonding techniques for microsensors. *Micromachining and micropackaging of transducers. Eds.* C. D. Fung, P. W. Cheung, W. H. Ko, and D. G. Fleming. Amsterdam: Elsevier.

21. Peeters, E. 1994. *Process development for 3D silicon microstructures, with application to mechanical sensor design*. PhD thesis, Catholic University of Louvain, Belgium.
22. Puers, B., A. Cozma, and E. Van De Weyer. 1993. *Intermediate project report*. KU Leuven.
23. Anthony, T. R. 1983. Anodic bonding of imperfect surfaces. *J Appl Phys* 54:2419–28.
24. Hok, B., C. Dubon, and C. Ovren. 1983. Anodic bonding of galium arsenide to glass. *Appl Phys Lett* 43:267–69.
25. Von Arx, J., B. Ziaie, M. Dokmeci, and K. Najafi. 1995. Hermeticity testing of glass silicon packages with high-density on-chip feedthroughs. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
26. Ito, N., K. Yamad, H. Okada, M. Nishimura, and T. Kuriyama. 1995. A rapid selective anodic bonding method. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
27. Sander, C. S. 1980. A bipolar-compatible monolithic capacitive pressure sensor. PhD thesis, #G558-10, Stanford University, December 1980.
28. Brooks, A. D., and R. P. Donovan. 1972. Low temperature electrostatic Si-to-Si seals using sputtered borosilicate glass. *J Electrochem Soc* 119:545–46.
29. Hanneborg, A., M. Nese, H. Jacobsen, and R. Holm. 1992. Silicon-to-thin-film anodic bonding. *J Micromech Microeng* 2:117–21.
30. Krause, P., M. Sporys, E. Obermeier, K. Lange, and S. Grigull. 1995. Silicon to silicon anodic bonding using evaporated glass. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
31. Elwenspoek, M., H. Gardeniers, M. de Boer, and A. Prak. 1994. *Micromechanics. Report no. 122830*. Twente, The Netherlands: University of Twente.
32. Esashi, M., A. Nakano, S. Shoji, and H. Hebiguchi. 1990. Low-temperature silicon-to-silicon bonding with intermediate low melting point glass. *Sensors Actuators A* A23:931–34.
33. Anthony, T. R. 1985. Dielectric isolation of silicon by anodic bonding. *J Appl Phys* 58:1240–47.
34. Ohashi, H., J. Ohura, T. Tsukakoshi, and M. Shimbo. 1986. Improved dielectrically isolated device integration by silicon-wafer direct bonding (SDB) technique. *Technical Digest: IEEE International Electron Devices Meeting (IEDM '86)*. Los Angeles, pp. 211–13.
35. Kissinger, G., and W. Kissinger. 1993. Void-free silicon-wafer-bond strengthening in the 200–400°C range. *Sensors Actuators A* A36:149–56.
36. Schmidt, M. A. 1994. Silicon wafer bonding technology for micromechanical devices. *Technical digest: 1994 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC.
37. Sun, L. G., J. Zhan, Q. Y. Tong, S. J. Xie, Y. M. Caim, and S. J. Lu. 1988. Cool plasma activated surface in silicon wafer direct bonding technology. *J Physique Colloq C* 49:79–82.
38. Shimbo, M., K. Furukawa, and K. Tanzawa. 1986. Silicon-to-silicon direct bonding method. *J Appl Phys* 60:2987–89.
39. Tong, Q.-Y., G. Cha, R. Gafiteanu, and U. Gosele. 1994. Low temperature wafer direct bonding. *J Micromechan Syst* 3:29–35.
40. Barth, P. W. 1990. Silicon fusion bonding for fabrication of sensors, actuators and microstructures. *Sensors Actuators A* A23:919–26.
41. Stengl, R., T. Tan, and U. Gosele. 1989. A model for the silicon wafer bonding process. *Jpn J Appl Phys Part I* 28:1735–41.
42. Backlund, Y., K. Ljungberg, and A. Soderbarg. 1992. A suggested mechanism for silicon direct bonding from studying hydrophilic and hydrophobic surfaces. *J Micromech Microeng* 2:158–60.
43. Lasky, J. B. 1986. Wafer bonding for silicon-on-insulator technologies. *Appl Phys Lett* 48:78–80.
44. Ismail, M. S., R. W. Bower, J. L. Veteran, and O. J. Marsh. 1990. Silicon nitride direct bonding. *Electron Lett* 26:1045–46.
45. Bower, R. W., M. S. Ismail, and B. E. Roberds. 1993. Low temperature Si_3N_4 direct bonding. *Appl Phys Lett* 62:3485–97.
46. Ismail, M. S., R. W. Bower, B. E. Roberds, and S. N. Farrens. 1993. *Seventh International Conference on Solid-State Sensors and Actuators (Transducers '93)*. Yokohama, Japan.
47. Anacker, W., E. Bassous, F. F. Fang, R. E. Mundie, and H. N. Yu. 1976. Fabrication of multiprobe miniature electrical connector. *IBM Tech Bull* 19:372–74.
48. Editorial. 1981. Sealing glass. *Corning Technical Publication*. Corning Glass Works.
49. Legtenberg, R., S. Bouwstra, and M. Elwenspoek. 1991. Low-temperature glass bonding for sensor applications using boron oxide thin films. *J Micromech Microeng* 1:157–60.
50. Field, L. A., and R. Muller. 1990. Fusing silicon wafers with low melting temperature glass. *Sensors Actuators A* A23:935–38.
51. Yamada, A., T. Kawasaki, and M. Kawashima. 1987. SOI wafer bonding with spin-on glass as adhesive. *Electronic Lett* 23:39–40.
52. Quenzer, H. J., and W. Benecke. 1992. Low-temperature silicon wafer bonding. *Sensors Actuators A* A32:340–44.
53. Guerin, L., M. A. Schaer, R. Sachot, and M. Dutoit. 1995. Proposal for new multichip-on-silicon packaging scheme. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
54. Tiensuu, A.-L., J.-Å. Schweitz, and S. Johansson. 1995. In situ investigation of precise high strength micro assembly using Au-Si eutectic bonding. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
55. Arquint, P., P. D. van der Wal, B. H. van der Schoot, and N. F. de Rooij. 1995. Flexible polysiloxane interconnection between two substrates for microsystem assembly. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
56. den Besten, C., R. E. G. van Hal, J. Munoz, and P. Bergveld. 1992. Polymer bonding of micro-machined silicon structures. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '92)*. Travemunde, Germany.
57. Munoz, J., A. Bratov, R. Mas, N. Abramova, C. Dominguez, and J. Bartroli. 1995. Packaging of ISFETs at the wafer level by photopatternable encapsulant resins. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
58. Becker, H., and C. Gärtner. 2000. Polymer microfabrication methods for microfluidic analytical applications. *Electrophoresis* 21:12–26.
59. Bower, R. W., M. S. Ismail, and S. N. Farrens. 1991. Aligned wafer bonding: A key to three dimensional microstructures. *J Electron Mater* 20:383–87.
60. Shoaf, S. E., and A. D. Feinerman. 1994. Aligned Au-Si eutectic bonding of silicon structures. *J Vac Sci Technol* A12:19–22.
61. Maszara, W. P., G. Goetz, A. Caviglia, and J. B. McKitterick. 1988. Bonding of silicon wafers for silicon-on-insulator. *J Appl Phys* 64:4943–50.

62. Nese, M., R. W. Bernstein, I.-R. Johansen, and R. Spooren. 1995. New method for testing hermeticity of silicon sensor structures. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
63. Minami, K., T. Moriuchi, and M. Esashi. 1995. Cavity pressure control for critical damping of packaged micro mechanical devices. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
64. Striny, K. M. 1988. Assembly techniques and packaging of *VLSI* devices. *VLSI Technology*. Ed. S. M. Sze, 566–611. New York: McGraw-Hill.
65. Traeger, R. K. 1976. Hermeticity of polymeric lid sealants. *26th Electronic Components Conference*. San Francisco.
66. Guckel, H., and D. W. Burns. 1984. Planar processed polysilicon sealed cavities for pressure transducer arrays. *Technical Digest: IEEE International Electron Devices Meeting (IEDM '84)*. San Francisco.
67. Guckel, H., and D. W. Burns. 1986. Fabrication techniques for integrated sensor microstructures. *IEEE International '86*. Los Angeles.
68. Guckel, H., and D. W. Burns. 1985. A technology for integrated transducers. *International Conference on Solid-State Sensors and Actuators.* Philadelphia.
69. Guckel, H., D. W. Burns, C. K. Nesler, and C. R. Rutigliano. 1987. Fine grained polysilicon and its application to planer pressure transducers. *Fourth International Conference on Solid-State Sensors and Actuators (Transducers '87)*. Tokyo, Japan.
70. Lin, L., K. M. McNair, R. T. Howe, and A. P. Pisano. 1993. Vacuum-encapsulated lateral microresonators. *Seventh International Conference on Solid-State Sensors and Actuators (Transducers '93)*. Yokohama, Japan.
71. Eaton, W. P., and J. H. Smith. 1995. A *CMOS-compatible, surface-micromachined pressure sensor* for aqueous ultrasonic application. *Smart Structures and Materials 1995: Smart Electronics (Proceedings of the SPIE)*. San Diego, CA.
72. Lebouitz, K. S., R. T. Howe, and A. P. Pisano. 1995. Permeable polysilicon etch-access windows for microshell fabrication. *Eighth International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden.
73. Guckel, H., D. W. Burns, and C. R. Rutigliano. 1986. In *Technical Digest: 1986 Solid State Sensor and Actuator Workshop.* Hilton Head Island, SC.
74. Erskine, J. C. 1983. Polycrystalline silicon-on-metal strain gauge transducers. *IEEE Trans Electron Devices* 30: 796–801.
75. *SSI Technologies.* 1995. Janesville, WI: SSI Technologies.
76. Cho, Y.-H., B. M. Kwak, A. P. Pisano, and R. T. Howe. 1993. Viscous energy dissipation in laterally oscillating planar microstructures: a theoretical and experimental study. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '93)*. Fort Lauderdale, FL.
77. Tang, W. C., T. H. Nguyen, and R. T. Howe. 1989. Laterally driven polysilicon resonant microstructures. *Sensors Actuators* 20:25–32.
78. Lin, L., T. C. Nguyen, R. T. Howe, and A. P. Pisano. 1992. Microelectromechanical filters for signal processing. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '92)*. Travemunde, Germany.
79. Hsu, C. H., and R. S. Muller. 1991. Micromechanical electrostatic voltmeter. *Sixth international conference on solid-state sensors and actuators (Transducers '91)*. San Francisco.
80. Bernstein, J., S. Cho, A. T. King, A. Kourepenis, P. Maciel, and M. Weinberg. 1993. A micromachined comb-drive tuning fork rate gyroscope. *Proceedings: Micro Electro Mechanical Systems (MEMS '93)*. Fort Lauderdale, FL.
81. Ikeda, K., H. Kuwayama, T. Kobayashi, T. Watanabe, T. Nishikawa, T. Yoshida, and K. Harada. 1990. Three-dimensional micromachining of silicon pressure sensor integrating resonant strain gauge on diaphragm. *Sensors Actuators A* A23:1007–10.
82. Ikeda, K., H. Kuwayama, T. Kobayashi, T. Watanabe, T. Nishikawa, T. Yoshida, and K. Harada. 1990. Silicon pressure sensor integrates resonant strain gauge on diaphragm. *Sensors Actuators A* A21:146–50.
83. Cohn, M. B., Y. Liang, R. T. Howe, and A. P. Pisano. 1996. Wafer-to-wafer transfer of microstructures for vacuum packaging. *Technical Digest:1996 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC.
84. Böhringer, K. F., R. S. Fearing, and K. Y. Goldberg. 1998. Microassembly. *Handbook of industrial robotics.* Ed. S. Nof. New York: Wiley, pp. 1045–66.
85. Nelson, B., and B. Vikramaditya. 1997. Visually guided microassembly using optical microscopes and active vision techniques. *Proceedings: IEEE Int. Conf. on Robotics and Automation (ICRA)*. Albuquerque, NM: IEEE.
86. Feddema, J. T., and R. W. Simon. 1998. CAD-driven microassembly and visual servoing. *Proceedings: IEEE Int. Conf. on Robotics and Automation (ICRA)*. Leuven, Belgium: IEEE.
87. Cohn, M. B., and R. T. Howe. 1997. Wafer-to-wafer transfer of microstructures using breakaway tethers. *IEEE solid-state sensor and actuator workshop.* Hilton Head, SC.
88. Singh, A., D. A. Horsley, M. B. Cohn, A. P. Pisano, and R. T. Howe. 1997. Batch transfer of microstructures using flip-chip solder bump bonding. *Digest: Int. Conf. on Solid State Sensors and Actuators*. Chicago: Transducers Research Foundation.
89. Yeh, H.-J., and J. S. Smith. 1994. Fluidic self-assembly for the integration of GaAs light-emitting diodes on Si substrates. *IEEE Photonics Tech Lett* 6:706–708.
90. Yeh, H.-J., and J. S. Smith. 1994. Fluidic self-assembly of microstructures and its application to the integration of GaAs on Si. *Proceedings of the IEEE International Workshop on Micro Electro Mechanical Systems (MEMS '94)*. Oiso, Japan.
91. Yeh, H.-J., and J. S. Smith. 1994. Integration of GaAs vertical-cavity surface-emitting laser on Si by substrate removal. *Appl Phys Lett* 64:1466–68.
92. Quaid, A. E., and R. L. Hollis. 1996. Cooperative 2-dof robots for precision assembly. *Proceedings: IEEE Int. Conf. on Robotics and Automation (ICRA)*. Minneapolis, MN: IEEE.
93. Zesch, W. 1997. Multi-degree-of-freedom micropositioning using stepping principles. PhD thesis, Swiss Federal Institute of Technology, Zurich, Switzerland.
94. Keller, C. G., and R. T. Howe. 1997. Hexsil tweezers for teleoperated micro-assembly. *Proceedings: IEEE Workshop on Micro Electro Mechanical Systems (MEMS)*. Nagoya, Japan.
95. Cohn, M. B. 1992. *Self-assembly of microfabricated devices*. US Patent 5,355,577.
96. Cohn, M. B., R. T. Howe, and A. P. Pisano. 1995. Self-assembly of microsystems using non-contact electrostatic traps. *Proceedings of the ASME International Congress and Exposition, Symposium on Micromechanical Systems (IC '95)*. San Francisco.

97. Böhringer, K.-F., M. B. Cohn, K. Goldberg, R. Howe, and A. Pisano. 1997. Electrostatic self-assembly aided by ultrasonic vibration. *AVS 44th National Symposium.* San Jose, CA.
98. Böhringer, K.-F., K. Goldberg, M. B. Cohn, R. Howe, and A. Pisano. 1998. Parallel microassembly with electrostatic force fields. *Proceedings: IEEE Int. Conf. on Robotics and Automation (ICRA)*. Leuven, Belgium.
99. Edman, C. F., C. Gurtner, R. E. Formosa, J. J. Coleman, and M. J. Heller. 2000. Electric-field-directed pick-and-place assembly. *HDI* October:30–35.
100. Kim, C.-J., A. P. Pisano, and R. S. Muller. 1992. Silicon-processed overhanging microgripper. *J Microelectromech Syst* 1:31–36.
101. Pister, K. S. J., M. W. Judy, S. R. Burgett, and R. S. Fearing. 1992. Microfabricated hinges. *Sensors Actuators A* 33:249–56.
102. Keller, C., and R. T. Howe. 1995. Hexsil bimorphs for vertical actuation. *Digest: Int. Conf. on Solid-State Sensors and Actuators.* Stockholm, Sweden.
103. Hosokawa, K., I. Shimoyama, and H. Miura. 1995. Dynamics of self-assembling systems: Analogy with chemical kinetics. *Artificial Life* 1:413–27.
104. Ashkin, A., J. M. Dziedzic, and T. Yamane. 1987. Optical trapping and manipulation of single cells using infrared laser beams. *Nature* 330:769–71.
105. Srinivasan, U., and R. Howe. 1997.
106. Brown, K. T., and D. G. Flaming. 1992. Advanced micropipette techniques for cell physiology. *IBRO handbook: Methods in neurosciences*. Chichester United Kingdom: Wiley.
107. Miltenyi, S., W. Mueller, W. Weichel, and A. Radbruch. 1990. High-gradient magnetic cell separation with MACS. *Cytometry* 11:231–38.
108. Stroscio, J. A., and D. M. Eigler. 1991. Atomic and molecular manipulation with the scanning tunneling microscope. *Science* 254:1319–26.
109. Schnelle, T., R. Hagedorn, G. Fuhr, S. Fiedler, and T. Müller. 1993. Three-dimensional electric field traps for manipulation of cells—calculation and experimental verification. *Biochim Biophys Acta* 1157:127–40.
110. Coackley, W. T. 1997. Ultrasonic separations in analytical biotechnology. *TIBTECH* 5:506–11.
111. Giddings, J. C. 1993. Field-flow fractionation: analysis of macromolecular, colloidal and particulate materials. *Science* 260:1456–64.
112. Fu, A., C. Spence, A. Scherer, F. H. Arnold, and S. R. Quake. 1999. A microfabricated fluorescence-activated cell sorter. *Nat Biotech* 17:1109–11.
113. Nakao, M., K. Tsuchiya, K. Matsumoto, and Y. Hatamura. 2001. Micro handling with rotational needle-type tools under real time observation. *CIRP Annals Manufact Tech* 50:9–12.
114. Hitakawa, H. 1988. Advanced parts orientation system has wide application. *Assembly Automation* 8:147–50.
115. Yasuda, K. 2000. Non-destructive mixing, concentration, fractionation and separation of μm-sized particles in liquid by ultrasound. *Micro total analysis systems 2000*. Eds. A. van den Berg, W. Olthuis, and P. Bergveld, 343–46. Enschede, The Netherlands: Kluwer Academic Publishers.
116. Darling, R. B., J. W. Suh, and G. T. Kovacs. 1998. Ciliary microactuator array for scanning electron microscope positioning stage. *J Vac Sci Technol A* 16:1998–2002.
117. Pister, K. S. J., R. Fearing, and R. Howe. 1990. A planar air levitated electrostatic actuator system. *Proceedings: IEEE Workshop on Micro Electro Mechanical Systems (MEMS)*. Napa Valley, CA.
118. Fujita, H. 1993. Group work of microactuators. *International Advanced Robot Program Workshop on Micromachine Technologies and Systems*. Tokyo, Japan.
119. Böhringer, K.-F., B. R. Donald, R. Mihailovich, and N. C. MacDonald. 1994. A theory of manipulation and control for microfabricated actuator arrays. *Proceedings: IEEE Workshop on Micro Electro Mechanical Systems (MEMS)*. Oiso, Japan.
120. Storment, C. W., D. A. Borkholder, V. Westerlind, J. W. Suh, N. I. Maluf, and G. T. A. Kovacs. 1994. Flexible, dry-released process for aluminum electrostatic actuators. *J Microelectromech Syst* 3:90–96.
121. Suh, J. W., S. F. Glander, R. B. Darling, C. W. Storment, and G. T. A. Kovacs. 1996. Combined organic thermal and electrostatic omnidirectional ciliary microactuator array for object positioning and inspection. *Proceedings: Solid-State Sensor and Actuator Workshop*. Hilton Head Island,SC.
122. Liu, C., T. Tsao, P. Will, Y. Tai, and W. Liu. 1995. In *Digest: Int. Conf. on Solid-State Sensors and Actuators.* Stockholm, Sweden: Transducers Research Foundation.
123. Liu, W., and P. Will. 1995. Parts manipulation on an intelligent motion surface. *IEEE/RSJ Int. Workshop on Intelligent Robots, and Systems (IROS)*. Pittsburgh, PA.
124. Popa, D. O., and H. E. Stephanou. 2004. Micro and meso scale robotic as sembly. *Proceedings of WTEC Workshop: Review of US Research in Robotics*. WTEC.
125. Wheeler, A. R., K. Morishima, D. W. Arnold, A. B. Rossi, and R. N. Zare. 2000. Single organelle analysis with integrated chip electrophoresis and optical tweezers. *Micro total analysis systems 2000.* Eds. A. van den Berg, W. Olthuis, and P. Bergveld, 25–28. Enschede, The Netherlands: Kluwer Academic Publishers.
126. Ashkin, A., and J. M. Dziedzic. 1985. Observation of radiation-pressure trapping of particles by alternating light beams. *Phys Rev Lett* 54:1245–48.
127. Ashkin, A., J. M. Dziedzic, J. E. Bjorkholm, and S. Chu. 1986. Observation of a single-beam gradient force optical trap for dielectrical particles. *Optics Letters* 11:228–90.
128. Perkins, T. T., D. E. Smith, R. G. Larson, and S. Chu. 1995. Stretching of a single tethered polymer in a uniform flow. *Science* 268:83–87.
129. Ashkin, A., and J. M. Dziedzic. 1987. Optical trapping and manipulation of viruses and bacteria. *Science* 235:1517–20.
130. Svoboda, K., and S. M. Block. 1994. Biological applications of optical forces. *Ann Rev Biophys Biomol Struc* 23:247–85.
131. Fuhr, G. R., and C. Reichle. 2000. Living cells in opto-electrical cages: characterisation, manipulation and force measurements. *Micro total analysis system*. Eds. A. van den Berg, W. Olthuis, and P. Bergveld. Enschede, The Netherlands: Kluwer Academic Publishers.
132. Yando, S. 1969. Method and apparatus for fabricating an array of discrete elements. US Patent 3439416.
133. Cohn, M. B., C. J. Kim, and A. P. Pisano. 1991. Self-assembling electrical networks: an application of micromachining technology. *Digest: Int. Conf. Solid-State Sensors and Actuators.* San Francisco: Transducers Research Foundation.
134. Hosokawa, K., I. Shimoyama, and H. Miura. 1996. Two-dimensional micro-self-assembly using the surface tension of water. *Ninth Annual International Workshop on Micro Electro Mechanical Systems (MEMS '96)*. San Diego, CA.

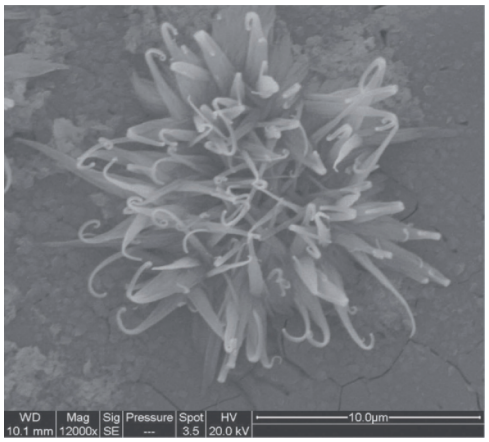
135. Edman, C. F., R. B. Swint, C. Gurtner, R. E. Formosa, S. D. Roh, K. E. Lee, P. D. Swanson, D. E. Ackley, J. J. Coleman, and M. J. Heller. 2000. Electric field directed assembly of an InGaAs LED onto silicon circuitry. *IEEE Photonics Tech Lett* 12:1198–200.
136. Braun, E., Y. Eichen, U. Sivan, and G. B. Yoseph. 1998. DNA-templated assembly and electrode attachment of a conducting silver wire. *Nature* 391:775.
137. Bashir, R. 2001. DNA nanobiostructures. *Materials Today* 4:30–9.
138. Lee, S. W., W. J. Chang, and R. Bashir, "Bottom-up" approach for implementing nano/microstructure using biological and chemical interactions. *Biotech Bioprocess Eng* 12:185–99.
139. Seeman, N. C. J. 1982. Nucleic acid junctions and lattices. *Theor Biol* 99:237–47.
140. Seeman, N. C. 1991. The use of branched DNA for nanoscale fabrication. *Nanotechnology* 2:149–59.
141. Seeman, N. C., Y. Zhang, and J. Chen. 1994. *J Vac Sci Technol A*12:1895.
142. Seeman, N. C. 1998. DNA nanotechnology: novel DNA constructions. *Annu Rev Biophys Biomol Struc* 27:225–48.
143. Winfree, E., F. Liu, L. Wenzler, and N. C. Seeman. 1998. Design and self-assembly of two-dimensional DNA crystals. *Nature* 394:539–44.
144. Mirkin, C. A., R. L. Letsinger, R. C. Mucic, and J. J. Storhoff. 1996. A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature* 382: 607–09.
145. Martin, B. R., D. J. Dermody, B. D. Reiss, M. Fang, L. A. Lyon, M. J. Natan, and T. E. Mallouk. 1999. Orthogonal self-assembly of colloidal gold-platinum nanorods. *Adv Materials* 11:1021–25.
146. Vorob'ev, A., P. Vaccaro, K. Kubota, S. Saravanan, and T. Aida. 2003. Array of micromachined components fabricated using "micro-origami" method. *Jpn J Appl Phys* 42:4024–26.
147. Kubota, K., T. Fleischmann, S. Saravanan, P. O. Vaccaro, and T. Aida. 2003. Self-Assembly of Microstage Using Micro-Origami on GaAs. *Jpn J Appl Phys* 42:4079–83.
148. Vorob'ev, A., P. O. Vaccaro, K. Kubota, T. Aida, T. Tokuda, T. Hayashi, Y. Sakano, J. Ohta, and M. Nunoshita. 2003. SiGe/Si microtubes fabricated on a silicon-on-insulator substrate. *J Phys D Appl Phys* 36:L71–L73.
149. Charlotte, P., P. Reverdy, L. Doppler, J. Bico, B. Roman, and C. N. Baroud. 2007. Capillary origami: Spontaneous wrapping of a droplet with an elastic sheet. *Phys Rev Lett* 98:156103.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

Selected Materials and Processes for MEMS and NEMS

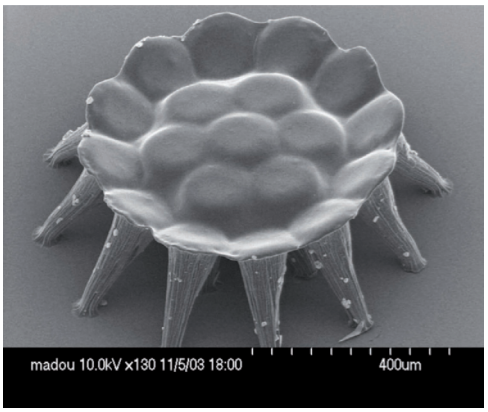
Micro glass blowing, a process developed by Andrei Shkel and others at the University of California, Irvine. The diameter of the glass spheres on a Si wafer is approximately 900 μm. (Courtesy of Dr. Andrei Shkel.)



Carbon microflowers, made in a process developed by Ashutosh Sharma and others at IIT Kanpur, India. An example of a carbon-based microelectromechanical system (C-MEMS) based on pyrolysis of resorcinol-formaldehyde gels (RF-gels). (Courtesy of Dr. Ashutosh Sharma, IIT Kanpur, India.)

Outline

Introduction
Selected MEMS and NEMS Materials
Quartz
GaAs
Diamond
SiC
Bulk and Surface Microstructuring of Ceramics
Polymer MEMS-Polyimides
Silicon Nitride
Amorphous Silicon and Hydrogenated Amorphous Silicon
Selected MEMS and NEMS Processes
Biocompatibility and Surface Modification
Surface Modifications
Questions
References



Microtable made with carbon MEMS at UCI (Courtesy of Dr. Marc Madou).

Introduction

Silicon is one of the most commonly used micromachining materials not only because of its excellent electronic and mechanical properties, the incredible degree of control one can exert over doping this semiconductor material, and the ease of insulating its surface with SiO_2 but also because it has been the dominant material in the integrated circuit (IC) industry ever since the early 1950s. However, silicon is not perfect for all applications, nor can it function on its own: the IC arsenal features a small group of additional materials, such as the metals Al, Cu, W, B, and P and the insulators SiO_2 and Si_3N_4. There are plenty of new demands that are challenging the continued dominance of silicon and will require the introduction of new materials that work in conjunction with silicon or necessitate the switch to alternative semiconductors altogether: higher speeds; better heat dissipation; merging of sensors and actuators with active electronics, optoelectronics, and quantum devices; lower power consumption; more environmentally friendly manufacturing processes, etc. In microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS), and in some application-specific ICs, alternative semiconductor materials already are being used, including gallium arsenide (GaAs), $Al_xGa_{1-x}As$ compounds, SiGe, CdSe, SiC, TiO_2, ZnO, and carbon nanotubes. A wide variety of polymers, ceramics and metals, foreign to the standard IC process also have been adopted, especially in bio-MEMS. In this chapter, we review some selected materials and processes that have affected and will continue to affect the direction MEMS and NEMS take on next. The chapter is headed by a description of possible substrate materials for MEMS and NEMS: quartz, GaAs, diamond, SiC, ceramics (with emphasis on Al_2O_3), polyimides, silicon nitride, and amorphous and hydrogenated amorphous silicon. For exciting new MEMS/NEMS techniques, we survey carbon-based MEMS (C-MEMS), bulk metallic glasses (BMGs), fast microfluidic compact disc (CD) prototyping, and micro glass blowing. Finally, we dedicate a section to biocompatibility and the various chemistries for surface derivation of solids to make them biocompatible. Studies of in vivo MEMS devices are regaining popularity, making a deeper understanding of biocompatibility a necessity.

Selected MEMS and NEMS Materials

Introduction

Miniaturization—a mantra in electrical engineering for a long time—has become a widely accepted discipline in the other engineering branches. This has led to a wider and wider array of materials used in miniaturization science. Here we dwell upon the importance of the substrate material, i.e., the principal material on which or in which to fabricate the MEMS or NEMS. In Chapter 1, this volume, we presented a section on substrate choice, including a comparison of important substrate materials for MEMS and NEMS (see Tables 1.9 and 1.10). These tables include materials that are merely "passive" substrates (e.g., plastics or cardboard support structures) and others that are "active" and may constitute an active member of a micromachine (e.g., GaAs or quartz).[1] If the substrate to be machined must combine good electronic properties with excellent mechanical behavior, single-crystal Si is still the favored material. The electronic qualifications of Si are well known through the resounding successes scored by the IC industry in this and the previous century (see Volume I, Chapter 4). But, when comparing the yield strength, density, specific strength (i.e., ratio of yield strength to density), Knoop hardness, Young's modulus, thermal conductivity, and thermal expansion coefficient for some of the most important technological materials (Volume I, Tables 4.15 and 4.16), one realizes that Si

also is an outstanding mechanical sensor/actuator substrate. Although a large percentage of MEMS and NEMS relies on single-crystal Si, and the IC industry is almost exclusively based on it, the need for all types of miniaturized devices with a wide variety of new functions is rapidly changing this picture. This is particularly true for chemical and biological sensors, actuators, harsh environment devices, and microfluidic structures. But even for mechanical devices, Si is far from the only choice, with quartz, for example, being an excellent contender. For biomedical applications and microfluidics, glass and polymers are becoming an increasingly favored miniaturization substrate. Fast ICs, radio frequency (RF)-MEMS, optoelectronics, and NEMS are pushing GaAs, $Al_xGa_{1-x}As$ compounds, and SiGe to the foreground. For harsh environments, ceramic micromachining is emerging as a new opportunity. A check-off list to determine the optimum substrate for a given micromachining task was introduced in Chapter 1, Table 1.11. Many MEMS/NEMS substrates were covered earlier in the book; some additional details on quartz, GaAs, diamond, SiC, ceramics (with emphasis on Al_2O_3), polyimides, silicon nitride, and amorphous and hydrogenated amorphous silicon, which were not treated in sufficient detail elsewhere, follow.

Quartz

Quartz Properties and Quartz Machining

Quartz is the only material known that possesses the following combination of properties:

- Piezoelectric
- Zero-temperature expansion coefficient cuts
- Stress-compensated cut exists
- Low loss [i.e., high-quality factor (Q)]
- Easy to process: low solubility in everything, under "normal" conditions, except fluoride and hot alkali etchants
- Hard but not brittle
- Abundant in nature; easy to grow in large quantities, at low cost, and with relatively high purity and perfection

Some quartz properties are summarized in Table 5.1, and in Figure 5.1, a quartz angular rate sensor (tuning fork-based), developed by Jan Söderkvist (Colibri Pro Development AB, Sweden) for automotive and other applications, is shown. The tines of the tuning fork sensor element shown are 2.5 mm long. The sensor is well suited, for example, to space applications because of its small size and ability to withstand harsh conditions.

The properties of quartz vary greatly with crystallographic direction: when a quartz sphere is etched deeply in an HF bath, the sphere takes on a triangular shape when viewed along the z-axis and a lens shape when viewed along the y-axis. The etching rate is more than 100 times faster along the fastest etching rate direction (the z-direction) than along the slowest direction (the slow-x-direction). Quartz, in some respects, presents us with an even more ideal sensor material than Si because of its nearly temperature-independent thermal expansion

TABLE 5.1 Some Important Physical Properties of Quartz

Property	Value ∥ Z	Value ⊥ Z	Temperature Dependence
Thermal conductivity [cal/(cm/s/°C)]	29×10^{-3}	16×10^{-3}	Decreases with T
Dielectric constant	4.63 (ε_1)	4.52 (ε_3)	Decreases with T
Thermal expansion coefficient (1/°C)	7.1×10^{-6}	13.2×10^{-6}	Increases with T
Electrical resistivity (Ω·cm)	0.1×10^{15} (ionic)	20×10^{15} (electronic)	Decreases with T
Piezoelectric constant, d (pC/N)	2.31 (d_{11})		
	0.73 (d_{14})		
Young's modulus (N/m^2)	9.7×10^{10}	7.6×10^{10}	Decreases with T
Density (kg/m^3)	2.66×10^3		
Maximum coupling coefficient, k	0.1		
Curie temperature (°C)	573		
Young's modulus E(GPa)	107		

∥ Z, parallel with Z; ⊥ Z, perpendicular to Z. T, temperature. See also Chapter 8, this volume, Table 8.7.

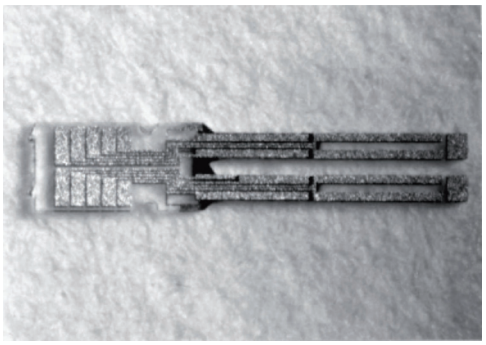


FIGURE 5.1 Quartz angular rate sensor (tuning fork) developed by Jan Söderkvist, Colibri Pro Development AB (Sweden), for automotive and other applications. The sensor is well suited to space applications because of its small size and ability to withstand harsh conditions. The tines of the tuning fork sensor element shown are 2.5 mm long.

coefficient. The thermal expansion coefficient of quartz is 7.8×10^{-6}/°C along the *z*-direction and 14.3×10^{-6}/°C perpendicular to the *z*-direction; the temperature coefficient of density therefore is -36.4×10^{-6}/°C. The temperature coefficients of the elastic constants range from -3300×10^{-6}/°C (for C_{12}) to $+164 \times 10^{-6}$/°C (for C_{66}). For the proper angles of cut, the sum of the different temperature coefficients may cancel, i.e., temperature-compensated cuts exist in quartz. Micromachined quartz is chemically and mechanically robust and may be used in high-temperature, high-shock applications where Si would fail. A variety of quartz micromachined piezoelectric devices also are on the market, such as electronic filters, resonators, and wristwatch tuning forks. Quartz exhibits very good temperature and long-term stability and therefore is ideal for high-precision oscillators. The frequency-accuracy, including all the environmental effects, after one year of aging, is of the order of few parts per million to tens of parts per million.

Prior to about 1956, the material used for quartz resonators was natural quartz, i.e., mined quartz. Natural quartz has been replaced by quartz grown in large single crystals in factories: "cultured quartz." Wafers up to 3 in. in diameter are commercially available as starting material. Although this quartz often is referred to as "synthetic quartz," nobody has yet found a way to synthesize single-crystal quartz directly from silicon and oxygen. Large quartz bars (typically ~15 cm long) of uniform size and shape are grown from small, irregularly shaped pieces of quartz. So, strictly speaking, the quartz is cultured quartz. Quartz is a common material in the earth's crust (e.g., sand is mostly quartz); however, the high-purity crystals needed for quartz growing are not as common. Most of the nutrient materials used by quartz growers are mined in Brazil and the United States.

Machining methods for quartz include mechanical machining, such as diamond saw cutting, grinding, lapping, polishing, and, relatively more recently, ultrasonic machining and wet and dry chemical etching. Lapping and polishing can be used for fashioning quartz plates as thin as 100 μm. These techniques do not lend themselves to making small, complex shapes, though. For the latter applications, ultrasonic machining has been applied to make, for example, high-precision bulk wave oscillators and complex-shaped piezoelectric resonator sensors in the millimeter size range, but the technique is not suitable for mass production. Wet chemical etching of quartz using HF/NH_4F on *z*-cut quartz can be used for submillimeter fashioning of complex quartz shapes and currently is used for mass production of wristwatch tuning fork resonators.[2] Etching rates depend strongly on the crystallographic orientation, with the *z*-cut exhibiting the highest etching rates: 1 mm/min at 50°C, with typical etch ratios in the range of 50:1 to 500:1 with respect to *x*-cut and *y*-cut plates.[3,4] Conventional photolithography is used to pattern shapes on quartz with chromium/gold for masking layers. High repeatability and automatic batch processing yield lower cost devices. Similar technologies are used for the fabrication of force sensors[5] and accelerometers. The tolerances of these quartz devices for mechanical sensing applications are, as expected from the temperature-independent thermal expansion coefficient, better than those of Si devices. Quartz also allows a greater number of different geometrical shapes than Si, and there is no need for deposition of an insulator between conductor and substrates. Dufour et al.[6] presented a balanced comparison between micromachined pressure sensors using quartz versus silicon vibrating beams and concluded that both materials are strong candidates for mass-produced micromachines. Companies such as Systron Donner[7] that have been selling traditional precision

Table 5.2 Accuracies of Various Oscillators

Oscillator Type[a]	Accuracy[b]	Typical Applications
Crystal oscillator (XO)	10^{-5} to 10^{-4}	Computer timing
Temperature-compensated crystal oscillator (TXCO)	10^{-6}	Frequency control in tactical radios
Microcomputer-compensated crystal oscillator (MCXO)	10^{-8} to 10^{-7}	Spread spectrum system clock
Oven-controlled crystal oscillator (OCXO)	10^{-8} (with 10^{-10} per g option)	Navigation system clock and frequency standard, MTI
Small atomic frequency standard (Rb, RbXO)	10^{-9}	C^3 satellite terminals, bistatic, and multistatic radar
High-peformance atomic standard (Cs)	10^{-12} to 10^{-11}	Strategic C^3, EW

[a] Sizes range from <5 cm^3 for clock oscillators to >30 L for cesium standards. Costs range from <$5 for clock oscillators to >$50,000 for cesium standards.
[b] Including environmental effects (e.g., –40°C to +75°C) and one year of aging.
MTI, moving-target indicator radar; C^3, command, control, and communications radar; EW, early-warning radar.

quartz products such as high-temperature, high-shock-resistance accelerometers are appreciating the potential competition from a Si technology and are developing silicon technology in parallel.

Hjort et al.[8] developed a most interesting, novel anisotropic dry/wet etching scheme for single-crystalline quartz. By bombarding quartz with fast heavy ions [e.g., ^{197}Au at 11.6 MeV/a and ^{129}Xe at 11.4 MeV/a specific energy (where a is the atomic mass unit)], deep nanosized amorphous cylindrical latent tracks (±10 mm wide) are formed. The bombarded wafers are then Au-Cr masked and etched in 20 M KOH at 143°C. By masking parts of the wafer prior to etching, microstructures can be generated along the ion track direction. The masked zone remains latent, and the tracks can be eliminated in an annealing step. When the ion tracks are etched in the mask windows, the latent pores are opened, and they widen until they overlap, creating the desired microfeatures. By irradiating the sample off its normal, slanted microstructures can be generated. The aspect ratio of the cylindrical nanotracks reaches a whopping 10^4, and after etching, aspect ratios larger than 100 result. Features as large as 80 µm were demonstrated in this fascinating new micromachining approach.

As mentioned earlier, quartz resonators constitute important quartz micromachining applications. From Table 5.2, accuracies of various oscillators range from 10^{-4} to 10^{-11}, and quartz can deliver even the highest accuracies required. Moreover, quartz comes with an excellent Q (see Chapter 8, this volume, Equation 8.28). Q is proportional to the decay time and is inversely proportional to the line width of the resonance peak. The higher the Q, the higher the frequency stability and accuracy capability of a resonator; i.e., high Q is a necessary but not a sufficient condition for high accuracy.*

Silicon MEMS (Si-MEMS) startup companies Discera, Silicon Clocks, and SiTime are currently targeting the quartz oscillator market. They are building packaged Si-MEMS oscillators, consisting of a MEMS resonator and the electrical circuits for the oscillator function and frequency and temperature compensation. The micromechanical resonator is a Si micromachined structure that vibrates at a specific frequency because of an external excitation (e.g., electrostatic). Today, the main technologies for oscillator clock products are quartz, ceramic and complementary metal-oxide-silicon (CMOS) silicon clocks. Si-MEMS is just barely emerging as a potential competitor. The market is still mostly served by quartz (70–80% of the value), with Japanese companies such as Kyocera, Rakon, and Epson Toyocom in control of this market. Silicon timing or CMOS clocks (i.e. non-MEMS, simply electronic) are available from, for example, Maxim or Linear Technologies and gradually are replacing

* If, e.g., $Q = 10^6$, then 10^{-10} accuracy requires the ability to determine the center of the resonance curve to 0.01% of the line width, and stability (for some averaging time) of 10^{-12} requires the ability to stay near the peak of the resonance curve to 10^{-6} of line width.

ceramic devices in applications with more relaxed specifications. Somewhere between CMOS clocks and quartz oscillators, Si-MEMS clocks now are looking for their piece of this multibillion-dollar pie (see Chapter 10, this volume).

The ~5-MHz resonant frequency resonator from SiTime is illustrated in Figure 5.2 and is made of single-crystal silicon. The resonator is ~200 µm on a side by ~10 µm thick. The trench gap is 0.4 µm and the Q is ~75 k at room temperature. The resonator is integrated into standard silicon CMOS chips, which makes the oscillator inexpensive to produce. All oscillator frequencies are derived from the same 5-MHz resonator. The oscillator is compensated by measuring the temperature with a bandgap thermometer on the same CMOS die. The temperature resolution is ~0.05°C, giving a frequency resolution of ~1.5 ppm.

An analysis of the competition between silicon and quartz for MEMS oscillators constitutes a good exercise to acquire a better insight into all of the arguments that need consideration in such a comparison (it is also another opportunity to test the decision tree presented in Chapter 1, "Decision Tree for the Optimized Micromanufacturing Option"). Si-MEMS oscillators are not new; the first publications mentioning MEMS resonators for oscillator applications appeared in the 1980s. Until recently, however, the poor temperature stability of silicon and the need for expensive ceramic or metal vacuum packaging prevented Si-MEMS oscillators from becoming a serious alternative to quartz. But, progress in processes, packaging and integration of the circuitry over the last five years has gone quite a way to address those issues.

Si-MEMS technology, in theory, offers major advantages over quartz in terms of size (especially for mobile applications), power consumption, integration cost, and ability to provide multiple operation frequencies. Let us look at each of these criteria in somewhat more detail, starting with the size argument (http://www.memsinvestorjournal.com/2006/10/mems_oscillator.html).

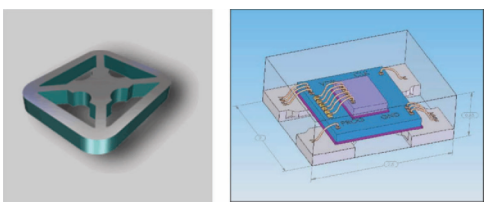


FIGURE 5.2 The ~5 MHz resonant frequency resonator from SiTime is made of single-crystal silicon.

Size Argument

Si-MEMS resonator elements have traditionally been much smaller than their quartz MEMS counterparts. This used to be the key selling argument in the mobile communication industry. A tiny MEMS oscillator often is pictured on top of a quartz oscillator that is 100 times its size to seal this argument. However, as is clear from the above discussed quartz micromachining efforts, the latest quartz MEMS products are very small as well. A new generation of quartz MEMS is emerging, with dimensions similar to those of the first Si-MEMS oscillators (e.g., $2 \times 2.5 \times 0.8$ mm for a quartz device from NDK compared with $2 \times 2.5 \times 0.85$ mm for the Si-MEMS structure from SiTime). The first generation Si-MEMS oscillators are pin-to-pin compatible with quartz products, allowing an easy one-on-one replacement. Since the size of the package will be the same, the main advantage of Si-MEMS oscillators in the "small oscillator segment" will not be size but price. The manufacturing and encapsulating of small quartz devices is expensive compared with Si-MEMS (see next). Printed circuit board (PCB) space also can be saved with Si-MEMS oscillators, which do not require additional bypass capacitors and resistors.

Low Cost Argument

Si-MEMS resonators can be manufactured at a lower cost than quartz MEMS oscillators since Si-MEMS leverages IC batch manufacturing techniques much better. The first commercial Si-MEMS oscillators are hybrid parts, though; e.g., SiTime's Si-MEMS resonators are wire bonded to an ASIC. SiTime uses an innovative wafer-scale packaging solution that is licensed from Bosch and is compatible with CMOS processing. It involves placing resonators on top of an SOI wafer and encapsulating them with an epitaxial layer of silicon, facilitating sealing at high temperatures and creating a barrier to moisture (see Chapter 4, this volume, on packaging, assembly, and self-assembly). The Si resonator is inexpensive

in volume and is estimated to be less than $0.10. The typical price of a Si-MEMS oscillator will start at $0.50 and may reach a few dollars in some instances.

Multifrequency Argument

Several RF-MEMS resonators with different frequencies can be manufactured on the same die, allowing for multifrequency devices while retaining small size and low cost. Crystals, on the other hand, are single-frequency devices. If multiple frequencies are needed such as in a cell phone, e.g., a 32 kHz oscillator for the real-time clock and several MHz oscillators for transmission, reception, and processing functions, several discrete-packaged crystals must be used.

High Q Argument

Si-MEMS devices exhibit a higher *Q* than CMOS clocks. However, Si-MEMS has not yet reached the quality factor of quartz (100,000–200,000 for quartz compared with 75,000 for Si-MEMS). The related phase noise of MEMS also is an issue that must be overcome to reach the performance required for temperature-compensated crystal oscillators (TCXOs).

Integration Argument

Contrary to the case with quartz, Si-based RF-MEMS oscillators are manufactured with CMOS-compatible processes. This could pave the way to true single-chip solution with built-in frequency references. Advantages in terms of power consumption (lower parasitics from the wires) and size are obvious. However, as stated above, the first commercially available Si-MEMS oscillators are actually hybrid parts.

Low Power Argument

The low power consumption of MEMS resonators is a key advantage in portable applications. However, MEMS solutions only reduce power consumption when monolithically integrated, because all of the parasitics associated with an off-chip component (i.e. wire bonding quartz or discrete MEMS) are eliminated. This advantage also is diminished in a hybrid part because of wire-bond parasitics.

Summary

Of the human-grown single crystals, quartz, at ~3000 tons per year, is second only to silicon grown in quantity (three to four times as much Si was grown annually as of 1997). Despite the mechanical/thermal advantages of quartz, quartz micromachining lags considerably behind that of Si. This is mainly due to the lack of plasma processes to structure quartz with the versatility achieved in the case of Si. Another reason that quartz is not competing better with Si as a mechanical MEMS substrate is the lack of large quartz substrates and its cost: a quartz wafer, 76.2 mm in diameter, 0.45 mm thick, and polished on one side costs about $200, while the average selling prices for 200- and 300-mm silicon wafers are about $45 and $200, respectively.

GaAs

GaAs Properties

The first gallium arsenide (GaAs) metal semiconductor field effect transistor (MESFET) was proposed in 1966 and the first GaAs high-electron mobility transistor (HEMT) was demonstrated in 1978 at the Bell Laboratories. However, GaAs is only now rising from its position as the technology that remains the technology of the future to true technological wonder. The main economic driver for widespread use of GaAs semiconductors was the emergence of the digital cellular phone, which needs faster IC technology than Si can provide, and ever since 1994, the GaAs industry has grown at least 25% per year. As the micromachining industry largely evolves in lockstep with the IC industry, we can expect GaAs micromachines to succeed as GaAs use penetrates the IC industry further. The growing importance of quantum mechanical devices also is pushing GaAs toward becoming a more important MEMS/NEMS substrate contender. As we pointed out in Volume I, Chapter 3, an important feature of the III-V technologies in general, and of GaAs in particular, is the possibility of forming compatible ternary and quaternary compounds. Using GaAs as the substrate material, the formation of $Al_xGa_{1-x}As$ compounds is especially attractive, since their lattice constants are nearly equal to those of GaAs, and

aluminum and gallium atoms are easily substituted in the lattice without causing too much strain. Today, epitaxy techniques have matured, and both high-quality MOCVD and molecular beam epitaxy (MBE) epitaxy are commercially available. Layers deposited with these techniques have very little induced strain, which allows them to be grown almost arbitrarily thick. In Volume I, Chapter 3, we describe GaAs-based quantum wells; Volume I, Chapter 5 covers double-barrier resonant tunneling diodes and quantum cascade lasers, all based on layers of gallium aluminum arsenide sandwiched between gallium arsenide layers. With its direct bandgap, GaAs can be used to emit light efficiently, whereas silicon has an indirect bandgap and is very poor at emitting light (see Volume I, Chapter 3). Nonetheless, recent advances in nanotechnology, as we described in Volume I, Chapters 3 and 5, may make silicon light-emitting diodes and lasers possible after all. Another important application of GaAs (and III-V compounds in general) is for high-efficiency solar cells. In 1970, Zhores Alferov (see Volume I, Figure 3.60) and his team in the USSR created the first GaAs heterostructure solar cells. Today, the combination of GaAs with germanium and indium gallium phosphide is the basis of a triple-junction solar cell that holds the record for solar light conversion efficiency of over 32% and can operate also with light as concentrated as 2000 suns.* This type of solar cell powers the robots *Spirit* and *Opportunity*, which are exploring the Mars surface.

In Chapter 1, this volume, Table 1.10, we compare some of the material properties of single-crystal SiC, Si, GaAs, and diamond with relevance to electronics and MEMS/NEMS. Gallium arsenide, it emerges, is a very promising MEMS/NEMS substrate for a variety of reasons. These include well-known optoelectronic properties such as a direct bandgap transition (see above) and high-mobility electrons. The higher saturated electron velocity and higher electron mobility of GaAs compared with those of Si allow it to function at frequencies in excess of 250 GHz. GaAs HEMTs are a typical high-electron mobility application, and the same property also renders the material ideal for the measurement of magnetic fields through the Hall effect (see Volume I, Equation 3.64). Also, GaAs devices generate less noise than silicon devices when operated at those high frequencies. Because of its high switching speed, GaAs, it would seem, also is more ideal for computer applications than Si. In the 1980s it often was suggested that the IC market might switch from Si to GaAs, which it never really did (from there its unfortunate designation as "the technology that remains the technology of the future"). Cray even built a GaAs-based computer in the early 1990s, the Cray-3, but the company filed for bankruptcy in 1995. Because it has a larger bandgap than Si, GaAs is considered to be a better material for high-temperature electronics. GaAs devices can be operated at higher power levels than equivalent silicon devices because GaAs has a higher breakdown voltage. Pressure, temperature, and vibration sensors, making use of the influence of external pressure and temperature on the bandgap, have already been built.

GaAs has piezoelectric properties comparable to those of quartz, and various physical effects give higher piezoresistive values than those of silicon. The piezoelectric effect enables GaAs-based piezoelectric transducers, and one may excite any vibration mode used in quartz if the proper crystal orientation is observed. Overall, the material is less attractive for purely mechanical MEMS devices, though, since the yield load is smaller by a factor of two compared with silicon.[9] We summarize some of the most important GaAs properties in Table 5.3.

GaAs Micromachining

Besides wet[9] and dry[10] bulk micromachining of GaAs, some GaAs surface micromachining work has been reported.[11] Because of the zinc blend structure of GaAs (same as Si), the wet etching of {111} crystal planes proceeds more slowly than all others. In wet bulk micromachining of GaAs, a unique triangular prism-shaped bridge can be obtained through anisotropic etching by appropriate placement of the etch mask (see Figure 5.3). This unique profile, not possible in silicon, is formed because of the

* Where one sun equals 1000 W/m² at the point of arrival. One sun is a unit of power flux.

TABLE 5.3 GaAs Properties

Property	GaAs
Atomic or molecular weight	144.63
Atoms or molecules/cm^3	4.42×10^{22}
Breakdown field, V/cm>	$\sim 4 \times 10^5$
Chemical resistance	Poor
Crystal structure	Zinc blend
De Broglie electron wavelength (Å)	240
Debye temperature	360
Density, g/cm^3	5.32
Dielectric constant (high frequency)	10.89
Dielectric constant (static)	12.9
Effective density of states: valence band N_v (cm^{-3})	7.0×10^{18}
Effective density of states: conduction band N_c (cm^{-3})	4.7×10^{17}
Effective mass: electrons	0.063 m_0
Effective mass: holes	0.051 m_0
Electron affinity (eV)	4.07
Energy gap (eV)	1.42 direct
Energy gap (eV)	1.43
Intrinsic carrier concentration, n_i (cm^3)	1.8×10^6
Intrinsic mobility: Electron, cm^2/V-s	8600
Intrinsic mobility: Hole, cm^2/V-s	400
Knoop hardness (kg/mm^2)	600
Lattice constant, nm	0.5653
Maximum operating temperature (°C)	460
Melting point (°C)	1238
Number of atoms in 1 cm^3	4.42×10^{22}
Optical phonon energy	0.035
Physical stability	Fair; sublimation of arsenic is a problem
Saturated electron drift velocity ($\times 10^7$ cm/s)	2
Symmetry group	T_d^2– F43 m
Temperature coefficient of expansion (°C^{-1})	6.8×10^{-6}
Thermal conductivity (W/cm-°C)	0.46
Yield strength (GPa)	2.0
Young's modulus (GPa)	75

See also http://www.ioffe.rssi.ru/SVA/NSM/Semicond/GaAs/index.html.

crystallographic polarity in some axial directions, and the differences between the etch rates of the A{111} planes and the other low-index planes.[12]

Although slower etching planes exist, GaAs does not present real etch stop planes, as observed in silicon when using, for example, KOH-based solutions (see Volume II, Chapter 4). In the latter etchants, the {111} Si planes present etch rates of up to 400 times lower than the other planes. Surface micromachined sensor and actuator structures made of metal on a GaAs substrate have been demonstrated in schemes compatible with normal GaAs IC processing. The most frequently used sacrificial wet and dry etch systems for III-V crystals are presented in the review papers of K. Hjort[14] and S. D. Collins[15] and in the work by Karam et al.[16] It was shown that to craft surface micromachined structures in epitaxial GaAs, $Al_xGa_{1-x}As$ ($x \geq 0.5$) may act as a sacrificial

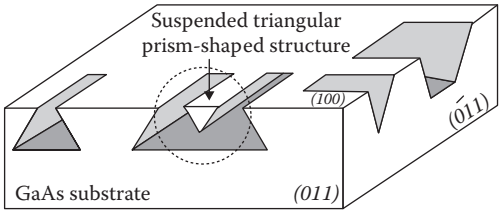


FIGURE 5.3 Free-standing triangular bridge in GaAs.[13]

layer. By using several steps of MOCVD epitaxial layer regrowth, structures of polysilicon-like complexity were built.[17,18] Note that not only different material compositions but also different dopant concentrations and damaged regions of GaAs can act as stop and sacrificial layers.

In Volume II, Chapter 3, we detail the GaAs single-crystal reactive etching and metallization process, which includes chemically assisted ion beam etching (CAIBE) and reactive ion etching (RIE) for vertical and undercut dry etches, respectively. Suspended and movable GaAs structures are produced with a 25:1 aspect ratio of vertical depth (10 μm) to lateral width (400 nm) (Volume II, Figure 3.43).

Ribas from TIMA developed a front-side bulk GaAs micromachining approach compatible with standard GaAs microelectronics technologies for the fabrication of integrated, low-cost, high-volume microsystem applications (http://cmp.imag.fr/products/mems/thesis.pdf).[13] Dr. Ribas also developed a set of computer-aided design tools for GaAs MEMS layout level design, such as cross-section and three-dimensional layout viewers, layout generators, and a bulk GaAs etching simulator within the Mentor Graphics environment. In the same context of GaAs IC compatibility, Karam et al.[16] are investigating gallium arsenide micromachining techniques using high-electron mobility transfer (HEMT) and metal Schottky field effect transistor (MESFET) foundry processes.

In an interesting example of self-assembly, two-dimensional arrays of self-positioning, free-standing GaAs structures were fabricated using a new method of surface micromachining called *micro-origami*. This technique is based on epitaxial growth of III-V multilayers on GaAs substrates followed by photolithography and wet etching, as we described in Chapter 4, this volume, on packaging, assembly, and self-assembly.

Progress in processes to deposit GaAs epitaxially on Si could benefit both IC and micromachining industries. GaAs grown epitaxially on Si substrates has great potential to lower the costs, improve the manufacturability, and improve the thermal characteristics of GaAs-based ICs. GaAs/Si wafers are larger and stronger and lighter than GaAs wafers.[19] This combination opens the possibility of monolithic integration of GaAs-based ICs with Si-based ICs. GaAs-on-Si has significant market potential, both as a substitution technology for manufacturing traditional GaAs devices and as a new technology for monolithic integration of GaAs devices and silicon integrated circuits. As a substitution technology, GaAs-on-Si holds the potential to fundamentally change the economics of manufacturing GaAs integrated circuits and GaAs optical components. As a technology for monolithic integration of GaAs devices and silicon integrated circuits, GaAs-on-Si technology is even more interesting. In fiber optics, integration of GaAs lasers and photodetectors with silicon control circuitry would become feasible. Because of these obvious advantages, many attempts have been made over the past thirty years to grow GaAs on Si substrates. The GaAs lattice constant (5.65 Å) is about 4% larger than silicon's (5.43 Å), and this difference makes it difficult to grow high-quality GaAs directly on silicon. Instead, Motorola researchers place a thin layer (about 50 Å) of strontium titanate ($SrTiO_3$) between the two semiconductors.[20] Strontium titanate's lattice constant falls roughly halfway between GaAs and Si. Moreover, an amorphous layer (10–20 Å thick) forms between Si and $SrTiO_3$. This layer, it is hypothesized, absorbs the lattice mismatch strain between the two materials, allowing the crystalline $SrTiO_3$ to form a normal lattice without distortion from the underlying Si. Because the $SrTiO_3$ layer is very thin, the amorphous interlayer also absorbs the mismatch strain between the $SrTiO_3$ and GaAs layers.

In 2001, scientists from Motorola reported on supposedly device-quality GaAs epitaxial layers with low dislocation densities, close to that of bulk GaAs crystals (~$10^4/cm^2$), grown onto Si by molecular beam epitaxy (MBE) (Figure 5.4).[21] In spite of the initial worldwide publicity, Motorola decided to stop its activity in this area in 2003.[22] The presence of a low dislocation density in the epitaxial GaAs layers has never been independently confirmed and remains controversial. So it is back to the drawing board. Another approach to integrating GaAs on a Si substrate was pursued by Yeh and Smith, who assembled GaAs laser diodes in micromachined wells in a Si substrate by trapping them as discussed in Chapter 4, this volume (see Figure 4.44).[23,24]

FIGURE 5.4 Motorola Labs created the world's first 8-in. GaAs on silicon wafer and worked with epitaxial wafer manufacturer IQE to create the world's first 12-in. GaAs on silicon wafer.

In the micromachining world, possible applications of GaAs MEMS are optical shutters or choppers, actuators in monolithic microwave integrated circuits (MMIC), sensors using piezo- or optoelectrical properties of GaAs, or applications favoring integration of micromechanical devices with electronic circuitry for fast signal processing, high operating temperature, or high radiation tolerance. GaAs micromachining might constitute a good method by which to incorporate micromirrors in resonant optical cavities for tunable lasers. Also, micromachining using gallium arsenide and group III-V compound semiconductors is a practical way to integrate RF switches, antennas, and other custom high-frequency components with ultrahigh-speed electronic devices for wireless communications (see Chapter 10, this volume).

Summary

Overall, it remains hard for GaAs to compete with Si as a semiconductor substrate, but GaAs did start making major commercial inroads for high-frequency applications with the introduction of mobile phones. Silicon presents at least three major advantages over GaAs, explaining why it is so hard for GaAs to compete. To begin, silicon is abundant and cheap to process. Silicon's greater physical strength allows working with larger wafers (maximum of ~300 mm compared with ~150 mm diameter for GaAs). Silicon is abundantly available in the Earth's crust in the form of all types of silicate minerals, and the economy of scale available to the silicon IC industry discourages a more rapid adaptation of GaAs. The second major advantage of Si over GaAs is the existence of silicon dioxide—one of the best and most stable insulators that may be created with a very low defect concentration at interface Si/SiO_2. Silicon dioxide can easily be incorporated onto silicon circuits, and such layers adhere very well to the underlying Si. GaAs does not form such a stable adherent insulating layer (its native oxides are typically leaky). The third, perhaps most important advantage of silicon over GaAs is that it possesses a much higher hole mobility. This high hole mobility allows for the fabrication of higher-speed p-channel field effect transistors required for CMOS logic. Because they lack a fast CMOS structure, GaAs logic circuits have much higher power consumption, which has made them unable to compete with silicon logic circuits.

As suggested before, the micromachining industry largely evolves with and piggybacks on the IC industry, so we can expect GaAs micromachines to succeed as GaAs use penetrates the IC industry further. The most effective approach to the cost reduction of any device is to increase the diameter of GaAs wafers. Along this line, Anadigics Inc., which manufactures RF ICs for the wireless handset and broadband communications markets, is building a 6-in. GaAs fab in China, a certain sign that GaAs has arrived (cost per 6-in. GaAs wafer is projected at $300).

The SiGe devices market has grown more rapidly than some anticipated, and III-V producers have some reason to be concerned about the transition from Si to SiGe instead of using GaAs in a number of key applications. With the talk of SiGe (see Volume I, Figure 4.59) matching performance with GaAs and beating it on cost, GaAs still edges out SiGe in terms of efficiency, so for those applications where this is less important—e.g., the shorter-range wireless networks, such as WiFi and Bluetooth—SiGe is going head to head with RF-CMOS.

Diamond

Introduction

Diamond, the hardest of all natural materials, is a wide bandgap semiconductor material with

excellent thermal, tribological, and mechanical properties. The thermal conductivity of diamond is higher than that of any other material near room temperature, making it an ideal material for heat sinking. Diamond components may be as much as 10,000 times more wear-resistant than those made from silicon and come with a much higher thermal and chemical stability. Diamond is also a biocompatible material, so that it could be used in the body as a drug-dispensing unit without initiating an allergic reaction. This is because carbon (the elemental ingredient of diamond) is chemically benign (see also below, under C-MEMS). If it could be micromachined easily, the exceptional physical and chemical properties of diamond would promise to expand the range of applications for microdevices. It is an ideal candidate for MEMS applications, particularly in harsh environments. In addition, diamond reduces stiction compared with silicon (diamond's low coefficient of friction is comparable to that of Teflon), perhaps making diamond-surface micromachined devices easier to manufacture. Diamond can be doped to change it from an insulator to a semiconductor: p-doping works fine, but it remains difficult to make an n-type diamond. We compare the properties of diamond with those of silicon, GaAs, and SiC in Table 5.4.

Diamond Deposition and Diamond Micromachining

Chemical vapor deposition (CVD) of single-crystalline diamond has been the subject of a great deal of research since the early 1980s. Very high-quality diamond thin films with negligible levels of defects and impurities (even superior to natural diamonds and high-pressure, high-temperature diamonds) are achievable when using microwave plasma chemical vapor deposition. However, high- quality homoepitaxial single-crystal CVD diamonds are still very expensive and are available only on small areas (3×3 or 4×4 cm^2) because of the need of single-crystalline diamond substrates.[25,26]

Alternatively, polycrystalline diamond film, deposited by CVD, is one potentially viable approach

Table 5.4 Properties of Diamond and Other Semiconductors at 293 K

	Diamond	Silicon	GaAs	SiC
Lattice constant (Å)	3.567	5.431	5.653	a = 3.086; c = 15.117
Hardness (kgf/mm^2)	7,000–10,000	1,000		1,875–3,980
Density (g/cm^3)	3.52	2.30	5.32	3.10
Refractive index	2.4195	3.448	3.4	2.65–2.69
Thermal conductivity (W/cm·K)	20	1.41	0.455	4
Specific heat (J/g·K)	0.52	0.70	0.35	0.65
Optical phonon energy (meV)	163	63	35	100
Longitudinal phonon velocity (cm/s)	1.8×10^6	9×10^5	5.2×10^5	1.3×10^6
Coefficient of thermal expansion	0.8×10^{-6}	2.5×10^{-6}	5.9×10^{-6}	2.9×10^{-6}
Debye temperature (K)	1,860	645	344	1,200
Electron saturated velocity (cm/s)	1.5×10^7	1×10^7	0.8×10^7	2×10^7
Hole saturated velocity (cm/s)	1.05×10^7	9×10^6	1×10^7	1×10^7
Electron mobility (cm^2/V·s)	2,000 (500)	1,420	8,800	600
Hole mobility (cm^2/V·s)	1,500 (500)	470	400	650
Effective mass				
Electron	0.57	1.1	0.068	0.45
Hole	1.2	0.8	0.5	1.0
Dielectric constant	5.7	11.9	12.5	10
Intrinsic resistivity (Ω·cm)	$>10^{15}$	10^5	10^8	$>10^{15}$
Breakdown field (V/cm)	$1–20 \times 10^6$	3×10^5	3.5×10^5	$1–5 \times 10^6$
Bandgap (eV)	5.47	1.12	1.42	2.2
Effective density of states				
Conduction band (cm^{-1})	1×10^{19}	2.8×10^{19}	4.7×10^{17}	7×10^{18}
Valence band (cm^{-1})	3×10^{19}	1.04×10^{19}	7×10^{18}	2.5×10^{19}

to high-temperature, harsh-environment MEMS[27] (see also http://www.nasatech.com/Briefs/Feb99/NPO20529.html). Chemical vapor deposition of diamond on Si substrates is compatible with the integration of Si-based electronics. The most common polycrystalline diamond CVD process involves a flowing mixture of methane and hydrogen, typically at a total pressure of 45 Torr (6 kPa) and a substrate temperature of 800–950°C. Diamond heteroepitaxially deposited on Si is polycrystalline (Figure 5.5) with a large number of grain boundaries, a high intrinsic stress, and a rough surface, making it difficult to use in MEMS. The grain size of typical diamonds grown from a methane and hydrogen gas source by microwave plasma CVD on Si is around 30 nm to 10 µm.[28] Problems with polycrystalline diamond films also include difficulties in growing good-quality diamond films and making reliable ohmic contacts to the material, as well as poor reproducibility.[29] This leaves us with ultrananocrystalline diamond (2–5 nm grain size)[30] and amorphous diamond[31] as the two most suitable diamond candidates for MEMS applications.

At Argonne National Laboratories, "ultrananocrystalline" diamond films have been deposited by a novel CVD process that may overcome some of the problems mentioned above (http://www.anl.gov/techtransfer/Available_Technologies/Chemistry/uncd_flc.html). Argonne's patented CVD method was developed at first using fullerenes as the carbon source. Fullerene powder was vaporized and introduced into an argon plasma, causing the fullerenes to fragment into two-atom carbon molecules (dimers). Silicon or other substrate materials are "primed" with fine diamond powder. When the carbon dimers settle out of the plasma onto the substrate, they arrange into a film of small diamond crystals about 3–5 nm in diameter. Subsequent work showed that the same result can be achieved by introducing methane into an argon plasma as long as little or no additional hydrogen is present. The Argonne method differs in two major respects from other CVD methods for making diamond film. First, the molecular building block is different (carbon dimers rather than methyl radicals), and second, little or no hydrogen is present in the plasma (~1%), whereas in other methods the plasma contains 97–99.5% hydrogen along with methane. The ultrananocrystalline films produced by Argonne's method are completely free of intergranular flaws and nondiamond secondary phases that degrade the properties of conventionally produced diamond films, have crystals about 50–200 times smaller than those in other films, and are 10–20 times smoother than conventional high-purity CVD diamond films.

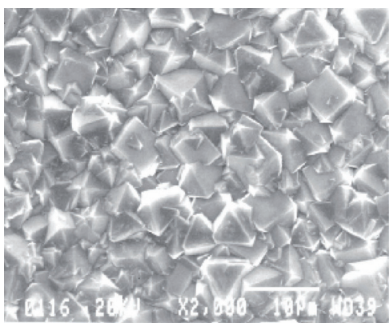


FIGURE 5.5 A typical SEM image of a CVD diamond surface on a Si substrate. (Courtesy of Dr. Chunlei Wang, Florida International University.)

For surface micromachining, a "starter" (nucleation) layer of nanocrystalline diamond powder on a silicon substrate is patterned, and the diamond film forms only in the areas where the nucleation layer remains. Flat, free-standing diamond films can also be grown on silicon dioxide. The diamond layer may be released from the substrate by etching away the silicon dioxide. The result is free-standing diamond structures as little as 300 nm thick with features as small as 100 nm and friction coefficients as low as 0.01. Argonne is also developing deposition methods that permit integrated fabrication of complete devices, such as micrometer-sized pinwheels, gears, turbines, and micromotors, without manual assembly. Also, hollow, three-dimensional (3D) structures can be fabricated in diamond surface micromachining. In this case, silicon is used as the sacrificial material (e.g., a disc or pin) for the desired diamond component. The diamond film is grown on the Si structure and conforms to its shape. The silicon is etched away, leaving a diamond cylinder or tube. It has been shown that this method is capable of making convex hollow structures (e.g., a hollow pyramid)—shapes that cannot be produced by conventional silicon lithographic methods.

Sandia researchers have explored amorphous diamond for micromachining. Amorphous diamond is the second hardest substance known, after crystalline diamond (http://www.sandia.gov/media/NewsRel/NR2000/diamond.htm). They use pulsed laser deposition, and a simple proof-of-principle device—a diamond interdigitated comb—takes about 3 h to fabricate. Subsequent annealing of the diamond film device down to zero stress (to prevent warpage) takes a few minutes. Before this work, amorphous diamond itself had been impractical because its tremendous internal stresses—hundreds of atmospheres—made it impossible for the material to stand alone or to coat thickly on any but the strongest substrates.

Diamond-based MEMS, such as cantilevers, membranes, and microturbines, can be fabricated using two strategies: lithographic patterning and selective deposition.[31,32] In the selective deposition method, diamond microstructures are fabricated by selective diamond growth on Si using a patterned SiO_2 mask. With this selective growth approach, piezoresistors and thermal resistor structures have been patterned.[32]

In lithographic diamond patterning, diamond films are synthesized on an appropriate substrate such as Si. This is followed by diamond surface micromachining processes similar to the ones used in the fabrication of poly-Si-MEMS. Diamond layers can be masked by photoresist or metals, followed by dry etching (RIE) in oxygen-containing plasmas.

Diamond films also can be directly patterned using focused ion beam (FIB) milling without the use of a mask.[33] Figure 5.6 shows CVD diamond microscale p-i-p electronic devices fabricated on homoepitaxially grown single-crystalline layers by FIB etching; the latter devices were used to investigate carrier transport properties of diamond at high electric field strengths (more than10^7 V/cm).[33]

A number of other diamond-based devices have been investigated; these include basic electronics elements, sensors, and actuators.[34–36] Although their cost, performance, and processibility are still not compatible with Si-based MEMS, diamond as a complementary MEMS material remains very promising, especially for use in extreme environments.

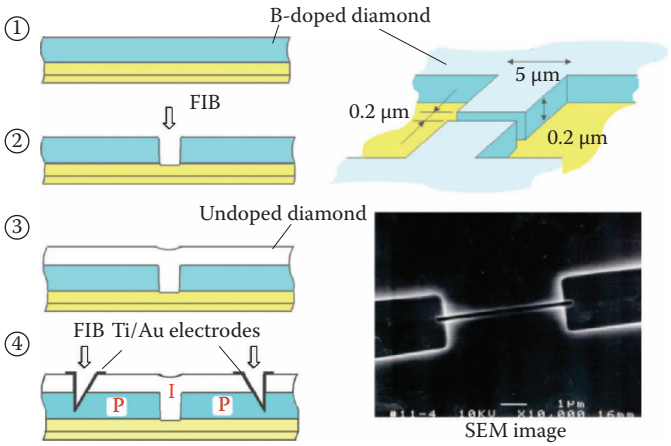


FIGURE 5.6 Schematic fabrication processes for a diamond p-i-p structure consisting of a narrow *intrinsic* (undoped) region sandwiched between two p-type (boron-doped) regions. Top view schematic drawing and a scanning electron microscope picture show narrow (0.2-μm) grooves formed in the 0.2-μm-thick p layer (at step 2). These grooves are filled with i diamond (at step 3).[33]

SiC

Like diamond, silicon carbide (SiC) is well known for its mechanical hardness, chemical inertness, high thermal conductivity, and electrical stability at temperatures well above 300°C, making it another excellent candidate for high-temperature MEMS. Both exhibit piezoresistive properties. In comparison with diamond, other SiC features are that it can be doped both p-type and n-type fairly easily and that it allows a natural oxide to be grown on its surface, which make SiC an attractive candidate for alternative semiconductor MEMS material. Because of its large bandgap and good carrier mobility, it can be employed in high-temperature and high-power applications. In addition, the fabrication technology for SiC active electronic devices is based mostly on processes established in the Si microelectronics industry. One property that makes SiC films particularly attractive for micromachining is that these films easily can be patterned by dry etching using Al masks. Patterned SiC films actually can be used as passivation layers in the micromachining of the underlying Si substrate (SiC can withstand both KOH and HF etching).[37]

SiC crystallizes in many different polytypes, which differ from one another in the stacking sequence of a repeat unit consisting of two planes

of close-packed Si and C atoms. The two most common SiC polytypes are 3C-SiC and 6H-SiC. The 3C polytype, also known as β-SiC, is the only polytype with a cubic structure. 3C-SiC crystallizes in a ZnS-type structure; hence, it can be deposited on Si. Historically, most research has focused on developing 6H-SiC as a semiconductor material for high-temperature and high-power electronics. However, the small wafer size (<2 in.) and the inability to grow epitaxial layers on substrates other than 6H-SiC have hindered the development of the 6H polytype. Over the last seven years we have witnessed a growing interest in 3C-SiC as a MEMS material. As we saw above, unlike 6H-SiC, the 3C-SiC polytype can be epitaxially grown on single-crystal silicon substrates. Large-area substrates enable low-cost batch processing, essential in making SiC MEMS devices viable for mass production applications (e.g., automotive). It was at Case Western Reserve University (CWRU) that the first successful depositions of spatially uniform, single-crystal, 3C-SiC films on 4-in. (100) silicon wafers were made. These epitaxial 3C-SiC films on Si were used to craft structures such as diaphragms and cantilever beams. Since SiC films are highly resistant to KOH and EDP etching, dry etching and Al masks are employed for patterning. Because these diamond films must be grown directly on single-crystal silicon, surface micromachining with a sacrificial layer was not possible at first. The solution was again invented at CWRU and involves growing polycrystalline cubic silicon carbide (poly-SiC) films of about 2 μm thickness by an APCVD process on 4-in. polysilicon-coated (100) silicon wafers.[38,39] Surface micromachining is achieved by using the underlying polysilicon film as the sacrificial layer. The poly-SiC deposition process is carried out at 1280°C in a cold-wall, RF-induction heated, vertical APCVD reactor. After deposition, the SiC film is polished to reduce the surface roughness (Ra) from ~400 Å in the as-grown film to <40 Å. The poly-SiC lateral resonant device shown in Figure 5.7 was made by this surface micromachining approach, and the finished device exhibits Qs as high as 215,000 at pressures below 10^{-5} Torr (see also "Quality Factor" in Chapter 8, this volume, on actuators). Resonant frequency drifts of less than 18 ppm/h have been observed, and device operation at elevated temperatures as high as 950°C

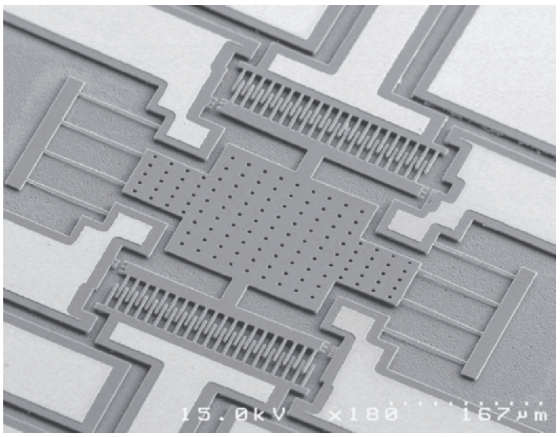


FIGURE 5.7 Scanning electron microscope (SEM) micrograph of a released poly-SiC lateral resonant device. The suspension beam length and width are nominally 100 and 2.5 μm, respectively. Exposed poly-SiC shows up as dark gray, while Ni metallization appears light gray. (Courtesy of Dr. M. Mehregany, Case Western Reserve University.)

has been achieved. The poly-SiC surface micromachining steps are outlined in Figure 5.8.

In Chapter 1, this volume, Table 1.10, we compared some of the material properties of single-crystal SiC with Si, GaAs, and diamond with relevance to electronics and MEMS/NEMS. For an update on this

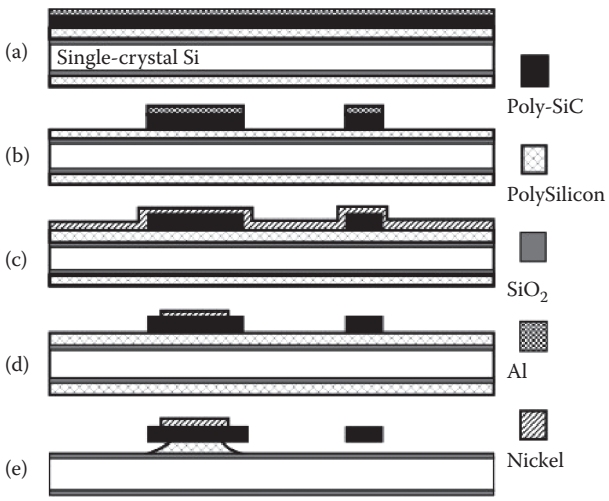


FIGURE 5.8 Schematic description of the poly-SiC resonator fabrication process showing cross-sections after a 5000-Å-thick layer of aluminum is sputter deposited over the polished poly-SiC (a); the Al layer is patterned using lithography and anisotropic CHF_3/O_2/He plasma etching to transfer the Al pattern into the poly-SiC (b); the Al layer is stripped and a 7500-Å-thick layer of Ni is deposited by sputtering (c); the Ni film is patterned for contacts using photolithography and wet etching (d); and poly-SiC is released in 40 wt% KOH at 40°C (e).

table for SiC properties, consult http://www.ioffe.rssi.ru/SVA/NSM/Semicond/SiC/index.html.

Further advances in SiC deposition and patterning techniques have led to the development of single and multilevel implementations similar to the polysilicon MUMPs (multiuser MEMS process). The multiuser silicon carbide (MUSiC) is an eight-mask, four polycrystalline SiC (poly-SiC) layer surface micromachining process.

The State of Ohio has funded high-temperature MEMS efforts both at Case Western Reserve University (http://mems.case.edu) and at NASA Glenn (http://www.grc.nasa.gov/WWW/SiC/SiC.html) and made these two centers leaders in the field.

Bulk and Surface Microstructuring of Ceramics

Ceramics encompass a vast array of materials, and a concise definition of ceramics is almost impossible. A workable definition of ceramics is that of a class of refractory, inorganic, and nonmetallic materials. This class of materials can be divided into traditional and advanced ceramics. Traditional ceramics include clay products, silicate glass, and cement, whereas advanced ceramics consist of carbides (SiC; see above), pure oxides (quartz; see above: Al_2O_3, etc.), nitrides (Si_3N_4), nonsilicate glasses, and many others. Ceramics offer many advantages compared with other materials. They are harder and stiffer than steel, more heat and corrosion resistant than metals or polymers, and less dense than most metals and their alloys, and their raw materials are both plentiful and inexpensive. Ceramic materials display a wide range of properties, which facilitates their use in many different product areas and ceramic MEMS, and NEMS represents another promising development.

The inherent chemical resistance of refractory materials makes their anisotropic micromachining very difficult and expensive.[40,41] In Volume II, Chapter 9, we promoted the use of plasma spray deposition for making oxygen-sensitive ZrO_2 membranes in batch mode (see Volume II, Figure 9.24). Another potential solution for working with these poorly machinable materials is by casting or by gas phase deposition into molds. MIT's micromachined gas turbines, for example, were made of thick (10–200 μm) SiC or silicon nitride, deposited by CVD into Si molds (http://web.mit.edu/aeroastro/faculty/index.html). Sol-gel-type solidification of ceramic liquid precursors in molds, as covered in Volume II, Chapter 8, is another available option—less expensive but yielding low- to medium-resolution microstructures only. Deep UV resists may be used as alternative, sacrificial molds. In most cases a high-temperature sintering step will significantly shrink the dimensions of the high-temperature part, reducing achievable accuracy. In Volume II, Chapter 10, on LIGA, we described how UV and LIGA techniques provide some of the highest-resolution ceramic parts. Finally, in Volume II, Chapter 7, we introduced atomic layer deposition by which precursor gases or vapors alternately are pulsed onto the substrate surface to make Al_2O_3 (see Volume II, Figures 7.47 and 7.48).

Here we introduce Al_2O_3 and TiO_2 micromachining work for high-accuracy ceramic MEMS with the potential for a plethora of novel micro- and nanomachining applications.

Micromachining of alumina ceramic appears particularly appealing and might have advantages over other high-temperature MEMS substrates, such as quartz, 3C-SiC, diamond, and GaAs (see Chapter 1, this volume, Table 1.10), in terms of cost, manufacturability, and performance.

Routkevitch et al.[42] and Li et al.[43] have made fascinating microstructures from nanoporous anodic alumina films (also known as porous anodic alumina). Anodic alumina is a self-organized material with unique pore morphology, composition, and properties. Regular porous films, from less than 100 nm thick to more than 300 μm, are formed when aluminum metal is anodized in solutions of phosphoric, oxalic, or sulfuric acid (in low concentrations—from 2 wt% to 10 wt%). Pore diameter is tunable from a few nm to several hundred nm, and pore densities vary from 10^9 to 10^{11} cm^{-2}. For typical anodization voltages (50–100 V), the pore diameter ranges from 0.15 to 0.3 μm, and by sintering above 830°C in air, the amorphous anodic alumina may be converted into crystalline material, thereby increasing its chemical and thermal

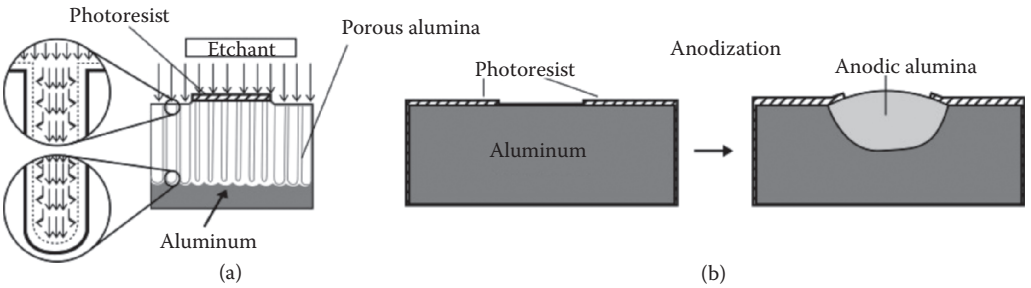


FIGURE 5.9 Schematic of two approaches for micromachining of anodic alumina. (a) High-resolution micromachining. (b) Lower-resolution microstructure prepared by localized anodization of aluminum. (Based on Routkevitch, D., A. Govyadinov, and P. Mardilovich. 2000. High aspect ratio, high resolution ceramic MEMS. *2000 International Mechanical Engineering Congress and Exposition,* 1–6.[42])

stability dramatically. The schematic in Figure 5.9 summarizes two approaches for micromachining of porous anodic alumina. In the higher-accuracy micromachining method (Figure 5.9a), the alumina film is anisotropically etched after patterning with conventional lithography. As prepared, anodic alumina dissolves readily at a pH lower than 4.2 and at a pH higher than 9.9, and the dissolution process is highly anisotropic. The walls of the porous pipes through the alumina film (see Figure 5.10) consist of an anion-contaminated outer layer and a purer inner alumina layer. The outer anion contaminated layer etches 25 times as fast as the purer inner layer. This etching rate difference and the anisotropic nature of the pores themselves, which provides access for the etchant throughout the entire thickness of the film, results in microstructures with aspect ratios as high as 50 (Tan et al.[41] projected the possibility of a 1000:1 aspect ratio) and a resolution better than 2 µm. The lateral underetching of the photoresist mask stops once the purer alumina at the wall of the first pore closed off at the top by the mask is reached by the etchant. Obviously, the smaller the pore diameter, the better the expected resolution of the technique. Routkevitch et al.[42] developed a second type of alumina micromachining process based on local anodization of aluminum metal. This process is simpler but has a somewhat poorer resolution limit (~20 µm) as pictured schematically in Figure 5.9b. In this case, the sidewalls do not remain vertical, since aluminum under the mask gets oxidized, and the alumina expands as compared with the initial volume of aluminum. Either the aluminum metal or the aluminum oxide may be etched selectively to reveal textured surfaces, with one being the exact imprint of the other. For selective etching of anodic alumina, a mixture of chromic and phosphoric acids is used, and aluminum may be selectively etched away in a solution of $CuCl_2$ in HCl. This work presents a whole new approach to microstructuring ceramic and metal surfaces.

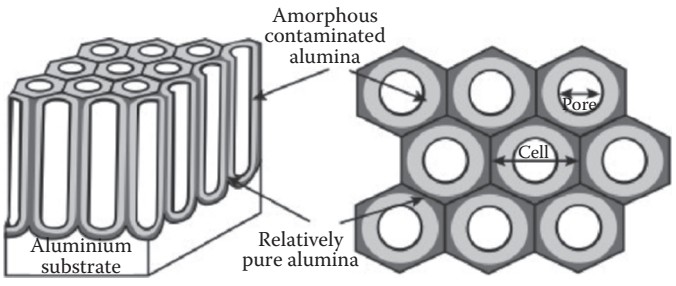


FIGURE 5.10 Schematic morphology and chemical composition of amorphous alumina. (Based on Routkevitch, D., A. Govyadinov, and P. Mardilovich. 2000. High aspect ratio, high resolution ceramic MEMS. *2000 International Mechanical Engineering Congress and Exposition,* 1–6.[42])

The nanoporous morphology of this ceramic MEMS material presents a controllable, very high surface area (up to 100 m^2/g) with numerous opportunities for applications such as a mechanical filter or as a support for catalysts and gas sensors, the latter of which require a high number of triple points (i.e., sites where gas reactants, catalyst, and support can meet). Recently, the method has been used in template-assisted nanowire growth as well.

Mechanical filters based on porous alumina are called *Anodisc™ Membrane Filters*. These aluminum oxide membrane filters are available in three nominal sizes (47, 25, and 13 mm diameter) and three pore sizes (0.2, 0.1, and 0.02 µm). Inorganic mechanical

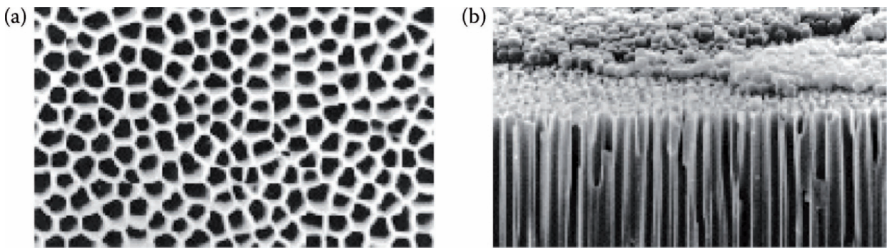


FIGURE 5.11 (a) Top view of a 0.2-μm Anopore™ inorganic membrane. (b) A cross-sectional view (below right) showing 0.23-μm latex microspheres retained on the membrane's surface. Membrane thickness is approximately 60 μm. Note the extraordinarily high pore density, which makes very high flow rates possible—unusual for membrane filters with pore sizes this small.

filters have several advantages over polymer membranes (e.g., polycarbonate sheets, about 10 μm thick, with pore sizes in the range of 15–12,000 nm), including higher flow rates, sieving of particles at the surface rather than trapping at an uncertain depth within the membrane and temperature and electron beam radiation resistant (see Figure 5.11) For more information, see http://www.2spi.com/catalog/spec_prep/filter2.shtml.

Porous alumina-based humidity sensors have been demonstrated.[44] The insulating material also makes for an excellent low-power and very-high-temperature microheater substrate with demonstrated temperatures up to 1200°C, power consumption as low as 50 μW/°C, and a very fast response.[44] The latter application represents an important alternative to Si as a microheater substrate in Taguchi sensors (see Volume II, Figures 9.11 and 9.12); Si is limited in temperature range and is a rather good heat conductor, so that it needs to be passivated for this application (it is also considerably more expensive).

The nanopores can also serve as a template for the deposition of nanostructures of a material of choice, such as metals, semiconductors, and polymers.[45] This is demonstrated in Figure 5.12 for the electrochemical deposition of nanowire thermoelectric material Bi_2Te_3. The process allows the oxide barrier layer to be removed/dissolved so that pores are in contact with the substrate. Filling rates of up to 90% have been achieved (T. Sands, HEMI group, http://www.mse.berkeley.edu/groups/Sands/HEMI/nanoTE.html).

The growth of aligned arrays of carbon nanotubes (CNTs) inside the pores of the self-organized nanoporous alumina templates is an especially intriguing application. Govyadinov et al.[46] synthesized arrays of CNTs (deposited with catalytic CVD) in alumina pores with diameters in the sub-100-nm range and densities from 10^8 to 10^{11} tips per cm^2. These prototype CNT arrays have been used as field emitters (carbon has a low work function) and shown to exhibit field emission at a low threshold of 3–5 V/μm and emission current densities up to 100 mA/cm^2 in low-vacuum conditions.

Other applications of porous alumina include microchannel plates, mesoscopic engines, filters, membranes, and nozzles and apertures.

Nanostructures similar to the ones shown in porous alumina have been produced by photoelectrochemical etching of polycrystalline n-type semiconductors, such as TiO_2, CdSe, CdS, and ZnO.[47] The most detailed work was performed on TiO_2. Results demonstrate an underlying anisotropic etching mechanism substantially different from the one

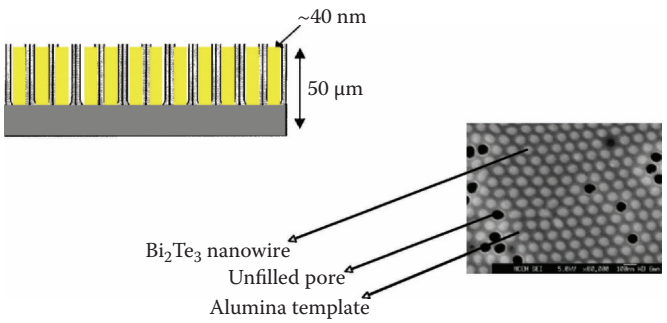


FIGURE 5.12 Nanoporous alumina works well for the electrochemical deposition of thermoelectric materials (Bi_2Te_3) and for metals. The process allows for the removal/dissolution of the oxide barrier layer so that pores are in contact with substrate. Filling rates of the pores of up to 90% have been achieved. (From T. Sands/HEMI group.)

at play in anodized alumina films.[48] Anodization of polycrystalline TiO_2 in a sulfuric acid solution under UV irradiation causes photoelectrochemical corrosion or photoetching. After a prolonged photoetching procedure, regularly ordered submicrometer porous structures (called *nanohoneycombs*) are observed at the TiO_2 surface. The quadrangle cells of the revealed nanohoneycomb are a few hundred nanometers wide and several micrometers deep. The thin walls (10–20 nm) of the pattern consist of {110} crystal faces. Etching selectively dissolves the grains, while the grain boundaries remain. The explanation for this selective etching behavior is readily understood from the fact that the holes created in the valence band of the semiconductor by the UV light readily corrode the bulk of the grains but recombine with electrons at the grain boundaries, leaving those intact. Besides their use in a dye-sensitized solar-cell and as a photocatalyst, fewer applications have been suggested for these new nanostructures than for porous alumina.

Conceptually, the alumina and TiO_2 micromachining approaches described here are somewhat similar to the quartz micromachining approach discussed earlier in which high-aspect ratio micromachining is achieved in single-crystal quartz by etching along the nuclear tracks left after bombardment with heavy ions. Obviously, alumina micromachining is a lot less complicated and costly, and so is the TiO_2 etching procedure, although the latter entails a rather difficult to control photoetching step and results in much less regular patterns than porous alumina.

Polymer MEMS-Polyimides

Introduction

Quartz and ceramic micromachining, discussed above, principally address harsh environment MEMS. Polymer materials also have some distinct advantages over more traditional engineering materials used in MEMS. Incentives for using polymers include increased fracture strength, low Young's modulus, high elongation, and reduced material costs. Furthermore, polymers are often inert and biocompatible, making them attractive for use in biological and chemical applications. As we have seen, in Volume I, Chapter 6, on fluidics, polymers are the most-used substrate for all types of fluidic platforms. We saw plenty of other examples of polymer MEMS in the section on soft lithography and dry photoresists in Volume II, Chapters 1 and 2, on lithography, in Volume II, Chapter 10, on replication and LIGA, and in Chapter 4, this volume, on packaging, assembly, and self-assembly. We conclude this section on new MEMS and NEMS substrate materials by considering the many applications of polyimides as an example of polymer MEMS.

Polyimide MEMS

Polyimides are part of a group of high-temperature polymers that have been in commercial use since the 1970s. Because of their structure, involving very strong carbon ring bonds, they do not melt and flow as most thermosets and thermoplasts do. Polyimides are synthesized from two monomers, a dianhydride and a diamine. Commercial products are supplied as soluble polyamic acid (PAA) intermediates, which undergo a thermal imidization with the evolution of water to form the insoluble polyimide. Polyimide films feature excellent thermal stability (up to ~450°C for short periods of time, $T_g > 300$°C), low dielectric properties ($\varepsilon = 3.3$) and a resistivity of ~10^{16} ohm-cm, superior chemical resistance, toughness, wear resistance, and flame retardance. Because of their flexibility they also exhibit interesting mechanical properties, and the material is easy to process. However, its tendency to absorb moisture causes the polyimide layers to swell and their dielectric constants to increase significantly. Another negative aspect of polyimides is their short shelf-life stability, which still needs to be addressed by organic chemists. Various polyimide analogs have been developed to avoid some of these inherent undesirable properties.

Current macro applications of nonphotosensitive polyimides include ball-bearing separators and mechanical seals. Polyimides in the microworld are used as passivating and interlayer dielectrics, planarizing compounds, reactive ion masks, alpha-particle barriers, humidity-sensitive materials,[49,50] liquid crystal display, color filters[51] (by incorporating dyes), and other optical elements, such as

waveguides. In planarization, polyimides smooth the undulation caused by topographic features on the wafer, so that the top-imaging layer has a much smaller thickness variation (see Volume II, Chapter 1, on photolithography). A typical polyimide, designed for MEMS applications, packaging dielectric and as substrate material is PI 2611, a low-moisture, low-stress (coefficient of thermal expansion (CTE) = 5 ppm), and high modulus of elasticity (8.5 GPa) polyimide. This polyimide can be patterned using dry etching techniques as well as TMAH etching (for example, using the developers MF 319 and AZ 300). Film thicknesses in the range 4–8 μm can be obtained in a single coat with good uniformity by varying spin speeds in the range 5000–1000 rpm. The polyimide substrate can withstand metal sputtering such as Cr, Au, and Al without substantial damage.

An organic ciliary array of thin-film polyimide bimorph microactuators exploiting combined thermal and electrostatic control was employed to implement sensorless manipulation strategies for small objects. The tasks of parts translation, rotation, orientation, and centering were demonstrated using small integrated circuit (IC) dice by Kovacs et al. (see Figure 5.13).[52,53] To induce motion on a part that is placed on an array of cilia, the cilia are actuated in a cyclic, gait-like fashion. In each cycle, the part is moved in a certain direction by the motion of the actuators that are in contact with it. The speed of the moving part depends on the (horizontal) displacement of the actuators per cycle, as well as the frequency of cycle repetition. It also depends on the surface properties and weight of the moving part.

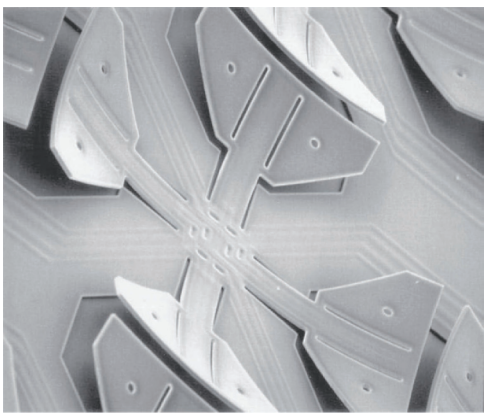


FIGURE 5.13 Polyimide cilia motion pixel (scanning electron microscope micrograph). Four actuators in a common center configuration make up a motion pixel. Each cilium is 430 μm long and bends up to 120 μm out of the plane. (Figure by John Suh.[52])

In our own work, we have used polyimide sheets as precursor material to design carbon structures in C-MEMS (see below), and in building flat panel displays at SRI International, Capp Spindt used polyimide posts to separate the glass plate from the Si field emitters (see Volume II, Figure 7.90).

Photosensitive polyimides have found many applications in both integrated circuitry and in MEMS and NEMS. Photosensitive polyimide is typically a negative tone resist and in use as photoresist, polyimide precursors, called polyamic acids (*PAAs*), are spun on the wafer; upon exposure to UV light, cross-linking results (more recently introduced positive polyimide resists are discussed in Volume II, Chapter 1, on photolithography; see Volume II, Figure 1.28). During development, the unexposed regions are dissolved, and final curing by further heat treatment leads to a chemical transformation (known as *imidization*) of the remaining cross-linked material, which yields polyimide. At a temperature of 275°C, more than 99% of the polyimide precursor is converted to polyimide. As with other negative resists, an oxygen effect is noted in the exposure. The actinic sensitivity of the resist usually is confined to 365 nm.

Polyimide photoresists often are used as permanent photoresists; that is, as resists that after exposure and curing can be left behind as a structural and/or functional component. For example, they have been used as a dielectric, especially when the uniquely low dielectric constant of the polyimide gives devices a decreased capacitance, resulting in increased speed for electronic applications.[54] In multiple spin-coats, thick polyimide films can be deposited (>100 μm). Since the films are so transparent, UV lithography permits the fabrication of high aspect ratio features (see LIGA-like processes in Volume II, Chapters 1 and 10). An important commercial application of both photosensitive and nonphotosensitive polyimides is in multiple chip modules.[55] As we explained in Volume II, Chapter 1 on lithography in the multiple chip module application, the use of photosensitive polyimides dramatically simplifies the manufacturing procedure as compared with non-photosensitive option, as illustrated in Figure 1.28.[55]

In a finished device, up to five layers of metallization are insulated by 15–20-μm-thick polyimide films on a ceramic wiring board.

Photosensitive polyimides also come as dry-film photoresists. Dry photoresists were introduced by DuPont as far back as 1968. These dry films come in rolls and consist of a photopolymerizable layer sandwiched between a polyester support film and a separator sheet (see Volume II, Figure 1.33). They do not need to be spin coated on, as they can, after removal of the separator sheet, be laminated onto a substrate. Also, the polyester cover sheet protects the resist film from oxygen diffusion, which would inhibit the cross-linking reaction. Both chlorinated and aqueous-base-developable dry film resist systems are available. The current resolution capabilities of dry film resists are less than 75 μm. The fact that the laminated dry film can bridge or "tent" over holes and cavities not only benefits the manufacture of printed wiring boards (PWBs) but can be exploited in the manufacture of biosensors as well, as we illustrate in Chapter 1, this volume, Example 1.1. For other photosensitive dry resist systems, see Table 1.12.

Exploration of IC and MEMS applications of modified polyimides also has started. A common approach to reducing the dielectric constant of polyimides is the inclusion of organofluorine components, in the form of pendant perfluoroalkyl groups. The polyamic acid precursor for fluorinated polyimides is commonly based on hexafluorodianhydrideoxydianiline. The inclusion of fluorinated monomers in the polyimide backbone has been found to reduce their moisture absorption and dielectric constant. Unfortunately, these polyimides generally show an increased susceptibility to chemical attack, making their use in multilayer fabrication questionable. Recently, however, newer formulations have improved the chemical resistance of the polyimide and have demonstrated a unique wet etch capability for vias with aspect ratios approaching 1.2:1.

Special perfluorinated polyimides were synthesized at the NTT laboratories in Japan. These remarkable materials are resistant to soldering temperatures (260°C) and are highly transparent at the wavelengths of optical communications (1.0–1.77 mm). In addition, their low dielectric constants and refractive indices match the conventional fluorinated polyimides, whereas their birefringence is lower.[56] As we saw in Volume II, Example 3.4, dry plasma etching of these modified polyimides results in very high aspect ratio microstructures with interesting optical properties.

Silicon Nitride

Introduction

Silicon nitride (Si_xN_y) is a commonly used material in microcircuit and microsensor fabrication because of its many superior chemical, electrical, optical, and mechanical properties. The material provides an efficient passivation barrier to the diffusion of water and to mobile ions, particularly Na^+. It also oxidizes slowly (about 30 times less than silicon) and has highly selective etch rates over SiO_2 and Si in many etchants. Some applications of silicon nitride are optical waveguides (nitride/oxide), encapsulant (diffusion barrier to water and ions), insulator (10^{16} Ω-cm, high dielectric strength, field breakdown limit of 10^7 V/cm), mechanical protection layer, etch mask, oxidation barrier, and ion implant mask (density is 1.4 times that of SiO_2). Silicon nitride also is hard, with a Young's modulus higher than that of Si, and can be used, for example, as a bearing material in micromotors.[57]

Silicon nitride can be deposited by a wide variety of the CVD techniques reviewed in Volume II, Chapter 7—APCVD, low-pressure chemical vapor deposition (LPCVD), and plasma-enhanced chemical vapor deposition (PECVD)—and its intrinsic stresses can be controlled by the specifics of the deposition process. Silicon nitride and silicon oxide deposited with these techniques usually exhibit too much residual stress, which hampers their use as mechanical components. However, CVD of mixed silicon oxynitride can produce substantially stress-free components.

Nitride often is deposited from SiH_4 or other Si containing gases and NH_3 in a reaction such as the following.

$$3SiCl_2 + 4NH_3 \rightarrow Si_3N_4 + 6HCl + 6H_2$$

Reaction 5.1

In this CVD process, the stoichiometry of the resulting nitride can be moved toward a silicon-rich composition by providing excess silane or dichlorosilane compared with ammonia.

PECVD Nitride

Plasma-deposited silicon nitride, also plasma nitride or SiN, is used as the encapsulating material for the final passivation of devices. The plasma-deposited nitride provides excellent scratch protection, serves as a moisture barrier, and prevents sodium diffusion. Because of the low deposition temperature, 300–350°C, the nitride can be deposited over the final device metallization. Plasma-deposited nitride and oxide both act as insulators between metallization levels, which is particularly useful when the bottom metal level is aluminum or gold. The silicon nitride that results from PECVD in the gas mixture of Reaction 5.1 has two shortcomings: high hydrogen content (in the range of 20–30 atomic percent) and high stress. The high compressive stress (up to 5×10^9 dyn/cm^2) can cause wafer warping and voiding and cracking of underlying aluminum lines.[58] The hydrogen in the nitride also leads to degraded MOSFET lifetimes. To avoid hydrogen incorporation, source gases containing little or no hydrogen, such as nitrogen, may be employed instead of ammonia as the nitrogen source. Also, a reduced flow of SiH_4 results in less Si-H in the film. The hydrogen content and the amount of stress in the film are closely linked. Compressive stress, for example, changes to tensile stress upon annealing to 490°C in proportion to the Si-H bond concentration. By adding N_2O to the nitride deposition chemistry an oxynitride forms with lower stress characteristics; however, oxynitrides are somewhat less effective as moisture and ion barriers than nitrides.

One key advantage of PECVD nitride, besides the low temperature aspect, is the ability to control stress during deposition. Silicon nitride deposited at the typical 13.56 MHz plasma excitation frequency exhibits tensile stress of about 400 MPa. When depositing the film at 50 kHz, a compressive stress of 200 MPa is typical. We discussed the effect of RF frequency on nitride stress, hydrogen content, density, and the wet etch rate in more detail in Volume II, Chapter 7. We concluded that low frequency (high energy bombardment) results in films with low compressive stress, lower etch rates, and higher density. Therefore, the effect of ion bombardment is more pronounced at lower pressures, and better quality CVD films ensue, characterized as films with a low wet etch rate (high film density) and low compressive stress. Experimental results show that as the reactor pressure is lowered, film stress goes from tensile to compressive (Volume II, Figure 7.59) and wet etch rates decrease. Often, stress dynamically changes when the film is exposed to the atmosphere and subsequent heating; the higher the ion bombardment during PECVD, the higher the stress. Stress also is affected by moisture exposure and temperature cycling (for a good review, see Wu and Rosler[59]).

We compare properties of silicon nitride formed by LPCVD and PECVD in Table 5.5, and Table 5.6 highlights typical PECVD process parameters. The refractive index of the deposited silicon nitride films is a measure of impurity content and overall quality. Its value ranges from 1.8 to 2.5 for PECVD films as compared with 2.01 for stoichiometric LPCVD (see below). The higher numbers indicate excess silicon, and lower numbers represent excess oxygen.

LPCVD Nitride

In the IC industry, stoichiometric silicon nitride (Si_3N_4) is LPCVD deposited at 700–900°C and at 200–500 mTorr, and it functions as an oxidation mask and as a gate dielectric in combination with thermally grown SiO_2. In micromachining, LPCVD nitride serves as an important mechanical membrane material and isolation/buffer layer. The standard source gases are SiH_2Cl_2 and ammonia, and at the above specified temperatures and pressures, this leads to a deposition rate of about 30 Å/min. The deposition of the amorphous silicon nitride is reaction-limited, so it deposits on both sides of the wafers with equal thickness. Dichlorosilane is used instead of silane because it results in more uniform film thickness, and the wafers can be spaced closer together for larger loads. For stoichiometric silicon nitride, typical gas flows are in a 10:1 ratio of ammonia to dichlorosilane. Stoichiometric silicon nitride has a large residual tensile stress of about 10^{10} dynes/cm^2, and as a consequence, film thickness

Table 5.5 Properties of Silicon Nitride

Deposition	LPCVD	Plasma-Enhanced CVD
Temperature (°C)	700–800	250–350
Density (g/cm^3)	2.9–3.2	2.4–2.8
Pinholes	No	Yes
Throughput	High	Low
Step coverage	Conformable	Poor
Particles	Few	Many
Film quality	Excellent	Poor
Dielectric constant	6–7	6–9
Resistivity (ohm-cm)	10^{16}	10^6–10^{15}
Refractive index	2.01	1.8–2.5
Atom % H	4–8	20–25
Energy gap	5 eV	4–5 eV
Dielectric strength (10^6 V/cm)	10	5
Etch rate in conc. HF	200 Å/min	
Etch rate in BHF	5–10 Å/min	
Residual stress (10^9 dyne/cm^2)	1 T	2 C–5 T
Poisson ratio	0.27	
Young's modulus	270 GPa	
TCE	1.6×10^{-6}/°C	

Sources: Based on 1) Adams, A. C. 1988. In *VLSI technology,* ed. S. M. Sze. New York: McGraw-Hill[60]; 2) Sinha, A. K., and T. E. Smith. 1978. Thermal stresses and cracking resistance of dielectric films. *J Appl Phys* 49:2423–26[61]; and 3) Retajczyk, T. F. J., and A. K. Sinha. 1980. Elastic stiffness and thermal expansion coefficients of various refractory silicides and silicon nitride films. *Thin Solid Films* 70:241–47.[62]

Note: C = compressive; T = tensile.

often is limited to a few thousands angstroms. By increasing the Si content in silicon nitride, the tensile film stress reduces (even to compressive), the film turns more transparent, and the HF etch rate lowers. Such films result by increasing the dichlorosilane to ammonia ratio.[63] At an ammonia to dichlorosilane ratio of approximately 1 to 6, the films are nearly stress free, for deposition temperatures of 850°C and pressures of 500 mTorr.[64] Figure 5.14 illustrates the effect of gas flow ratio and deposition temperature on stress and the corresponding refractive index and HF etch rate.[63]

Table 5.6 Silicon Nitride Plasma-Enhanced Chemical Vapor Deposition Process Conditions

Flow (sccm)	SiH_4	190–270
	NH_3	1900
	N_2	1000
Temperature (°C)	T.C.	350 or 400
	Wafer	~330 or 380
Pressure (Torr)		2.9
RF power (watts)	1100	
Deposition rate (Å/min)		1200–1700
Refractive index		2.0

Source: Wu, T. H. T., and R. S. Rosler. 1992. Stress in PSG and nitride films as related to film properties and annealing. *Solid State Technol* May, 65–71. With permission.[59]

Silicon-rich or low-stress nitride has emerged as an important micromechanical material. Low residual stress means that relatively thick films can be deposited and patterned without fracture. Low etch rate in HF means that films of silicon-rich nitride survive release etches better than stoichiometric silicon nitride. The etch characteristics of LPCVD Si_xN_y are summarized in Table 5.7. The properties of a LPCVD Si_xN_y film, deposited by the reaction of $SiCl_2H_2$ and NH_3 (5:1 by volume) at 850°C, were already summarized in Table 5.5.

Amorphous Silicon and Hydrogenated Amorphous Silicon

Amorphous silicon behaves quite different from the coarse- and the fine-grained polysilicon discussed in Volume II, Chapter 7. Amorphous Si can be stress annealed at temperatures as low as 400°C.[65] This low-temperature anneal makes the material

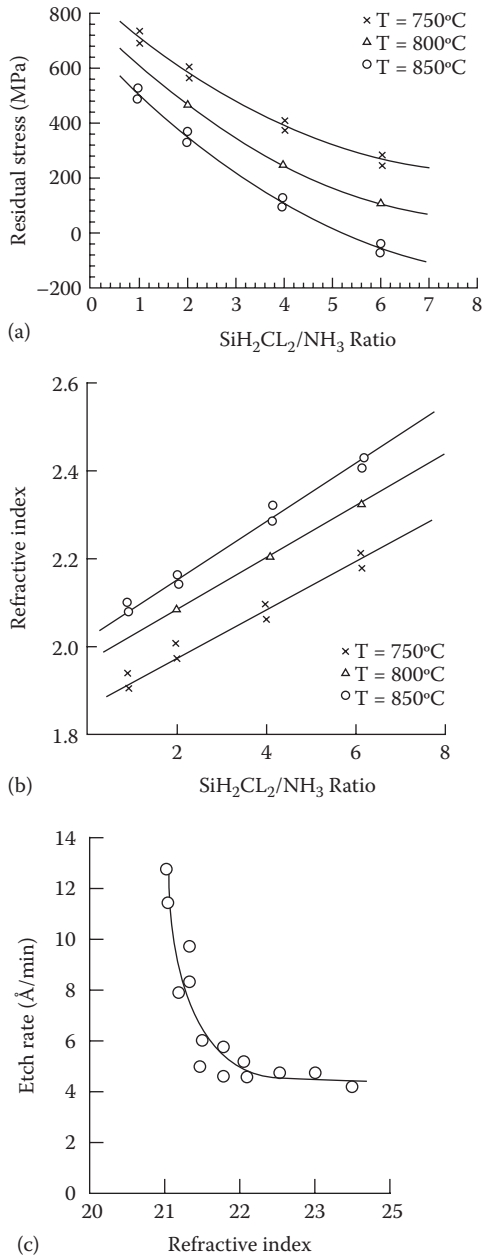


FIGURE 5.14 Silicon nitride low-pressure chemical vapor deposition parameters. (a) Effect of gas-flow ratio and deposition temperature on stress in nitride films. (b) The corresponding index of refraction. (c) The corresponding HF etch rate. (From Sakimoto, M., H. Yoshihara, and T. Ohkubo. 1982. Silicon nitride single-layer x-ray mask. *J Vac Sci Technol* 21:1017–21. With permission.[63])

compatible with almost any active electronic component. Unfortunately, little is known about the mechanical properties of amorphous Si.

Hydrogenated amorphous Si (α-Si:H), with its interesting electronic properties, is even less well understood in terms of its mechanical properties. If the mechanical properties of hydrogenated amorphous Si were found to be as good as those of polysilicon, the material might make a better choice than polysilicon as a MEMS material, given its better electronic characteristics.

TABLE 5.7 Etching Behavior of Low-Pressure Chemical Vapor-Deposited Si_3N_4

Etchant	Temperature (°C)	Etch Rate (Å/min)	Selectivity of Si_3N_4:SiO_2:Si
H_3PO_4	180	100	10:1:0.3
CF_4–4% O_2 plasma		250	3:2.5:17
BHF	25	5–10	1:200:±0
HF (40%)	25	200	1:>100:0.1

The amorphous polysilicon material produces a high breakdown strength (7–9 MV/cm) oxide with low leakage currents (vs. a low breakdown voltage and large leakage currents for polycrystalline Si oxides). Amorphous polysilicon also attains a broad maximum in its dielectric function without the characteristic sharp structures near 295 and 365 nm (4.2 and 3.4 eV) of crystalline polysilicon. Approximate refractive index values at a wavelength of 600 nm are 4.1 for crystalline polysilicon and 4.5 for amorphous material.[60] As deposited, the material is under compression, but an anneal at temperatures as low as 400°C reduces the stress significantly, even leading to tensile behavior.[65]

Like polysilicon, amorphous silicon exhibits a strong piezoresistive effect. The gauge factor is about five times smaller than that of single-crystal Si, but the temperature coefficient of the resistance (TCR) is lower than that of Si.

Hydrogenated amorphous silicon enables the fabrication of active semiconductor devices on foreign substrates at temperatures between 200°C and 300°C. The technology, first primarily applied to the manufacturing of photovoltaic panels, now quickly is expanding into the field of large-area microelectronics, such as active matrix liquid crystal displays (AMLCD). It is somehow surprising that micromachinists have not taken more advantage of this material to either power surface micromachines or implement electronics cheaply on nonsilicon substrates. In Figure 5.15 we compare polycrystalline-based solar cells with ultrathin amorphous Si on glass and ultrathin amorphous Si on flexible substrates.

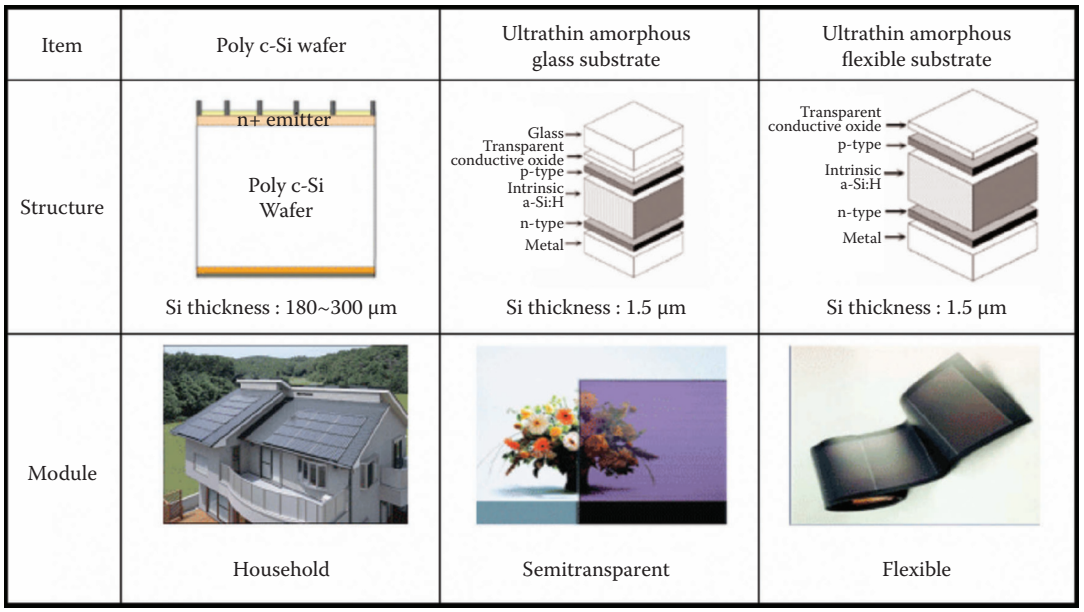


FIGURE 5.15 Comparison of polycrystalline-based solar cells with ultrathin amorphous Si on glass and ultrathin amorphous Si on flexible substrates. (From http://www.displaybank.com.)

Spear and Le Comber[66] showed that, in contrast to α-Si, α-Si:H could be doped both n- and p-type. Singly bonded hydrogen, incorporated at the silicon dangling bonds, reduces the electronic defect density from ~10^{19}/cm^3 to ~10^{16}/cm^3 (typical H concentrations are 5–10 atomic percent—several orders of magnitude higher than needed to passivate all the Si dangling bonds). The lower defect density results in a Fermi level that is free to move, unlike in ordinary amorphous Si, where it is pinned. Other interesting electronic properties are associated with α-Si:H—exposure of α-Si:H to light increases photoconductivity by four to six orders of magnitude, and its relatively high-electron mobility (~1 cm^2/V s^{-1}) enables fabrication of useful thin-film transistors. Lee et al.[67] noted that hydrogenated amorphous silicon solar cells are an attractive means to realize an onboard power supply for integrated micromechanical systems. They point out that the absorption coefficient of α-Si:H is more than an order of magnitude larger than that of single-crystal Si near the maximum solar photon energy region of 500 nm. Accordingly, the optimum thickness of the active layer in an α-Si:H solar cell can measure 1 μm, much smaller than that of single-crystal Si solar cells. By interconnecting 100 individual solar cells in series, the measured open circuit potential reaches as high as 150 V under AM 1.5 conditions, a voltage high enough to drive onboard electrostatic actuators.

Hydrogenated amorphous silicon is manufactured by plasma-enhanced chemical vapor deposition (PECVD) from silane. Usually, planar RF-driven diode sources using SiH_4 or SiH_4/H_2 mixtures are used. Typical pressures of 75 mTorr and temperatures between 200°C and 300°C allow silane decomposition with Si deposition as the dominant reaction. Decomposition occurs by electron impact ionization, producing many different neutral and ionic species.[68] Deposition rates for usable device quality α-Si:H generally do not exceed ~2–5 Å/s, due to the effects of temperature, pressure, and discharge power. Table 5.8 gives state-of-the-art parameters for α-Si:H prepared by PECVD. Although its semiconducting properties are inferior to those of single-crystal Si, the material is finding more and more applications. Some examples are thin film transistor (TFT) switches for picture elements in AMLCDs,[69] page-wide TFT-addressed document scanners, and high-voltage TFTs capable of switching up to 500 V.[70] An excellent source for further information on amorphous silicon is the book *Plasma Deposition of Amorphous Silicon-Based Materials*.[71]

Table 5.8 Typical Optoelectronic Parameters Obtained for Plasma-Enhanced Chemical Vapor Deposition α-Si:H

	Symbol	Parameter
Undoped		
Hydrogen content		~10%
Dark conductivity at 300 K	σ_D	~10^{-10} $(\Omega\text{-cm})^{-1}$
Activation energy	E_σ	0.8–0.9 eV
Pre-exponent conductivity factor	σ_0	>10^3 $(\Omega\text{-cm})^{-1}$
Optical bandgap at 300 K	E_g	1.7–1.8 eV
Temperature variation of bandgap	$E_g(T)$	2–4 × 10^{-4} eV/K
Density of states at the minimum	g_{min}	>10^{15}–10^{17} cm^3/eV
Density of states at the conduction band edge		~10^{15}/cm^3
ESR spin density	N_s	~10^{21}/cm^3-eV
Infrared spectra		2000/640 cm^{-1}
Photoluminescence peak at 77 K		~1.25 eV
Extended state mobility		
Electrons	μ_n or μ_e	>10 cm^2/V-s
Holes	μ_p or μ_h	~1 cm^2/V-s
Drift mobility		
Electrons	μ_n or μ_e	~1 cm^2/V-s
Holes	μ_p or μ_h	~10^{-2} cm^2/V-s
Conduction band tail slope		25 meV
Valence band tail slope		40 meV
Hole diffusion length		~1 μm
Doped amorphous		
n-type[a]	σ_D	10^{-2} $(\Omega\text{-cm})^{-1}$
	E_g	~0.2 eV
p-type[b]	σ_D	10^{-3} $(\Omega\text{-cm})^{-1}$
	E_g	~0.3 eV
Doped microcrystalline		
n-type[c]	σ_D	≥1 $(\Omega\text{-cm})^{-1}$
	E_g	≤0.05 eV
p-type[d]	σ_D	≥1 $(\Omega\text{-cm})^{-1}$
	E_g	≤0.05 eV

Source: Crowley, J. L. 1992. Plasma enhanced CVD for flat panel displays. *Solid State Technol* February, 94–98. With permission.[68]

[a] 1% PH_3 added to gas phase.
[b] 1% B_2H_6 added to gas phase.
[c] 1% PH_3 added to dilute SiH_4/H_2, or 500 vppm PH_3 added to SiF_4/H_2 (8:1) gas mixtures. Relatively high powers are involved.
[d] 1% B_2H_6 added to dilute SiH_4/H_2.

Selected MEMS and NEMS Processes

Introduction

In this section we review selected MEMS/NEMS processes that might affect the direction that MEMS/NEMS takes next. Many relatively new and important processes, such as soft lithography (see Volume II, Chapter 1) and atomic layer deposition (see Volume II, Chapter 7), are covered elsewhere in this work. The following additional selection is rather arbitrary and complements the new materials/processes covered elsewhere. For exciting new MEMS/NEMS techniques, we survey carbon-based MEMS (C-MEMS), bulk metallic glasses (BMG), fast microfluidic CD prototyping, and micro glass blowing.

C-MEMS

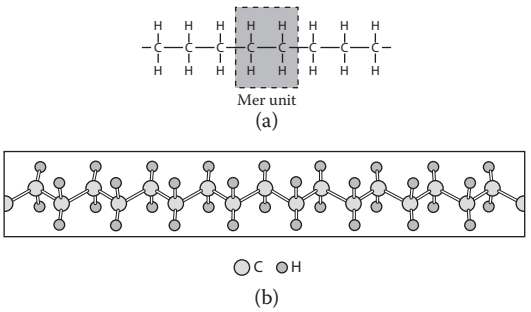
Introduction

In nature, carbon occupies a very special place as the essential building block of life. In human enterprise, too, carbon in its various forms has acquired an enviable position. To just name a few, diamonds are made of carbon, and graphite, coke, diamond-like carbon, and glassy carbon are all forms of carbon. So are the more recently discovered buckyballs, carbon fibers, and carbon nanotubes (CNTs).[72,73] Carboneous materials have been used in various application fields because their widely different crystalline structures and morphologies enable different physical, chemical, mechanical, thermal, and electrical uses. For example, glassy carbon electrodes often are used because of their wide electrochemical stability window, low background currents and low cost. Graphite and hard carbons are well known for their utility in Li ion battery applications because of their Li intercalation/deintercalation capacity. Other properties, such as biocompatibility, chemically inertness, and ease of functionality have made carbon very attractive for biosensor and implantable device applications. Moreover, CNTs are of tremendous current interest in both fundamental research and for nanoelectronics applications, including flat panel displays, chemical sensors, and the next generation of transistors (see Volume I, Figure 1.30). Recently, new microfabrication methods for carboneous materials have received a lot of attention, especially with regard to their use as microelectrodes in electrochemical sensors and miniaturized energy storage/energy conversion devices. Some key advantages of carbon compared with Si are summarized in Table 5.9. Although silicon, human's foremost technological material, is abundant in silicate minerals, carbon is the key to life. It can rightfully be argued that carbon is becoming a more and more important material in human technology. Besides the common use of carbons in sensors and batteries, carbon nanotubes might form the basis for a future carbon-electronics industry. We speculate that C-MEMS, described here, might be the material to use to contact those carbon-nanotube-based devices.

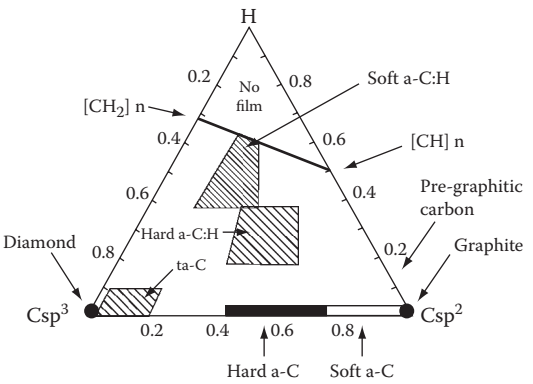
Additional important C-MEMS properties are included in Table 5.9: the material has a very wide

TABLE 5.9 Carbon: The Next Si?

• Polymerizes better than Si (see inset 1 below)
• Takes all types of forms: amorphous, graphite, nanotubes, etc. (see inset 2 below)
• Has a wide electrochemical stability window
• Exhibits biocompatibility
• Has a low cost
• Is chemically inert
• Is easy to derivatize
• Is well known for its battery and sensor application
• Carbon nanotubes connect via carbon-based microelectromechanical system
• Is nature's building block for living things



Inset 1



Inset 2

electrochemical stability window, exhibits excellent biocompatibility, and is low cost, and the surface of this very chemically inert material is easy to functionalize. The material has particular importance in bio-MEMS applications, including electronic DNA arrays (replacing the platinum shown in Nanogen's DNA array; see Figure 8.51 this Volume), glucose sensors, and microbatteries (see below).

Microfabrication of Carbon: C-MEMS Process

For microfabrication of carbon structures, IC processing technologies, such as focused ion beam (FIB) and reactive ion etching (RIE), are time consuming and expensive because of the need of high-vacuum systems. Low feature resolution and poor repeatability of the carbon composition, as well as the widely varying properties of the resulting devices, limit the application of screen printing with commercial carbon inks for carbon microfabrication.[73] Recently, more nontraditional carbon microfabrication methods were reported. Whitesides and co-workers described the use of soft lithography (see Volume II, Chapter 2) to fabricate glassy carbon microstructures.[74] In this approach, micromolding of a resin such as poly(furfuryl alcohol) in an elastomeric mold yields polymeric microstructures, the microstructures are converted to free-standing carbon by heat treatment (500–1100 °C) in an inert atmosphere (see Volume II, Figure 2.55).

The carbon microfabrication technique introduced by this author is based on the pyrolysis of shaped/patterned polymers (e.g., resists patterned with photolithography) at different temperatures and different ambient atmospheres.[75,76] Carbon, because it is brittle and hard, is a difficult material to machine. Polymers, on the other hand, can be machined easily in a wide variety of machine tools. The underlying principle of C-MEMS is to choose an easy to work with polymer precursor, machine, or pattern this precursor material and then convert it to carbon by pyrolysis.

In one embodiment of the C-MEMS process, a photoresist is patterned by photolithography and is subsequently pyrolyzed at high temperatures in an oxygen-free environment. By changing the lithography conditions, soft and hard baking times and temperatures, additives to the resist, pyrolysis time, temperature, and environment, C-MEMS permits a wide variety of interesting new MEMS and NEMS applications that employ structures having a wide variety of shapes, resistivities, and mechanical properties. The process is detailed in Figure 5.16.[73]

The photolithography to make C-MEMS structures, illustrated in Figure 5.16, includes the usual spin coating, soft bake (not shown), near-UV exposure, postbake (not shown), and development. A typical patterning process for a 200-μm-thick SU-8 photoresist film involves photoresist spinning at approximately 500 rpm for 12 s and then at 1400 rpm for 30 s, followed by a 10-min soft-bake at 65°C and finally a soft-bake for 80 min at 95°C. Exposure is performed in a Karl Suss MJB3 contact aligner for about 100 s. The postbake is carried out for 2 min at 65°C and for 30 min at 95°C. Development is

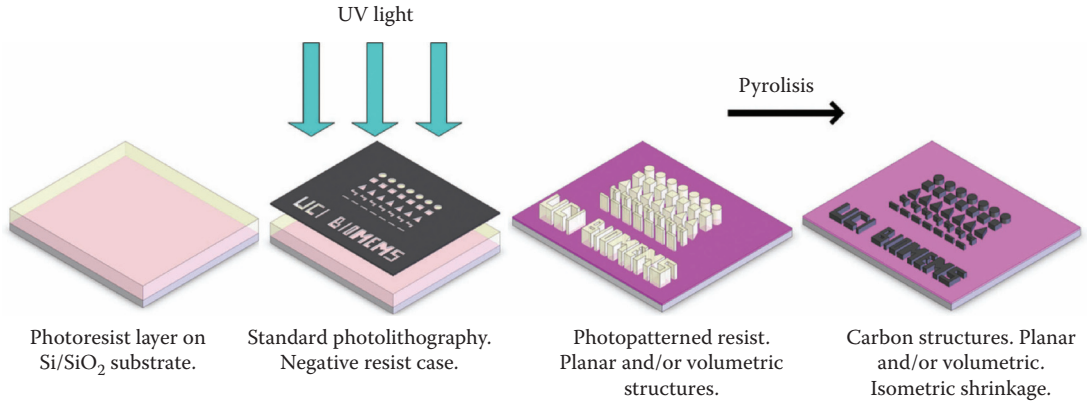


FIGURE 5.16 An illustration of one type of carbon-based microelectromechanical system: photolithography patterning of a polymer precursor and pyrolysis (see text for details).

carried out using a SU-8 developer from MicroChem (NANO™ SU-8 developer).

Photoresist-derived C-MEMS architectures are obtained in a two-step pyrolysis process in an open-ended quartz-tube furnace, in which samples are postbaked in an N_2 atmosphere at 300°C for about 40 min first and then heated in a N_2 atmosphere (2000 sccm) up to 900°C. At this point, the N_2 gas is shut off and forming gas [H_2 (5%)/N_2] is introduced (2000 sccm) for 1 h, and then the heater is turned off and the samples are cooled down again in an N_2 atmosphere to room temperature. The heating rate is about 10°C/min, and the total cooling time is about 10 h.

The resulting carbon material was found to be amorphous and glass carbon-like in electrochemical behavior.[77] The average height of the SU-8 posts shown in two typical scanning electron microscope (SEM) pictures (Figure 5.17a and 5.17b), is around 300 µm.

As shown in Figure 5.17c and 5.17d, after pyrolysis, the overall shape of the cylindrical posts largely is retained, and a typical aspect ratio of the carbon post achieved is around 10:1. It was found that the C-MEMS posts shrink much less during the pyrolysis process near the base of the structures than at the midsection because of the good adhesion of SU-8 to the substrate.

The processes described above easily can be extended to two-level or multilevel C-MEMS structures. This feature was used for making Li-ion-based C-MEMS batteries.[78] In the C-MEMS battery depicted in Figure 5.18, one layer of carbon constitutes the current collector for the rows of anode and cathode posts made in a second C-MEMS layer.[78]

In Figure 5.18, the current collector is 20 µm thick and the posts are 250 µm high. C-MEMS Li ion batteries, as shown here, hold the promise of changing the battery world in three significant ways (see also Chapter 9, this volume, on power and brains in miniature devices): 1) Smarter batteries: C-MEMS batteries contain a multitude of anodes and cathodes (we call them *baxels*, by analogy to *pixels*) that can be interconnected to provide the optimal current or voltage depending on the application need. In other words, we can perform in-battery smart load

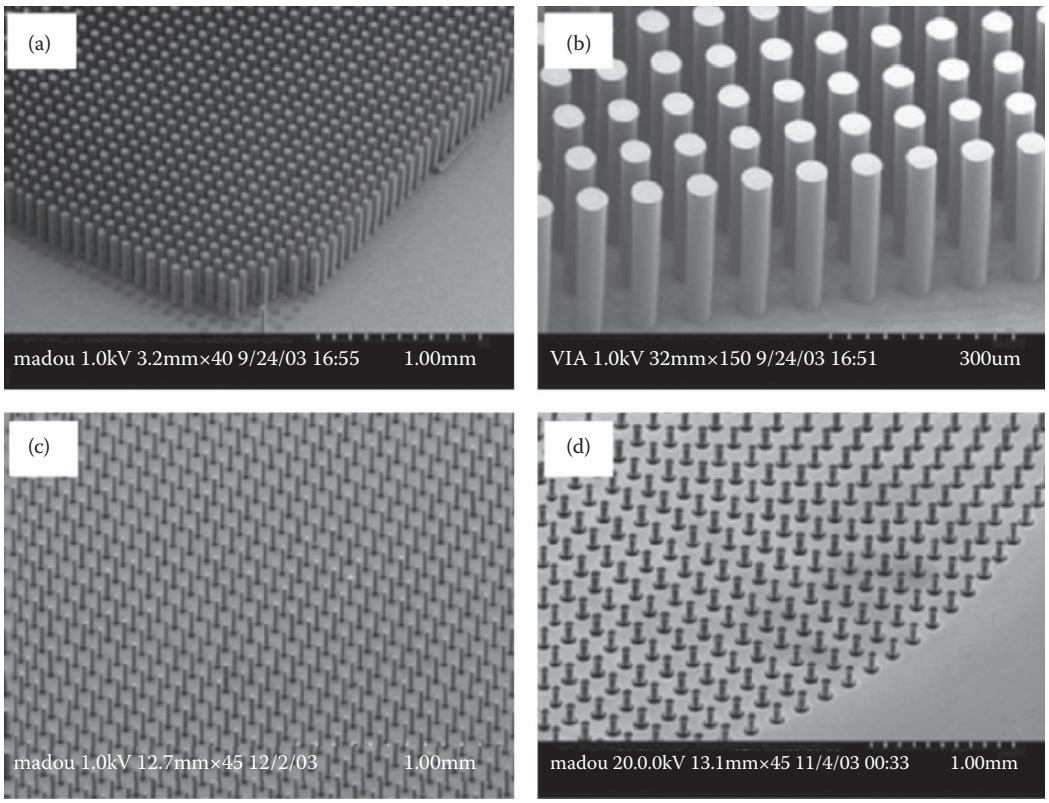


FIGURE 5.17 Typical scanning electron microscope photos. (a) and (b) SU-8 post arrays before pyrolysis. (c) and (d) Carbon post arrays after pyrolysis.

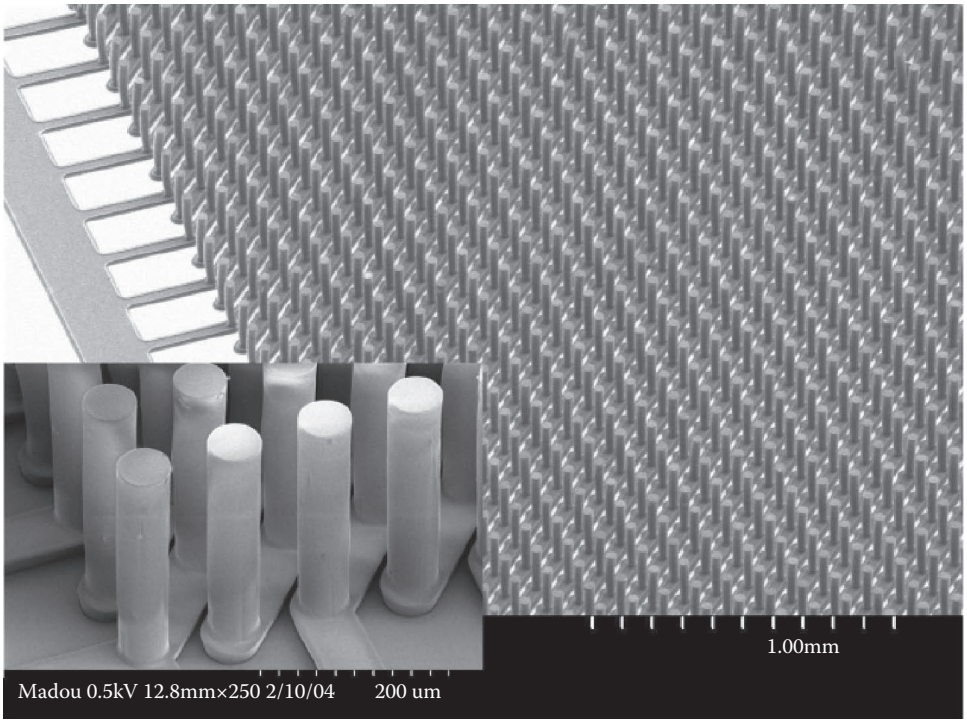


FIGURE 5.18 Typical two-level carbon-based microelectromechanical system electrodes with carbon contacts underneath. The inset image shows enlarged SU-8 two-level structure.[78]

leveling; 2) Faster charging batteries: in charging the C-MEMS battery, Li ions only need to intercalate into the thin individual carbon posts rather than intercalating into one big chunk of carbon; 3) Theoretically, the capacity of the C-MEMS batteries, depending on the aspect ratio of the posts and the thickness of the electrolyte, can be significantly higher than that of current Li ion batteries.[79]

In order to build yet higher aspect ratio C-MEMS structures (up to 40:1), we developed a three-level C-MEMS process. In a multilayer C-MEMS process, each layer of SU-8 photoresist is exposed and baked separately, but the whole assembly is developed all at the same time in the last step. Typical SEM photos of three-level C-MEMS devices are shown in Figure 5.19, with a three-level C-MEMS structure in which the first two are in good alignment but the third is out of alignment.

By carefully controlling the lithography processing parameters, a variety of complex 3D C-MEMS structures, such as suspended carbon wires, bridges, plates, and self-organized posts (carbon flowers) and ribbons (networks), were also microfabricated by Madou and coworkers.[72] These intricate 3D structures are fabricated using nontraditional process recipes, such as overexposure, underdevelopment, photoresist additives, directed flow of the developer, and exploitation of surface tension in the developing photoresist patterns. Suspended carbon fibers are shown in Figure 5.20, and a C-MEMS table was shown in the inset in the heading of this chapter. The diameter of the suspended carbon fibers in Figure 5.20 is submicrometer. The wire resistance between two posts was measured at room temperature and in air using two point probes, and a typical resistivity

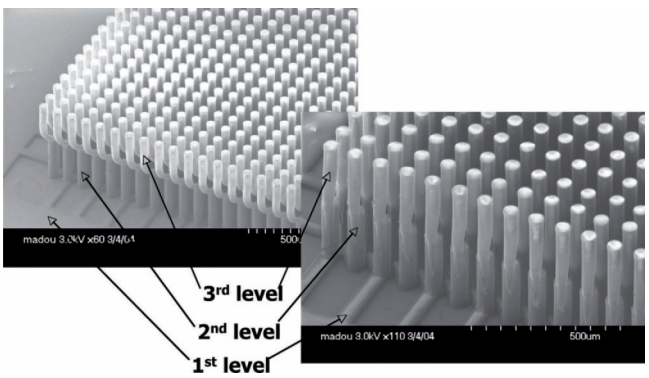


FIGURE 5.19 Typical scanning electron microscope photos of three-level carbon-based microelectromechanical system (C-MEMS) devices: a three-level C-MEMS structure with two levels in good alignment but with a third level out of alignment.

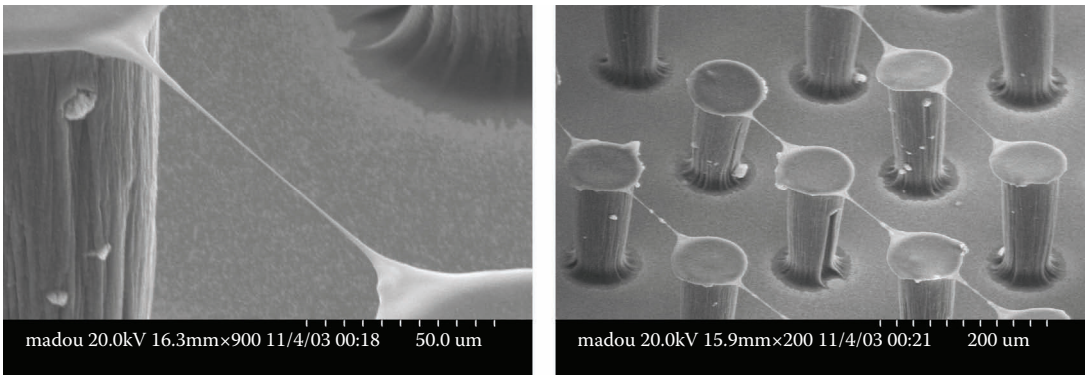


FIGURE 5.20 Carbon-based microelectromechanical system suspended carbon wires. Washcloth lines.

of 1.8×10^{-2} Ωcm was obtained. The Madou team is currently working on measuring the influence of chemical reactions on the suspended wire's resistivity. The aim is to turn these amorphous carbon "washcloth lines" into "washcloth line sensors" by derivatizing the C-MEMS surface with a monolayer of a sensor molecules (e.g., proteins, DNA). It is anticipated that the fact that these nanowires are removed from the supporting substrate and are thus suspended in air or liquid will lead to significant advantages over other nanowire-based sensors typically embedded on a solid substrate. This arrangement exposes the sensor to its sensing environment in all directions and avoids deleterious substrate influences, such as contamination and charge shunting.

In Figure 5.21, we illustrate self-organized groups of carbon posts (bunched posts)—a feature most readily observed when starting with high-aspect ratio SU-8 arrays with posts higher than 300 µm. It should be noted that when bunching of the SU-8 posts occurs, the bowed posts in each bunch remain structurally intact when converted to carbon. This means that the aggregation/bunching takes place after UV exposure, i.e., during postbake, development, or the drying after development. Bunching could in principle occur at the postbake temperature, as that temperature (65°C) is above the glass transition temperature of unexposed SU-8 (50–55°C); at those temperatures, the unexposed areas reflow, which could enable the posts to move toward each other. From SEM observations, after the postbake, the mask patterns are transferred with high fidelity. Clearly no aggregation or bunching is happening during postbake. The opportunity to control and optimize self-organized two-dimensional textures obviously presents itself only during the development process itself. We speculate at this point that the dominant cause for the bunching of resist posts is surface tension. When the developer solution is removed, it gently pulls posts that are tall and close enough together into symmetric patterns.

The electrochemical characteristics of carbon produced by pyrolysis of photoresists have been extensively studied.[76,77] It is found that electrochemical reactions on pyrolyzed photoresist exhibit reaction kinetics very similar to those on glassy carbon. We also found that C-MEMS electrodes can

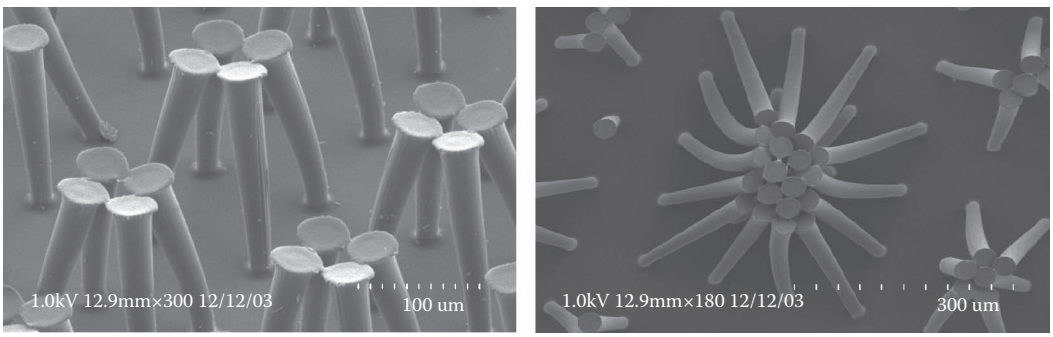


FIGURE 5.21 Carbon flowers. Self-assembly of polymer structures converted into carbon.

be reversibly charged and discharged with Li ions,[78] which led to the construction of C-MEMS Li ion batteries, as shown in Figure 5.18. The advantage of using photoresists as the starting material for the microfabrication of various carbon structures is that the photoresists can be very finely patterned by lithography techniques, and hence, a wide variety of repeatable shapes are possible. Moreover, different temperature treatments result in different resistivities and mechanical properties. Molecular rectification and conductance switching in pyrolyzed carbon-based molecular junctions were studied by McCreery et al.[80] They found that the current/voltage behavior showed strong and reproducible rectification and strong dependence on temperature and scan rate. Pyrolyzed photoresist carbon films were demonstrated to be useful in electroanalytical and microfluidic applications; for example, pyrolyzed carbon has been used in conjunction with a microchip capillary electrophoretic device.[81]

Madou and co-workers, as explained in Chapter 9, this volume, have presented theoretical arguments demonstrating that making fractal-like carbon electrodes will lead to better sensors and more energy efficient batteries and fuel cells.[82] In ongoing work, our team is attempting to make what we call a fractal garden. An artistic rendition of this concept is shown in Figure 5.22. In this fractal garden, each element in the array of fractal electrodes can be addressed individually. These electrodes maximize surface area while minimizing internal resistance and can be implemented gainfully for the sensors, batteries, and fuel cells of the future.

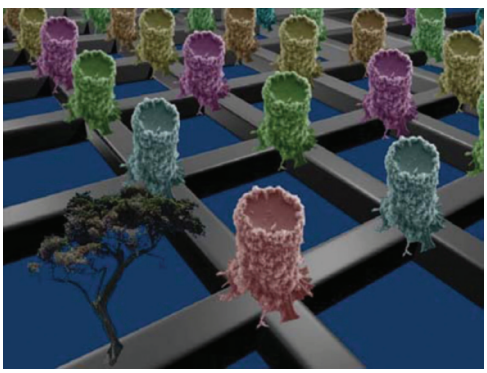


FIGURE 5.22 An artistic rendition of a carbon-based microelectromechanical system fractal garden (see also Chapter 9, this volume).

In Volume I, Chapter 6, on fluidics, we introduced 3D C-MEMS dielectrophoresis (see Volume I, Example 6.2). The main concept behind using 3D carbon electrodes here was to reduce the mean distance of any targeted particle contained in a channel or chamber to the closest electrode surface.[83]

The author believes that C-MEMS/NEMS technology can find applications in several MEMS fields as a substitute for silicon-based devices. This technology can be revolutionary not only in energy applications, such as biofuel cells, microbatteries, and supercapacitors, but also in biological and chemical sensing applications. The C-MEMS/NEMS approach gives the development engineers unprecedented freedom in designing and manufacturing high-surface area conductive structures through the use of new MEMS materials and innovative fabrication techniques. The development of novel biofuel cells based on high-aspect ratio 3D carbon structures from millimeter to nanometer size, integrating "top-down" and "bottom-up" processing approaches and combining soft biological components with hard MEMS/NEMS structures, could have the effect of making the bio-MEMS world less Si-centric and start a trend of lithographically patterning materials other than Si. Beyond the impact on the MEMS/NEMS/bio-MEMS community, one expects a broader anticipated impact on industry and on the technology end user.

Bulk Metallic Glasses

Bulk metallic glasses (BMGs) are multicomponent alloys for which crystallization is avoided during cooling. Sluggish crystallization kinetics and a corresponding ease of solidification into the amorphous state have been observed for a wide range of alloys, including compositions based on Zr, Cu, Ni, Ti, Fe, Mg, Pd, Pt, and Au.[84–94] Because of their random amorphous structure and lack of dislocation formation during deformation, these materials exhibit very high strength and elasticity, outstanding resilience, good corrosion resistance, and a near-perfect elastic behavior. For example, a typical BMG based on zirconium yields at 2000 MPa after it has elastically deformed by 2%.[95] Crystallization of a BMG must be avoided during any kind of processing, since it degrades the properties of the alloy. One

consequence of the sluggish crystallization kinetics in BMG forming alloys is the presence of a supercooled liquid region (SCLR). In this temperature region, the amorphous BMG first relaxes into a viscous liquid before it eventually crystallizes. This behavior results in a large window of material processing opportunities, where the BMG, in its supercooled liquid state, can be processed under pressures and temperatures similar to those used for plastics,[96] with several minutes of processing time allowed before crystallization sets in and starts degrading the alloy properties. Thus, thermoplastic forming (TPF) is carried out by reheating the BMG into its supercooled liquid region, and a relatively small forming pressure is sufficient to fill even the smallest and most complicated features while avoiding crystallization.[97] In other words, we can process the metal as if it were a polymer, as demonstrated in Figure 5.23.

Thermoplastic forming with BMGs has a strong potential as a technology for MEMS and NEMS fabrication.[99–100] The amorphous structure of a BMG makes it an ideal material for small geometries, since it lacks an intrinsic limitation like the grain size in crystalline materials or the large molecular weight of polymers. The BMG is homogeneous and isotropic down to the atomic scale. The absence of crystallization during solidification and the low processing temperature during TPF dramatically reduce solidification shrinkage to levels below 0.2% for most BMGs of technical relevance.[97] This low shrinkage, together with the fact that fast cooling is not required after TPF, allows one to precisely replicate mold features.

The supercooled liquid state of the BMG can be utilized for other processes, one of which is hot separation.[98] Thermoplastic forming into a mold cavity results in precise replication of micro features, which, however, are connected to a typically large reservoir of amorphous material. For MEMS fabrication and also for three-dimensional microparts, this reservoir must be separated from the part. Hot separation, where a scraper separates reservoir from part as shown in Figure 5.23, is carried out in the SCLR. This leaves a plain surface on which further fabrication steps can be carried out or a three-dimensional part after the mold is etched. The low-viscosity supercooled liquid state of the BMG can also be used to erase features in the BMG,[101] as shown in Figure 5.24. This is accomplished through surface tension forces alone. Such a process has potential for use as a rewrite mechanism for data storage, as well as to smooth the surface of microparts (where the finish is limited by the surface roughness of the mold).

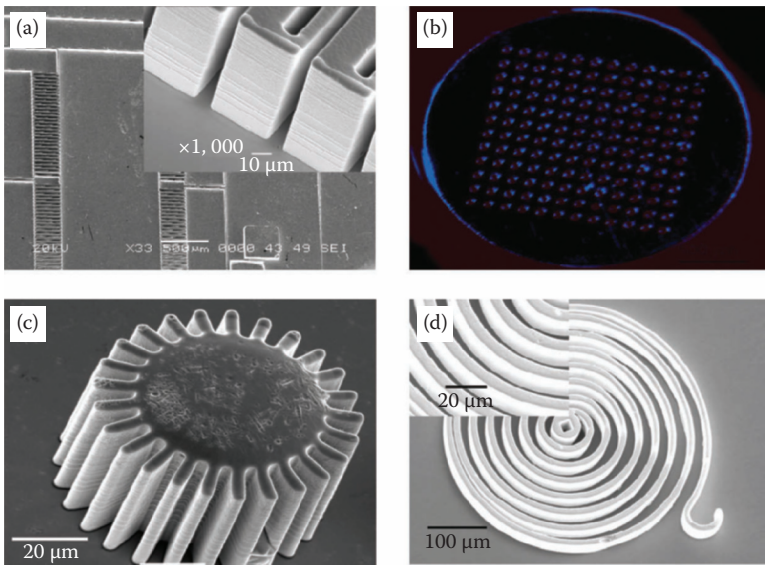


FIGURE 5.23 Examples of structures created by thermoplastic forming of bulk metallic glass (BMG). The top images depict precise replication of typical microelectromechanical system features with zirconium-based BMG (a) and a microporous membrane (b). Separation of the microparts from the BMG reservoir using a hot-separation technique results in three-dimensional microparts (c and d).[98] (Courtesy of Dr. Jan Schroers, Yale University.)

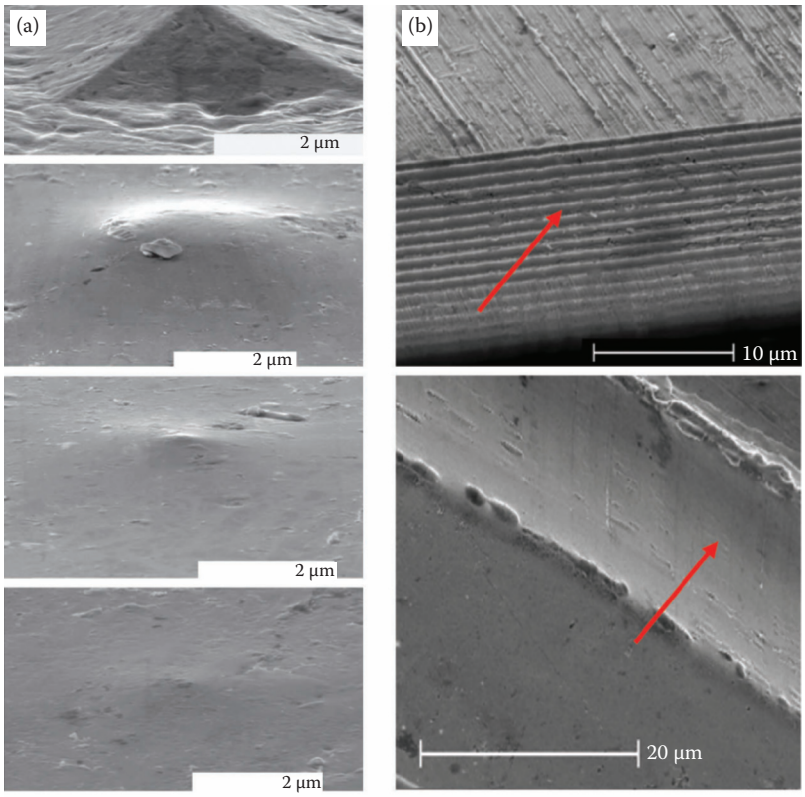


FIGURE 5.24 Forces exerted by surface tension alone can be used to erase features (a) or smoothing of the surface (b) in bulk metallic glass while processing it in the supercooled liquid region.[101]

Because of favorable chemical selectivity in BMGs, sacrificial etching and patterning can be performed without damaging the amorphous material.[98] For example, Zr- and Pt-based BMGs are resistant to most acids and bases, making them good candidates for postprocessing (selective wet-etching) of silicon after the TPF process.

Not only can BMGs be used to replicate a mold, but they can also be used as the mold material itself from which various materials are replicated. Even other BMGs can be replicated on a BMG mold, as long as they soften at lower temperatures than the mold material. BMGs typically lose their strength gradually with increasing temperature, but once the glass transition temperature is reached, where the glassy BMG relaxes into a metastable liquid, the strength significantly drops. Consequently, clever use of BMGs with decreasing T_g values allows the processing sequence depicted in Figure 5.25, where a BMG is used to replicate a conventional mold, and when processed at a lower temperature, as a mold itself.[98]

MEMS and microstructure applications are dictated by the BMG properties, processability, and process compatibility. Mechanical properties and processing temperature ranges for a few selected BMGs are summarized in Table 5.10. Their very high yield strength values, which by far exceed those of any other metal used for MEMS, are beneficial for almost every application. In particular, medical probes, skin-based drug delivery patches, sensors and actuators in high-shock environments, and microthrusters can all benefit from high strength. The high yield strength also implies a high hardness. This leads to less wear and is particularly interesting for MEMS relays and RF dielectric switches, which require robust performance after 10+ million contact cycles. The absence of dislocations also results in a very large elastic strain limit. Within the elastic region the material approaches perfect elastic behavior, reflected in low losses upon cyclic loading, which is particularly useful for resonators. Together, the high yield strength and large elastic strain limit suggest an outstanding resilience. For example, the 17-MPa resilience of a Zr-based BMG is about 1000 times higher than the resilience of nickel. Even the best conventional metallic spring material, high-carbon

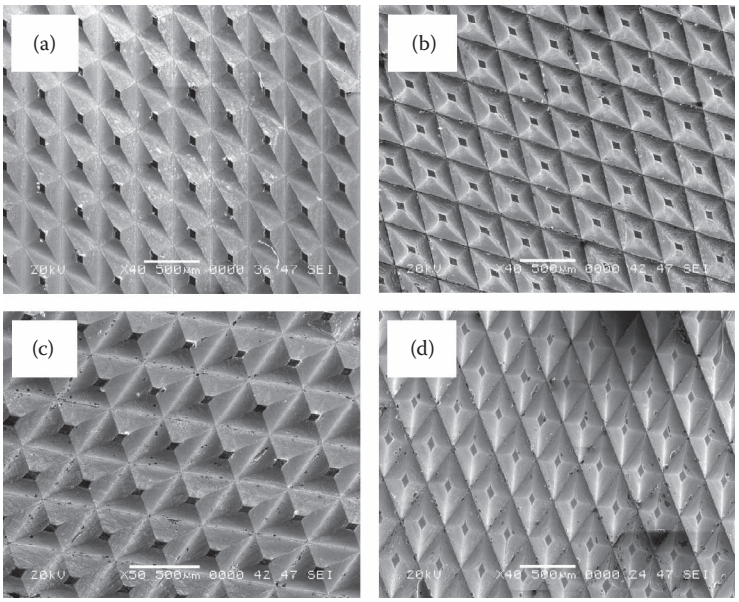


FIGURE 5.25 Sequential use of various bulk metallic glasses (BMGs) with different softening behavior as mold and molded material. The nickel master (a) is replicated at 450°C with a Zr-based BMG (b). This structure is then used as a mold for a Pt-based BMG formed at 280°C (c). A Au-based BMG with an even lower glass transition temperature of 130°C (softening behavior) is used to replicate the Pt-based BMG structure. The Au-based BMG is processed at 160°C, and perfect replication of the Pt-based BMG mold can be achieved (d).[102]

spring steel, possesses a resilience below 2 MPa. Obviously, their outstanding resilience makes BMGs excellent materials for any spring application involving energy storage and harvest, as well as for valve applications.

BMGs are available in a wide range of densities. The high density of $Pt_{57.5}Cu_{14.7}Ni_{5.3}P_{22.5}$, 15.3 g/cm^3, makes it an attractive candidate for improving the sensitivity of accelerometers.

The ability to use BMGs as a material to replicate parts on a mold and as a mold material itself suggests an alternative to the LIGA process (see Volume II, Chapter 10) for the mass production of microparts. BMG molds could supplant disposable molds.

Summarizing, BMGs are an entirely new class of material best described as an ultrahigh-strength material that can be processed like a plastic. There are many reasons why this class of materials might become an important MEMS/NEMS material:

- No intrinsic size limitation (like grain boundaries in crystalline materials)
- Perfectly isotropic
- Very large strength and strain
- Perfectly elastic (high Q factor)
- Some are very inert in most environments
- Very low processing temperature (60–400°C for various BMGs)

TABLE 5.10 Selected Properties and Processing Temperatures for BMGs with Potential for Microelectromechanical System and Microstructural Applications

BMG Material	Density (g/cm^3)	Yield Strength (MPa)	ε_{el} (%)	Y (GPa)	G (GPa)	B (GPa)	E_{el} (MPa)	ν	Thermoplastic Forming Temperature (°C)
$Au_{49}Ag_{5.5}Pd_{2.3}Cu_{26.9}Si_{16.3}$	11	1200	1.5	66.38	23.43	132.3	7.5	0.41	150–190
$Pt_{57.5}Cu_{14.7}Ni_{5.3}P_{22.5}$	15.3	1470	1.3	94.8	33.3	243	8	0.42	250–300
$Zr_{44}Ti_{11}Cu_{10}Ni_{10}Be_{25}$	6.1	1860	1.9	96.7	35.6	114	17	0.36	380–470

Source: Jan Schroers, Yale University.

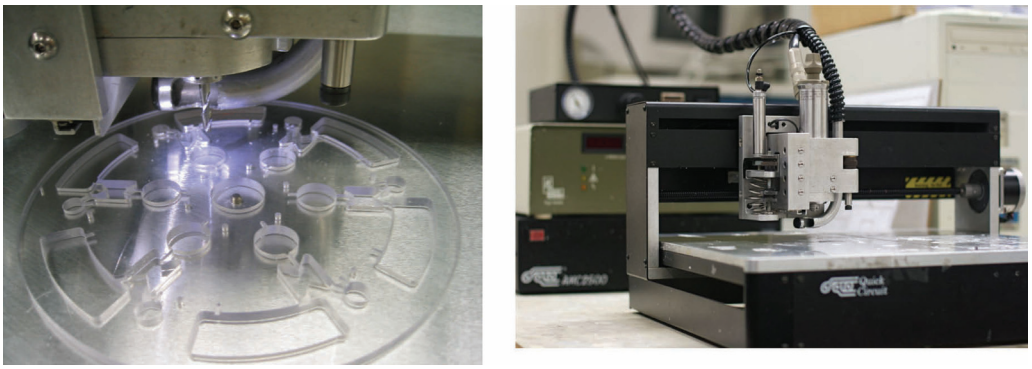


FIGURE 5.26 A computer numerically controlled milling machine. Quick Circuit 5000 machine cuts fluidic design in a polycarbonate disc.

- Very low flow stress
- Features as small as 30 nm can be replicated
- Molding process reduces internal stresses and porosity to an undetectable level

Fast Microfluidic CD Prototyping

In Volume I, Chapter 6, we analyzed CD-based microfluidics.[103] Here we detail some simple methods to implement fast prototyping of microfluidic CDs in a research laboratory setting.

Many microfluidic applications utilize relatively large volumes (on the order of 20 µL to several milliliters) that are not always compatible with typical microfabrication techniques. This holds especially true for biological applications, including sample preparation and nucleic acid analysis (see also Chapter 10, this volume, under sample preparation). Centrifugal platforms in which microfluidic CDs are used for biological analysis require these larger volumes and may feature small conduits and DNA analysis chambers, which means that a combination of macro- and microfabrication techniques must be used to create these devices.

Computer numerical control (CNC) machining can be used to create most macro features (>100 µm) on a microfluidic CD. Inexpensive plastics, such as polycarbonate (PC) or ABS, are used as substrates in which channels down to ~1 mm in width can be machined using standard endmill and drill bits (Figure 5.26, CNC machine, Quick Circuit 5000). The polymer substrate thicknesses can be as small as 200 µm. Using a layering approach, different sheets of thin machined plastic CDs can be laminated together using pressure-sensitive adhesive (PSA) layers (PSA-FLEX mount DFM 200 clear V-95). The PSA layers may be as thin as 100 µm and are cut using a CNC cutter-plotter (Figure 5.27, CNC plotter, Graphtec CE-2000-60). The use of PSA layers allows for channels as narrow as 250 µm. When sandwiched between two plastic sheets, these channels in the PSA connect the larger CNC machined reservoirs in the PC discs.

Even the simplest standard microfluidic CD consists of no fewer than five layers, as illustrated in Figure 5.28. Usually the fluidic compact discs made in our lab are composed of three optically clear polycarbonate disks held together by two pressure-

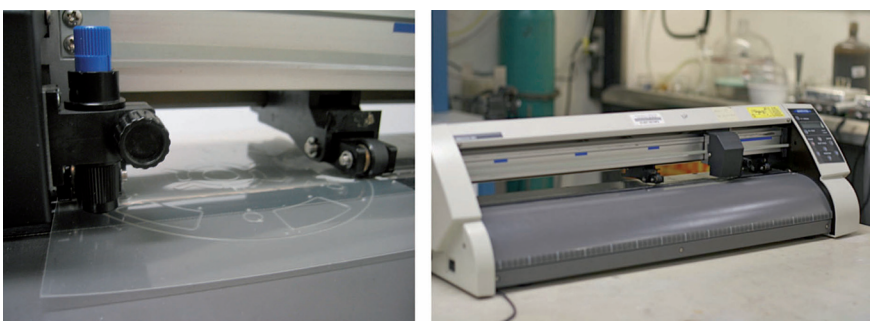


FIGURE 5.27 A roll-feed cutter plotter cuts out the pressure-sensitive adhesive layer designs. Graphtec CE-2000–60.

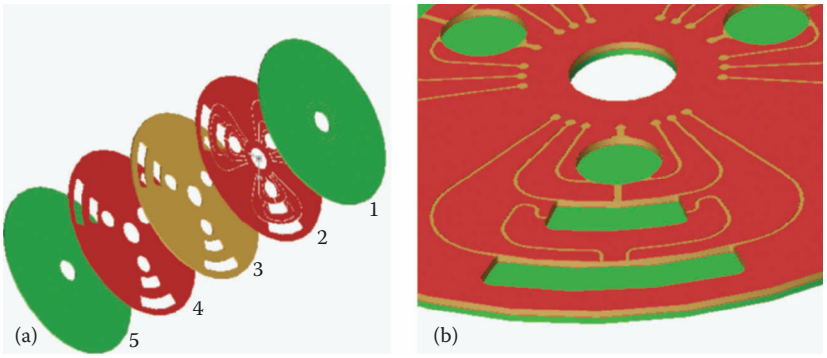


FIGURE 5.28 Computer numerical control (CNC) machined compact disc (CD) assembly. (a) Illustration showing five layers fabricated by a CNC machine and vinyl cutter: 1, top plate (1.2 mm); 2, pressure-sensitive adhesive layer (100 μm); 3, middle plate (1.2 mm); 4, pressure-sensitive adhesive layer (100 μm); 5, bottom disc (1.2 mm). (b) Schematic illustration of assembled CD structure.

sensitive adhesive layers. The top plate (including venting holes), the middle plate (containing all fluidic chambers), and the bottom plate contain the alignment holes at the same positions. Specifically, the five layers shown in Figure 5.28 are as follows: 1) top polycarbonate CD with CNC-machined sample loading, sample removal, and air venting holes; 2) pressure-sensitive adhesive with channel features cut using a plotter; 3) middle polycarbonate CD with machined channel features; 4) pressure-sensitive adhesive with channel features cut using a plotter; 5) solid bottom polycarbonate CD to seal off the channels. Microfluidic CD platforms can involve more layers to accommodate more complex fluidics. Moreover, different devices and substances can be placed inside the CD during fabrication, such as beads, lyophilized reagents, or filters. The CDs also can be exposed to O_2 plasma treatment or functionalized with bovine serum albumin to create hydrophilic and hydrophobic surfaces, respectively. Once the appropriate pieces have been designed and machined, they are aligned centrally and radially and laminated together using the PSA layers. The fabrication process ends with running the CDs through an industrial press to ensure excellent adhesion and sealing between all CD layers. After the assembled layers are pressed together, they are fed through a laminator to press the layers together tightly and ensure a good seal (Figure 5.29).

While the majority of CD platforms developed in our laboratory utilize standard macromachining processes as described above, microfabrication is easily integrated onto the CD platform. Integration usually takes the form of creating microfluidic polydimethylsiloxane (PDMS) molds on 6-in. Si wafers using multilevel, thick-resist lithography (Figure 5.30). Briefly, thick photoresist is spun on a Si wafer, prebaked to remove solvents, exposed to ultraviolet (UV) to catalyze cross-linking, postbaked to finish cross-linking, and developed to remove unexposed

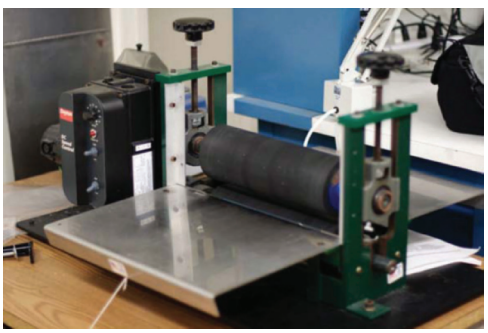


FIGURE 5.29 A laminator is used to press CNC machined discs together with pressure-sensitive adhesive layers.

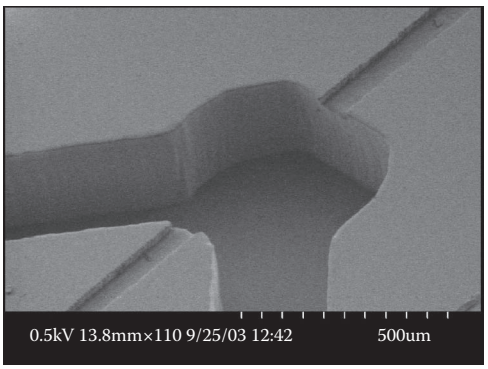


FIGURE 5.30 A scanning electron microscope image of a typical two-level polydimethylsiloxane flow cell structure. The microchannel is 25 μm deep, and the chamber is 250 μm deep.

photoresist. This creates the master mold (a negative of the desired channel features), upon which PDMS (a biocompatible elastomer) can be poured and cured. This process, termed soft lithography (see Volume II, Chapter 2), can be used to make many PDMS parts from a single Si mold.

Sample loading and venting holes can then be punched using small-gauge syringes, and the soft lithography PDMS parts (either 6-in. round entire PDMS CDs or individual microfluidic units) can be placed onto a polycarbonate CD, or in some cases bound to glass slides and secured with a specially designed slide holder CD. Most often, the passive binding properties of PDMS on glass/plastic are enough to prevent leakage during use. However, an O_2 plasma treatment process can be used to create a permanent bond with PDMS. Almost all processes and materials used to create the microfluidic CD platforms are relatively inexpensive, with the exception of the initial Si mold needed when incorporating microfabrication.

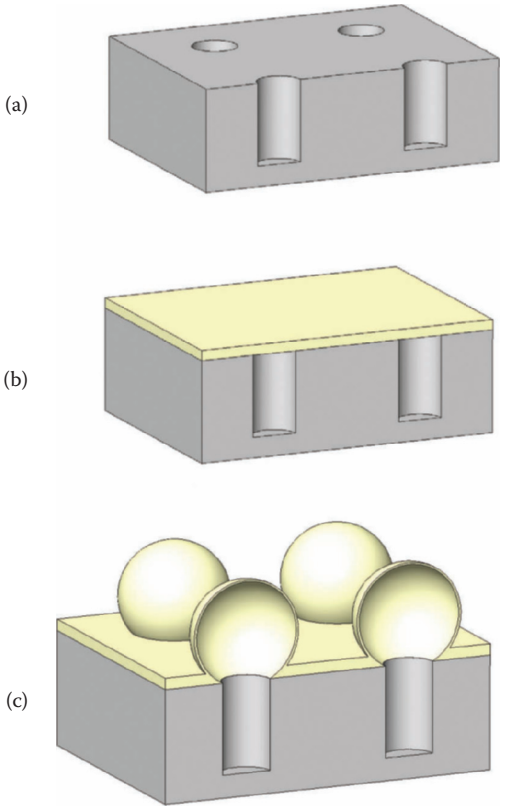


FIGURE 5.31 Micro glass blowing process sequence. (a) Step 1: etching of deep cavities in a silicon wafer. (b) Step 2: anodic bonding of glass wafer to the silicon wafer. (c) Step 3: blowing of glass spheres by heating sample inside furnace.

Micro Glass Blowing

Eklund and Shkel developed a wafer-level glass blowing process for the formation of three-dimensional glass microstructures.[104] In contrast to conventional glass blowing, where relatively large glass structures are shaped one at a time, the micro glass blowing process allows for parallel fabrication of thousands of microscopic glass parts.

The micro glass blowing fabrication process is illustrated in Figure 5.31. In a first step (Figure 5.31a), deep cylindrical cavities are dry etched in a silicon wafer. The etch mask (not shown) is removed, and a borosilicate glass wafer is anodically bonded to the silicon wafer (Figure 5.31b). Next, the parts are rapidly heated inside a furnace to a temperature above the softening point of the glass. This elevated temperature causes the pressure of the gas trapped inside the sealed cavities to increase. The glass deforms into spherical shapes because of the combination of the increased gas pressure and the low viscosity of the heated glass (Figure 5.31c).

The size of the blown glass spheres depends on the dimensions of the etched cavity, the furnace temperature, and the pressure and temperature at which the cavities are sealed. The total volume of the etched cavity and the blown glass sphere is proportional to the volume of the etched cavity, with the proportionality constant being the ratio between the final and initial temperatures. For example, if a cavity is sealed in atmospheric pressure and at room temperature and the furnace temperature is 900°C, the final volume is approximately four times as great as the volume of the etched cavity (1173 K vs. 295 K). The volume enclosed by only the glass sphere is therefore three times as great as the volume of the etched cavity.

In Figure 5.32 we show fabricated glass spheres, all with a diameter of approximately 900 µm. A layer of AZ P4620 photoresist was used as the etch mask during the deep reactive ion etching of the cylindrical cavities. For the sample shown in Figure 5.32a, a silicon wafer with a thickness of 450 µm was used, and the etched cavities were 350 µm deep and 750 µm in diameter. For the parts in Figure 5.32b and 5.32c, cavities with a depth of 750 µm and a

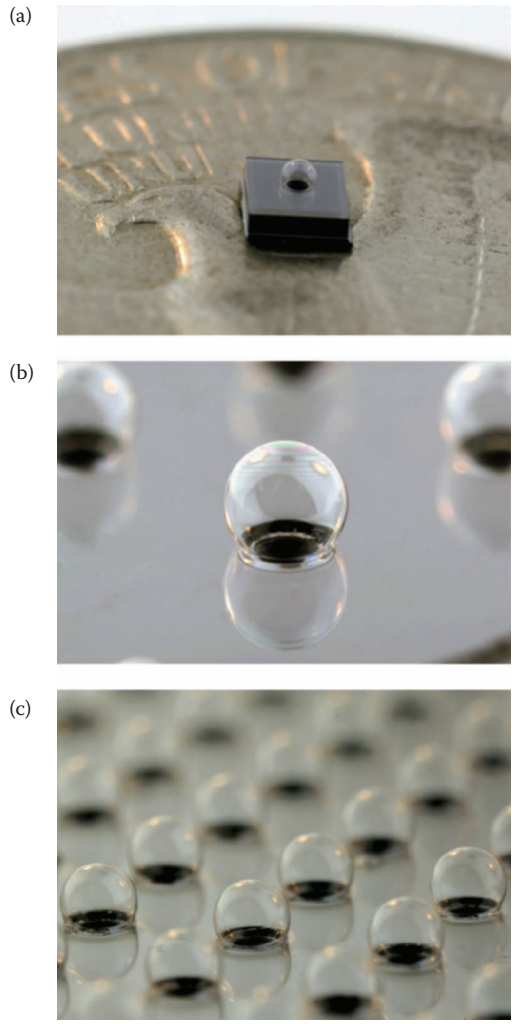


FIGURE 5.32 Fabricated glass structures. (a) Diced chip on US quarter coin. (b) Close-up of fabricated glass sphere. (c) Array of fabricated glass spheres.

diameter of 500 µm were etched in silicon wafers with a thickness of 1 mm. Pyrex 7740 glass wafers with an initial thickness of 100 µm were used for the anodic bonding of all samples, and the spheres were formed inside a furnace set to 850°C. This technology may be utilized for several different types of MEMS devices, including microfabricated glass lenses, three-dimensional microfluidic networks, and spherical gas confinement chambers for atomic MEMS.[105]

Biocompatibility and Surface Modification

Definition

Before new materials or devices can be used in conjunction with a biological system, whether cells or organisms, their effect on that system must be tested. In the simplest terms, a substance or device that has no effect on the time scale tested is said to be *biocompatible*. This can be verified at first by observing whether the material or device under study forms a suitable substrate for the growth of cultured cells. Such preliminary screening eliminates materials that have direct, harmful effects. To fully test the biocompatibility of a new material or device, one needs to test it on a wide range of different types of cells; for applications in humans, it also has to be tested on animals for any less direct effects. Biocompatibility is obviously a complex phenomenon. Some materials and devices are immediately toxic to all cells or organisms; some are toxic only to specific animals; others are harmful in the long term (e.g., carcinogens); others, while not toxic, may cause marked changes in the normal behavior of the system.[106] The definition of biocompatibility of a material or device also varies depending on how the material or device is to be used. If a device is to be implanted into a tissue, one may want the cells of the tissue to treat it as part of the tissue to stick to it and grow on it (not an advisable strategy for a chemical-type sensor where membranes might become blocked). But if a device is injected into the bloodstream, it is vital that the blood cells not stick to it, as this results in a clot, which could be fatal to the biological system. In the case of a medical nanorobot (perhaps a future version of a pharmacy-on-a-chip shown in Volume II, Figure 4.100) floating or swimming in the circulation, cell and tissue adhesion is again undesirable. The term *biocompatibility* evolved over time, and today an accepted working definition is the ability of a material or device to perform with an appropriate host response in a specific application.

History

Biocompatibility work has a long history. Two thousand years ago, gold was already used in dental applications; glass eyes and wooden teeth were introduced next, and plastics were developed in the early 1900s. After World War II, parachute cloth was used as vascular prosthesis. Finally, over the last two decades, active implants (implants providing more than a mechanical support function), such as

biosensors and drug delivery systems, have further challenged our understanding of biocompatibility issues. Indeed, in many of these active devices an efficient transport of chemicals to and from the device is essential, something that was not an issue for more traditional implants. Clogging of the pores or membranes in such active devices will render the device useless, though the device may remain perfectly harmless to the functioning of the biological system itself. Over the last fifteen years, considerable work has been directed to create nano- and microtextured surfaces that are cell-friendly and promote controlled cell growth. This promises to become another important application area for MEMS. One envisions substrates for cell culturing, the growth of nerves for neural implants, textures that encourage neural regeneration and the acceptance of prostheses, and microsubstrates for bioartificial organs and as a medium for the regeneration of tissues such as bone and tendon. Typical materials used or proposed, with or without nano- or microtexturing, include titanium, niobium, carbon, some ceramics, polyimide, some polyurethanes, Teflon, Parylene, Makrolon, polystyrene, PMMA, PEEK, polycarbonate, polypropylene, and silicone. Often, there are conflicting views (remember the breast implant controversy), and in vivo sensor applications remain so elusive that many companies have abandoned implantable devices all together and are focusing on ex vivo systems instead.[107] The field where perhaps the most progress has been made in terms of biocompatible materials evaluation is in the area of biomedical materials for drug delivery systems (see, e.g., Chapter 8, this volume, Table 8.20).[108–110] The biocompatibility of silicon was discussed in Volume I, Chapter 4.

Mechanism

A foreign material implanted in a host tissue causes a cascade of events to occur at the tissue/material interface. It starts with a noncovalent adsorption of plasma protein from blood onto the surface. This protein adsorption is quite a bit faster than the transport of host cells to the foreign surface. By the late 1960s, several researchers already knew that within 10 s of exposure to blood or plasma, a uniform ~6-nm layer of fibrinogen forms on surfaces such as Si, Pt, Ge, and Ta. Fibrinogen is a 340-kilodalton soluble plasma glycoprotein about 47.5 nm in length. It is the major surface protein. The protein-coated surface, in turn, mediates the types of cells that may adhere to the surface, which ultimately determines the type of tissue that forms in the vicinity of the implant. The type and state of adsorbed proteins, including their conformational changes, initiate coagulation via platelet adhesion, and subsequent collagen deposition leads to the encapsulation of the microdevice by fibrous tissue. Excessive fibrosis may cause tissue necrosis, granulomas, or tumor genesis.[111] The adsorption and denaturation of adsorbed fibrinogen molecules actually is commonly used as a biocompatibility indicator; the amounts of denatured fibrinogen accumulated on a surface correlates closely to the extent of biomaterial-mediated inflammation. Adsorbed amounts of denatured fibrinogen on surgical implants that are regarded "biocompatible" (such as titanium and stainless steel) are about ~2.1 mg/m^2 (the same as for CVD diamond).

There are two important viewpoints from which to approach biocompatibility:

1. The implant perspective: the protection of implanted devices against the corrosive effects of body fluids and tissue, their sterilizability, and continued operation.
2. The biological system perspective: the avoidance of blood clotting and long-term cytotoxicity, and the continued health of the host.

The former has been named *sensocompatibility*.[107] An important aspect of senscompatibility is biofouling, the adhesion of proteins and other organic matter on the sensor surface. As pointed out above, biofouling is a special concern for active implants such as biosensors (see, e.g., Volume II, Figure 4.92 for an in vivo pH, CO_2, and O_2 sensor), where an efficient transport of chemicals to and from the device is essential. Immune rejection during organ transplants may be prevented by pharmacologically depressing the immune system by the administration of cyclosporine. Such a strategy is, of course, not applicable for the long-term biocompatibility of implants, whether they are biosensors

or cell-transplant-based therapeutics—for example, Langerhans or pancreatic islets for type I diabetes,[112] cochlear implants, or heart pacemakers. Perhaps a pharmacy-on-the-chip, as shown in Volume II, Figure 4.100, could also monitor for biofouling and release anti-clotting chemicals such as heparin or herudin as required to maintain the function of the sensor. It should be kept in mind that biofouling occurs both for invasive and noninvasive sensors, but obviously intervention is easier with the latter. In Figure 5.33, we show a schematic of an amperometric invasive glucose sensor, perhaps the most widely studied and most "successful" implantable sensor to date.[107] It is assumed that the needle-type glucose sensor is embedded in tissue and in contact with blood (for details on the operation of amperometric glucose sensors, see Chapter 1, this volume).[113] Failure modes (many packaging related) include the following:

- Lead detachment (could be avoided with a telemetric device)
- Electrical short
- Membrane delamination
- Membrane biofouling
- Electrode passivation
- Enzyme degradation
- Membrane biodegradation
- Fibrous encapsulation
- Electrode fouling

Surface Modifications

Surface Modifications for Biocompatibility

Most success today for control of protein and cell adhesion on surfaces has been achieved by surface immobilization of polyethylene oxide or polyethylene glycol (PEG; one refers to a "pegylated surface") forming a water-soluble, nontoxic, and nonimmunogenic polymer film.[114,115] Zhang et al. modified silicon surfaces by covalent attachment of a self-assembled PEG film and demonstrated an effective suppression of both plasma protein adsorption and cell attachment to such modified silicon surfaces.[116] A pegylated surface may cut the amount of protein adsorbed down by a factor of ten. Glow-discharge plasma-deposited tetraethylene glycol dimethyl ether can reduce the fibrinogen adsorption to ~0.2 mg/m^2 (~350 molecules/μm^2) on many different types of substrates. A specific strategy to fabricate biocompatible ion-selective electrodes of the type shown in Volume I, Figure 7.45, implemented by Brooks et al.,[117] involves covalent attachment of heparin to the surface of derivatized cellulose triacetate membranes, which are subsequently impregnated with the potassium-selective ionophore valinomycin. The resulting ion-selective electrodes respond to potassium and have selectivity coefficients on the same order of magnitude as those of conventional poly(vinyl chloride)-based electrodes.

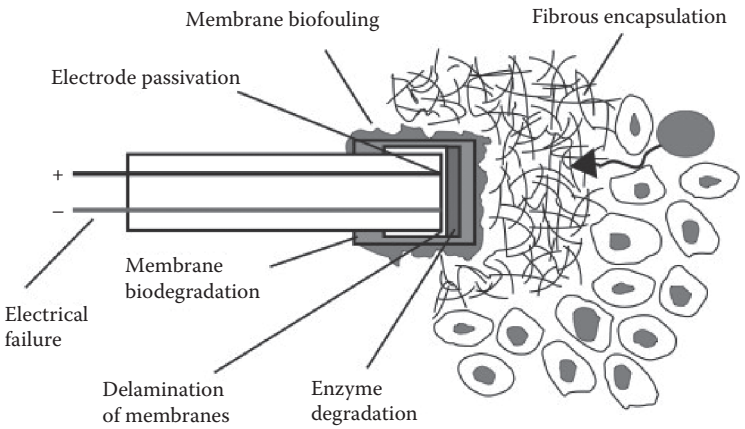


FIGURE 5.33 Schematic illustration of a blood-borne analyte exiting capillary and traversing to a needle-type glucose biosensor embedded in tissue. In addition to normal component failure—electrical failure, enzyme degradation, and membrane delamination—the sensor can fail for several physiologically related reasons, such as membrane biodegradation, electrode passivation, and reduction of analyte access due to fibrous encapsulation and membrane biofouling. (Based on Wisniewski, N., F. Moussy, and W. M. Reichert. 2000. Characterization of implantable biosensor membrane biofouling. *Fresenius J Anal Chem* 366:611–21.[107])

It was found that the heparin layer does not alter significantly the response characteristics of the electrodes. The biological activity of the immobilized heparin was measured in terms of its inactivation of blood coagulation factor Xa; the covalently anchored heparin was able to inactivate factor Xa.

Other approaches that have worked well to avoid protein adsorption are based on biomimetics. This involves the development of membranes that mimic the surface of biological membranes, that is, the design of "stealth" materials. The external region of a cell membrane, the glycocalyx, is studded with glycosylated molecules, which direct specific interactions such as cell-cell recognition and contribute to the steric repulsion that prevents undesirable nonspecific binding of other molecules or cells. By modifying a pyrolitic graphite surface with oligosaccharide surfactant polymers, which, like glycocalyx, provide a dense and confluent layer that mimics the nonadhesive properties of a cell layer, spontaneous adsorption on diverse hydrophobic surfaces is suppressed by at least ~90%.[118] The highly hydrated dextran in these layers protrudes into the aqueous phase and suppresses nonspecific adsorption of plasma proteins. Coating surfaces with phosphorylcholine groups to mimic red blood cell surfaces also leads to diminished protein adsorption. One such material is poly(methacryloyloxyethyl phosphorylcholine-co-*n*-butylmethacrylate) (see Figure 5.34). This polymer, which behaves as a hydrogel, shows very low protein and cell adsorption. Studies have revealed that sensors coated with this hydrogel demonstrate excellent sensor response characteristics, while scanning electron microscopy (SEM) of the surface of the sensors indicates no platelet adhesion. The antifouling characteristics are again believed to be due to the extreme hydrophilic nature of the surface layer; proteins have difficulty adsorbing on the surface because of the layer of bound water.[119] For a further review of in vivo biosensors, consult Frasier;[120] for surface modification strategies, see Wisniewski and Reichert.[121]

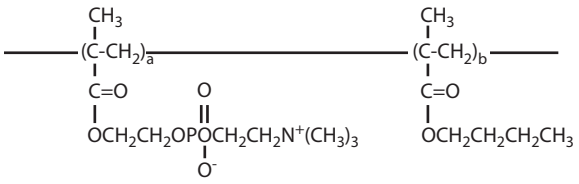


FIGURE 5.34 Structure of poly(methacryloyloxyethyl phosphorylcholine-co-*n*-butylmethacrylate).

Questions

Questions by Dr. Marc Madou assisted by Mr. Omid Rohani, UC Irvine

5.1: What properties make silicon a good mechanical sensor/actuator substrate?

5.2: What has prevented Si-MEMS oscillators from becoming a serious alternative to quartz so far?

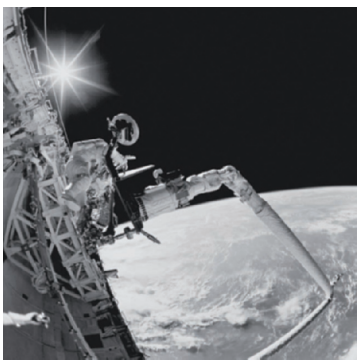
5.3: What are the possible applications of GaAs in general and GaAs MEMS in particular?

5.4: What are the two most suitable diamond variations for MEMS application? Explain briefly.

5.5: What method would you suggest for patterning convex hollow structures?

5.6: What are the main advantages of ceramics MEMS compared to other materials?

5.7: NASA's space shuttle has a long manipulator robot arm, also known as the Shuttle Remote Manipulator System or SRMS (see figure below), that permits astronauts to launch and retrieve satellites. It is also used to view and monitor the outside of the space shuttle using a mounted video camera. Select a suitable material for the arm.



5.8: List some of the applications of nonphotosensitive and photosensitive polyimides in microtechnology.

5.9: How does an increase in Si content in silicon nitride change the following parameters?
(a) Tensile film stress
(b) Film transparency
(c) HF etch rate

5.10: In the micro glass blowing method, what are the parameters on which the sizes of blown glass spheres depend?

5.11: Describe the merits and problems associated with micromachining in GaAs, Si, diamond, and 3C-SiC.

5.12: (a) Detail the C-MEMS process.
(b) What advantages might a C-MEMS world have compared to a Si-world?

5.13: What are bulk metallic glasses (BMGs)? What are their properties? How would you use them in MEMS and NEMS?

5.14: Present a definition of biocompatibility of a MEMS structure.

5.15: A thermistor is a device used to measure temperature by taking advantage of the change in electrical conductivity when the temperature changes. Select a material that might serve as a thermistor in the 500 to 1000°C temperature range.

5.16: Describe the ionic bonding between magnesium and chlorine. What are the consequences for the projected strength of the compound that forms and the electronic/ionic conductivity of such a compound?

5.17: Silica (SiO_2) is often given as an example of a covalently bonded material. In reality, silica exhibits both ionic and covalent bonding. What fraction of the bonding is covalent? Give some example applications in which silica is used. To answer this question first find a figure that plots electronegativity vs. atom type.

5.18: Silica is used for making long lengths of optical fibers (see figure below). Being a covalently and ionically bonded material, the strength of the Si-O bonds is expected to be high. Other factors such as susceptibility of silica surfaces to react with water vapor in atmosphere have a deleterious effect on the strength of silica fibers. Given this, what design strategies can you think of such that silica fibers could still be bent to a considerable degree without breaking?

5.19: Define each of the following terms: stress, strain, elastic modulus, Young's modulus of typical materials (ceramics, glasses, semiconductors, metals, and polymers), Poisson's ratio, and yield strength.

5.20: Why are p-type piezoresistors most commonly used in Si pressure sensors?

5.21: In the context of using piezoresistivity to sense strain, what are the advantages/disadvantages of using silicon versus silicon nitride for pressure sensor diaphragms? How does residual stress affect the sensitivity?

References

1. Senturia, S. D., and R. T. Howe. 1990. Mechanical properties and CAD *Lecture notes.* Boston: Massachusetts Institute of Technology.
2. Studer, B., and W. Zingg. 1990. In *4th European Frequency and Time Forum*. Neuchatel, Switzerland, pp. 635–58.
3. Tellier, C. R., and F. Jouffroy. 1983. Orientation effects in chemical etching of quartz plates. *J Mater Sci* 18:3621–32.
4. Ueda, T., F. Kohsaka, T. Lino, and D. Yamazoki. 1987. Theory to predict etching shapes in quartz and applications to design devices. *Trans Soc Inst Control Eng* 23: 1–6.
5. Chuang, S. S. 1983. In *Proceedings: 37th Annual Frequency Control Symposium 1983*, 248–54. Philadelphia: IEEE.
6. Dufour, M., M. T. Delaye, F. Michel, J. S. Danel, B. Diem, and G. Delapierre. 1991. A comparison between micromachined pressure sensors using quartz or silicon vibrating beams. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, June 24–27, 1991, pp. 668–71.
7. Editorial. 1994. Concord, CA: BEI Systron Donner.
8. Hjort, K., G. Thornell, R. Spohr, and J.-Å. Schweitz. 1996. Heavy ion induced etch anisotropy in single crystalline quartz. *Ninth Annual International Workshop on Micro Electro Mechanical Systems*. San Diego, February 11–15, 1996, pp. 267–71.

9. Ericson, F., S. Johansson, and J.-Å. Schweitz. 1988. Hardness and fracture toughness of semiconducting materials studied by indentation and erosion techniques. *J Mater Sci Eng* A105/106:131–41.
10. Takebe, T., T. Yamamoto, M. Fujii, and K. Kobayashi. 1993. Fundamental selective etching characteristics of $HF+H_2O_2+H_2O$ mixtures for GaAs. *J Electrochem Soc* 140:1169–80.
11. Zhang, Z. L., and N. C. MacDonald. 1993. Fabrication of submicron high-aspect-ratio GaAs actuators. *J Microelectromech Syst* 2:66–73.
12. Tarui, Y., Y. Komiya, and Y. Harada. 1971. Preferential etching and etched profile of GaAs. *J Electrochem Soc* 118:119–122.
13. Ribas, R. P. 1998. *Maskless front-side bulk micromachining compatible to standard GaAs IC technology*. Grenoble, France: TIMA Laboratory.
14. Hjort, K. 1996. Sacrificial etching of III-V compounds for micromechanical devices. *J Micromechanics Microeng* 6:370–375.
15. Collins, S. D. 1997. Etch stop techniques for micromachining. *J Electrochem Soc* 144:2242–2262.
16. Karam, J. M., B. Courtois, M. Holjo, J. L. Leclercq, and P. Viktorotovitch. 1996. Collective fabrication of gallium arsenide based microsystems. *SPIE: Micromachining and Microfabrication Process Technology II*. Austin, TX, October 14–15, 1996, pp. 315–26.
17. Hjort, K., J.-Å. Schweitz, and B. Hok. 1990. In *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '90)*. Napa Valley, CA, pp. 73–76.
18. Hjort, K., J.-Å. Schweitz, S. Andersson, O. Kordina, and E. Janzen. 1992. In *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '92)*. Travemunde, Germany, February 4–7, 1992, 83–86.
19. Morkoc, H., H. Unlu, H. Zabel, and N. Otsuka. 1988. Gallium arsenide on silicon: A review. *Solid State Technol* March: 71–76.
20. Ramdani, J., R. Droopad, Z. Yu, J. A. Curless, C. D. Overgaard, J. Finder, K. Eisenbeiser, J. A. Hallmark, W. J. Ooms, V. Kaushik, P. Alluri, and S. Pietambaram. 2000. Interface characterization of high-quality SrTiO3 thin films on Si(100) substrates grown by molecular beam epitaxy. *Applied Surface Science* 159–160:127–133.
21. Singer, P. 2001. GaAs-on-silicon, finally! *Semiconductor International* October.
22. LaPedus, M., and P. Clarke. 2003. Motorola closes GaAs-on-silicon wafer subsidiary. *EETimes*.
23. Yeh, H. J., and J. S. Smith. 1994. Fluidic self-assembly of microstructures and its application to the integration of GaAs on Si. *IEEE International Workshop on Micro Electro Mechanical Systems (MEMS '94)*. Oiso, Japan, January 25–28, 1994, pp. 279–84.
24. Yeh, H. J., and J. S. Smith. 1994. Integration of GaAs vertical-cavity surface-emitting laser on Si by substrate removal. *Appl Phys Lett* 64:1466–68.
25. Teraji, T., S. Mitani, C. Wang, and T. Ito. 2002. High-rate deposition of high-quality homoepitaxial diamond films (review). *New Diamond Front Carbon Tech* 12:355.
26. Wang, C., M. Irie, K. Kimura, T. Teraji, and T. Ito. 2001. Boron-doped diamond film homoepitaxially grown on high-quality chemical-vapor deposited diamond (100). *Jpn J Appl Phys* 40:4145.
27. Herb, J. A., M. G. Peters, S. C. Terry, and J. H. Jerman. 1990. PECVD diamond films for use in silicon microstructures. *Sensors Actuators A* A23:982–87.
28. Spear, K. E., and J. P. Dismukes. 1994. *Synthetic diamond: Emerging CVD science and technology*. New York: Wiley-Interscience.
29. Obermeier, E. 1995. In *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, pp. 178–81.
30. Auciello, O., J. Birrell, J. A. Carlisle, J. E. Gerbi, X. Xiao, B. Peng, and H. D. Espinosa. 2004. Materials science and fabrication processes for a new MEMS technology based on ultrananocrystalline diamond thin films. *J Phys Condens Matter* 16:R539–52.
31. Webster, J. R., C. W. Dyck, J. P. Sullivan, T. A. Friedmann, and A. J. Carton. 2004. Performance of amorphous diamond RF MEMS capacitive switch. *Electron Lett* 40:43–44.
32. Kohn, E., P. Gluche, and M. Adamschik. 1999. Diamond MEMS: A new emerging technology. *Diamond Relat Mater* 8:934–40.
33. Irie, M., S. Endo, C. L. Wang, and T. Ito. 2003. Fabrication and properties of lateral p-i-p structures using single-crystalline CVD diamond layers for high electric field applications. *Diamond Relat Mater* 12:1563–68.
34. Lari, T., A. Oh, N. Wermes, H. Kagan, M. Keil, and W. Trischuk. 2005. Characterization and modeling of non-uniform charge collection in CVD diamond pixel detectors. *Nuclear Instruments Methods Phys Res Sec A Accel Spectrom Detect Associated Equip* 537:581–93.
35. Muller, R., M. Adamschik, D. Steidl, E. Kohn, S. Thamasett, S. Stiller, H. Hanke, and V. Hombach. 2004. Application of CVD-diamond for catheter ablation in the heart. *Diamond Relat Mater* 13:1080–83.
36. Achard, J., F. Silva, H. Scheider, R. S. Sussmann, A. Tallaire, A. Gicquel, and M. C. Castex. 2004. The use of CVD diamond for high-power switching using electron beam excitation. *Diamond Relat Mater* 13:876–80.
37. Krotz, G., W. Legner, C. Wagner, H. Moller, H. Sonntag, and G. Muller. 1995. Silicon carbide as a mechanical material. *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, 186–89.
38. Fleischman, A. J., S. Roy, C. A. Zorman, M. Mehregany, and L. G. Matus. 1996. Polycrystalline silicon carbide for surface micromachining. *9th Annual International Workshop on Micro Electro Mechanical Systems*. San Diego, CA, February 11–15, 1996, pp. 234–38.
39. Roy, S., A. K. McIlwain, R. G. DeAnna, A. J. Fleischman, R. K. Burla, C. A. Zorman, and M. Mehregany. 2000. In *Solid-State Sensor and Actuator Workshop 22–25*. Hilton Head Island, SC: Transducers Research Foundation.
40. Mardilovich, P., D. Routkevitch, and A. Govyadinov. 2000. Hybrid micromachining and surface microstructuring of alumina ceramic. *Electrochem Soc Proc* 200-19:33–42.
41. Tan, S., M. Reed, H. Han, and R. Boudreu. 1995. High aspect ratio microstructures on porous anodic aluminum oxide. *Proceedings of the Eighth International Workshop on Micro Electro Mechanical Systems (MEMS-95)*. Amsterdam, January 1995, pp. 267–72.
42. Routkevitch, D., A. Govyadinov, and P. Mardilovich. 2000. High aspect ratio, high resolution ceramic MEMS. *2000 International Mechanical Engineering Congress and Exposition*. Orlando, FL, November 5–10, 2000, pp. 1–6.
43. Li, A., F. Muller, A. Birner, K. Nielsch, and U. Gosele. 1999. Fabrication and microstructuring of hexagonally ordered two-dimensional nanopore arrays in anodic alumina. *Adv Materials* 11:483–87.

44. Govyadinov, A., P. Mardilovich, K. Novogradecz, S. Hooker, and D. Routkevitch. 2000. Anodic alumina MEMS: Applications and devices. In *2000 International Mechanical Engineering Congress and Exposition*, Orlando, FL, November 5–10, 2000, pp. 1–6.
45. Mardilovich, P., D. Routkevitch, and A. Govyadinov. 2000. New approach for surface microstructuring. In *2000 International Mechanical Engineering Congress and Exposition*. Orlando, FL, November 5–10, 2000, pp. 1–5.
46. Govyadinov, A., P. Mardilovich, and D. Routkevitch. 2000. Field emission cathode from aligned arrays of carbon nanotubes. *Electrochem Soc Abstracts* 2000-2:Abstract 553.
47. Sugiura, T., T. Yoshida, and H. Minoura. 1998. Designing a TiO_2 nano-rhoneycomb structure using photoelectrochemical etching. *Electrochem Solid-State Lett* 1:175–77.
48. Sugiura, T., S. Itoh, T. Ooi, T. Yoshida, K. Kuroda, and H. Minoura. 1999. Evolution of a skeleton structures TiO_2 surface consisting of grain boundaries. *Electroanal Chem* 473:204–08.
49. Schubert, P. J., and J. H. Nevin. 1985. A polyimide-based capacitive humidity sensor. *IEEE Trans Electron Devices* ED-32:1220–24.
50. Ralston, A. R. K., C. F. Klein, P. E. Thoma, and D. D. Denton. 1995. A model for the relative environmental stability of a series of polyimide capacitance humidity sensors. *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, pp. 821–24.
51. Latham, W. J., and D. W. Hawley. 1988. Color filters from dyed polyimides. *Solid State Technol* 31:223–26.
52. Suh, J. W., S. F. Glander, R. B. Darling, C. W. Storment, and G. T. A. Kovacs. 1996. Combined organic thermal and electrostatic omnidirectional ciliary microactuator array for object positioning and inspection. *Proceedings: Solid-State Sensor and Actuator Workshop*. Hilton Head, SC, June 2–6, 1996, pp. 168–73.
53. Darling, R. B., J. W. Suh, and G. T. Kovacs. 1998. Ciliary microactuator array for scanning electron microscope positioning stage. *J Vac Sci Technol A* 16:1998–2002.
54. Makino, D. 1994. In *Polymers for microelectronics: Resists and dialectrics* (eds. Thompson, L. F., C. G. Willson, and S. Tagawa) Washington, DC: American Chemical Society.
55. Studt, T. 1992. Polyimides: Hot stuff for the '90s. *Res Dev* August: 30–31.
56. Ando, S., T. Matsuura, and S. Sasaki. 1994. Polymers for microelectronics: Resists and dielectrics, ed. L. F. Thompson, C. G. Willson, and S. Tagawa. *ACS Symposium Series* 537:304–22.
57. Pool, R. 1988. Microscopic motor is a first step. *Res News* October: 379–80.
58. Rosler, R. S. 1991. The evolution of commercial plasma enhanced CVD systems. *Solid State Technol* June: 67–71.
59. Wu, T. H. T., and R. S. Rosler. 1992. Stress in PSG and nitride films as related to film properties and annealing. *Solid State Technol* May: 65–71.
60. Adams, A. C. 1988. In *VLSI technology,* ed. S. M. Sze. New York: McGraw-Hill, pp. 259–60.
61. Sinha, A. K., and T. E. Smith. 1978. Thermal stresses and cracking resistance of dielectric films. *J Appl Phys* 49:2423–26.
62. Retajczyk, T. F. J., and A. K. Sinha. 1980. Elastic stiffness and thermal expansion coefficients of various refractory silicides and silicon nitride films. *Thin Solid Films* 70:241–47.
63. Sakimoto, M., H. Yoshihara, and T. Ohkubo. 1982. Silicon nitride single-layer x-ray mask. *J Vac Sci Technol* 21:1017–21.
64. Ibid.
65. Chang, S., W. Eaton, J. Fulmer, C. Gonzalez, B. Underwood, J. Wong, and R. L. Smith. 1991. Micromechanical structures in amorphous silicon. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, June 24–27, 1991, pp. 751–54.
66. Spear, W. E., and P. G. Le Comber. 1975. Substitutional doping of amorphous silicon. *Solid State Commun* 17:1193–96.
67. Lee, J. B., Z. Chen, M. G. Allen, A. Rohatgi, and R. Arya. 1995. A miniaturized high-voltage solar cell array as an electrostatic MEMS power supply. *J Microelectromech Syst* 4:102–08.
68. Crowley, J. L. 1992. Plasma enhanced CVD for flat panel displays. *Solid State Technol* February: 94–98.
69. Holbrook, D. S., and J. D. McKibben. 1992. Microlithography for large area flat panel display substrates. *Solid State Technol* May: 166–72.
70. Bohm, M. 1988. Advances in amorphous silicon based thin film microelectronics. *Solid State Technol* September.
71. Bruno, G., P. Capezzuto, and A. Madan, Eds. 1995. *Plasma deposition of amorphous silicon-based materials.* Boston: Academic Press.
72. Wang, C., and M. Madou. 2005. From MEMS to NEMs with carbon. *Biosensors and Bioelectronics* 20:2181–87.
73. Wang, C., G. Jia, L. H. Taherabadi, and M. J. Madou. 2005. A novel method for the fabrication of high-aspect ratio C-MEMS structures. *J MEMS* 14:348–58.
74. Schueller, O. J. A., S. T. Brittain, C. Marzolin, and G. M. Whitesides. 1997. Fabrication and characterization of glassy carbon MEMS. *Chem Mater* 9:125–39.
75. Madou, M., A. Lai, G. Schmidt, X. Song, K. Kinoshita, M. Fendorf, A. Zetl, and R. White. 1997. Carbon micromachining (CMEMS). In *Chemical and biological sensors and analytical electrochemical methods*, ed. A. Ricco, M. B., P. Vanysck, G. Horvai, and A. Silva. Pennington, NJ: Electrochemical Society Proceedings Series, pp. 61.
76. Kim, J., X. Song, K. Kinoshita, M. Madou, and R. White. 1998. Electrochemical studies of carbon films for pyrolyzed photoresist. *J Electrochem Soc* 145:2314–19.
77. Ranganathan, S., R. McCreery, S. M. Majji, and M. Madou. 2000. Photoresist derived carbon for microelectromechanical systems and electrochemical applications. *J Electrochem Soc* 147:277–82.
78. Wang, C., L. Taherabadi, G. Jia, M. Madou, Y. Yeh, and B. Dunn. 2004. C-MEMS for the manufacture of 3D microbatteries. *Electrochem Solid-State Lett* 7:A435–38.
79. Hart, R. W., H. S. White, B. Dunn, and D. R. Rolison. 2003. 3D microbatteries. *Electrochem Commun* 5:120–23.
80. McCreery, R., J. Dieringer, A. O. Solak, B. Snyder, A. M. Nowak, W. R. McGovern, and S. Duvall. 2003. Molecular rectification and conductance switching in carbon-based molecular junctions by structural rearrangement accompanying electron injection. *JACS* 125:10748–58.
81. Hebert, N. E., B. Snyder, R. L. McCreery, W. G. Kuhr, and S. A. Brazill. 2003. Performance of pyrolyzed photoresist carbon films in a microchip capillary electrophoresis device with sinusoidal voltammetric detection. *Anal Chem* 75:4265–71.
82. Park, B. Y., R. Zaouk, C. Wang, and M. J. Madou. 2007. A case for fractal electrodes in electrochemical applications. *J Electrochem Soc* 154:1–5.

83. Park, B. Y., and M. J. Madou. 2005. 3-D electrode design for flow-through dielectrophoretic systems. *Electrophoresis* 26:3745.
84. Inoue, A., T. Zhang, and T. Masumoto. 1990. Zr-Al-Ni amorphous-alloys with high glass-transition temperature and significant supercooled liquid region. *Materials Transactions Jim* 31:177–83.
85. Peker, A., and W. L. Johnson. 1993. A highly processable metallic-glass: Zr41.2ti13.8cu12.5ni10.0be22.5. *Appl Phys Lett* 63:2342–44.
86. Ponnambalam, V., S. J. Poon, and G. J. Shiflet. 2004. Fe-Mn-Cr-Mo-(Y,Ln)-C-B (Ln = lanthanides) bulk metallic glasses as formable amorphous steel alloys. *J Mater Res* 19:3046–52.
87. Lu, Z. P., C. T. Liu, J. R. Thompson, and W. D. Porter. 2004. Structural amorphous steels. *Phys Rev Lett* 92:245503.
88. Zhang, Q. S., W. Zhang, and A. Inoue. 2006. New Cu-Zr-based bulk metallic glasses with large diameters of up to 1.5 cm. *Scripta Materialia* 55:711–13.
89. Xu, D. H., G. Duan, W. L. Johnson, and C. Garland. 2004. Formation and properties of new Ni-based amorphous alloys with critical casting thickness up to 5 mm. *Acta Materialia* 52:3493–97.
90. Lin, X. H., and W. L. Johnson. 1995. Formation of Ti-Zr-Cu-Ni bulk metallic glasses. *J Appl Phys* 78:6514–19.
91. Inoue, A., A. Kato, T. Zhang, S. G. Kim, and T. Masumoto. 1991. Mg-Cu-Y amorphous-alloys with high mechanical strengths produced by a metallic mold casting method. *Materials Transactions Jim* 32:609–16.
92. Nishiyama, N., and A. Inoue. 1999. Supercooling investigation and critical cooling rate for glass formation in P-Cu-Ni-P alloy. *Acta Materialia* 47:1487–95.
93. Schroers, J., and W. L. Johnson. 2004. Highly processable bulk metallic glass-forming alloys in the Pt-Co-Ni-Cu-P system. *Appl Phys Lett* 84:3666–68.
94. Schroers, J., B. Lohwongwatana, W. L. Johnson, and A. Peker. 2005. Gold based bulk metallic glass. *Appl Phys Lett* 87:61912.
95. Johnson, W. L. 1999. Bulk glass-forming metallic alloys: Science and technology. *MRS Bull* 24:42–56.
96. Schroers, J., and N. Paton. 2006. Amorphous metal alloys form like plastics. *Adv Materials Processes* 164:61–63.
97. Schroers, J. 2005. The superplastic forming of bulk metallic glasses. *JOM* 57:35–39.
98. Schroers, J., Q. Pham, and A. Desai. 2007. Thermoplastic forming of bulk metallic glass—A technology for MEMS and microstructure fabrication. *J MEMS* 16:240–47.
99. Saotome, Y., K. Imai, S. Shioda, S. Shimizu, T. Zhang, and A. Inoue. 2002. The micro-nanoformability of Pt-based metallic glass and the nanoforming of three-dimensional structures. *Intermetallics* 10:1241–47.
100. Jeong, H. W., S. Hata, and A. Shimokohbe. 2003. Microforming of three-dimensional microstructures from thin-film metallic glass. *J MEMS* 12:42–52.
101. Kumar, G., and J. Schroers. 2008. Write and erase mechanism on bulk metallic glass. *Appl Phys Lett* 92:031901.
102. Schroers, J., Q. Pham, and A. Desai. 2007. Thermoplastic forming of bulk metallic glass—A technology for MEMS and microstructure fabrication. *J MEMS* 16:240–47.
103. Madou, M., J. Zoval, G. Jia, H. Kido, J. Kim, and N. Kim. 2008. Lab on a CD. *Annu Rev Biomed Eng* 8:601–28.
104. Eklund, E. J., A. M. Shkel, S. Knappe, E. Donley, and J. Kitching. 2008. Glass-blown spherical microcells for chip-scale atomic devices. *Sensors Actuators A* 143:175–80.
105. Eklund, E. J., and A. M. Shkel. 2007. Glass blowing on a wafer level. *J MEMS* 16:232–39.
106. Von Recum, A., ed. 1998. *Handbook of biomaterials evaluation: Scientific, technical, and clinical testing of implant materials.* New York: Hemisphere.
107. Wisniewski, N., F. Moussy, and W. M. Reichert. 2000. Characterization of implantable biosensor membrane biofouling. *Fresenius J Anal Chem* 366:611–21.
108. Kaetsu, I. 1995. Biocompatible and biofunctional membranes by means of radiation techniques. *Nucl Instrum Methods Phys Res B* 105:294–301.
109. Schlosser, M., and M. Ziegler. 1997. Biocompatibility of active implantable devices. In *Biosensors in the body: Continuous in vivo monitoring,* ed. D. M. Frasier. New York: Wiley.
110. Reichert, W. M., and A. A. Sharkawy. 1999. Biosensors. In Handbook of biomaterials evaluation (ed. von Recum, A.) 439–60. Philadelphia: Taylor and Francis.
111. Park, J. B., and R. S. Lakes. 1993. *Biomaterials: An introduction.* New York: Springer.
112. Desai, T. A., D. Hansford, and M. Ferrari. 1999. Characterization of micromachined silicon membranes for immunoisolation and bioseparation applications. *J Membr Sci* 159:221–31.
113. Madou, M. J., and S. R. Morrison. 1989. *Chemical sensing with solid state devices.* New York: Academic Press.
114. Mrksich, M., and G. M. Whitesides. 1996. Using self-assembled monolayers to understand the interactions of man-made surfaces with proteins and cells. *Annu Rev Biophys Biomol Struct* 25:55–78.
115. Ratner, B. D., and A. S. Hoffman. 1996. Thin films, grafts, and coatings. In *Biomaterials science: An introduction to materials in medicine,* eds. B. D. Ratner, A. S. Hoffman, F. J. Schoen, and J. E. Lemons. San Diego, CA: Academic Press, p. 105.
116. Zhang, M., T. Desai, and M. Ferrari. 1998. Proteins and cells on PEG immobilized silicon surfaces. *Biomaterials* 19:953–60.
117. Brooks, K. A., J. R. Allen, P. W. Feldhoff, and L. G. Bachas. 1996. Effect of surface-attached heparin on the response of potassium-selective electrodes. *Anal Chem* 68:1439–43.
118. Holland, N. B., Y. Qiu, M. Ruegsegger, and R. E. Marchant. 1998. Biomimetic engineering of non-adhesive glycocalyx-like surfaces using oligosaccharide surfactant polymers. *Nature* 392:799–801.
119. Ishihara, K., N. Nakabayashi, M. Sakakida, N. Kenro, and M. Shichiri. 1998. In *American Chemical Society Annual Meeting.* Orlando, FL.
120. Frasier, D. M., ed. 1997. *Biosensors in the body: Continuous in vivo monitoring.* New York: Wiley.
121. Wisniewski, N., and W. Reichert. 2000. Methods for reducing biosensor membrane biofouling. *Colloids Surf B: Biointerfaces* 18:197–21.

6

Metrology and MEMS/NEMS Modeling

Outline

Introduction

Metrology

Tools to See and Size: Dimensional Metrology

Tools to Measure Materials Properties: Property Metrology

Compositional Metrology

Summary: Microscope Comparisons

MEMS Based Metrology: In Situ Stress-Measuring Techniques

Biaxial Measurements of Mechanical Properties of Thin Films: Suspended Membrane Methods

Other Surface Micromachined Structures to Gauge Intrinsic Stress

Stress Nonuniformity Measurement by Cantilever Beams and Cantilever Spirals

Modeling

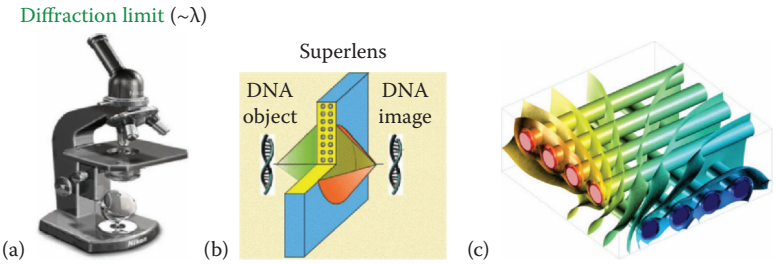
Finite Element Analysis

CAD for MEMS

Questions

Further Reading

References



From (a) microscopes to (b) superlenses. (c) Model of microelectromechanical system heat exchanger. (From http://www.comsol.com/showroom/gallery/252.php.) In a new field such as nanotechnology it is often the consultants and conference organizers making the only money. Another group often cashing in early are measuring instrument or metrology developers. Edwards compared this to the hardware stores and whorehouses making more money in the gold rush of 1849 than the actual gold miners. (Steven A. Edwards, in *The Nanotech Pioneers,* 2006.)

Steven A. Edwards

Introduction

This chapter is divided into two parts; one deals with metrology of microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS), the other with their modeling. Metrology is the science that deals with measurements. Tools for MEMS and NEMS metrology are either analytical techniques derived from analytical chemistry or size measuring techniques borrowed from the semiconductor industry. We start the metrology section by exploring techniques to visualize MEMS and NEMS devices and to measure sizes such as film thickness and critical dimensions (CDs), i.e., dimensional metrology. Dimensional metrology tools we cover are light microscopes (including dark-field, bright-field, phase contrast, and fluorescence), scanning confocal and two-photon microscopes, white light interferometers, ellipsometers, scanning electron

microscopes (SEMs), and a set of scanning proximal probe microscopes (SPMs) including scanning near-field optical microscopes (NSOMs or SNOMs), scanning tunneling microscopes (STMs), and atomic force microscopes (AFMs).

In the category of material property measuring equipment, we start with a review of specialized SPMs for material property determination. SPMs are used not only to visualize structures (see SNOM, STM, and AFM, above) but also to analyze a wide variety of material properties such as magnetic, thermal, and mechanical behavior. We then consider tools used to measure conductivity and carrier type of a material. MEMS devices often are mechanically dynamic by nature, hence the need to measure dynamic motion of MEMS devices. We study mechanical properties using interferometry with a stroboscopic light-emitting diode (LED) to freeze device motion and laser Doppler vibrometry to measure contactless velocity and displacement of vibrating structures.

Next are examples of metrology techniques to analyze structure of samples (e.g., crystallinity, defects, and surface roughness) and finally analytical techniques for chemical composition. In the structural metrology category, we review stylus profilometry, which can be used for surface roughness and transmission electron microscopy (TEM) with utility for crystallinity determination. As examples of elemental analysis of surfaces, we discuss energy dispersive analysis of x-rays (EDAX), a standard procedure for identifying and quantifying elemental composition of specimen surfaces to shallow depths, and scanning Auger microscopy (SAM), which can identify every atom from lithium on up.

We close off the metrology section by looking at MEMS approaches to metrology. This includes on-chip MEMS devices to measure physical properties such as stress, Young's modulus, Poisson ratio, natural frequency, etc.

Modeling is the use of mathematical equations to simulate and predict real events and processes. In the modeling section, we introduce finite element analysis and review software packages helpful in the design phase of miniaturized devices, that is, software for computer-aided design (CAD) in MEMS. The emphasis in this modeling section is on MEMS modeling. NEMS and BIONEMS require molecular and quantum effect modeling. These techniques are used to construct, display, manipulate, simulate, and analyze molecular structures and to calculate their properties. The topic is beyond the scope of the current work, and only some useful software packages will be referred to.

Metrology

Introduction

The history of the last two industrial revolutions—first the machine revolution (1800–1920) and second the semiconductor revolution (1950–2010)—reveals that a nanotechnology revolution might not happen until the considerable lag of new nanometrology technologies has been remedied. During the machine revolution, the milli-inch (1 mil = 25 μm) was the measuring stick, and it could be gauged by vernier calipers. The semiconductor industry today has as a dominant measure 50 nm, accessible through electron imaging and SPMs. The nanometer revolution is centered at the 1 nm length-scale, for which we do not yet have the appropriate high-throughput tools. Today, high-throughput imaging or measuring sub-30-nm features, using a conventional SEM or SPMs (say, an AFM or an STM), results in fuzzy, noisy images. Fast, sharp, low-noise images at the nanoscale will take Herculean efforts. This does not bode well for the swift arrival of so-called *nanomanufacturing*, where rapid, accurate, high-resolution measurements are essential for profitability. On the positive side, in research settings, measurements, images, and motion can be controlled to within 10 picometers (10^{-11} m).

MEMS and NEMS metrology includes CD measurements of metal lines on an accelerometer, establishing surface roughness of a pressure sensitive membrane, determining compositional profiles of a pn-junction, and measuring magnetic, mechanical, electrical, and optical properties (see Figure 6.1). Accordingly, metrology equipment is divided into dimensional metrology (I), material property studies (electrical, optical, magnetic, mechanical, etc.) (II), structure identification (surface roughness, defects, crystallinity) (III), and chemical composition

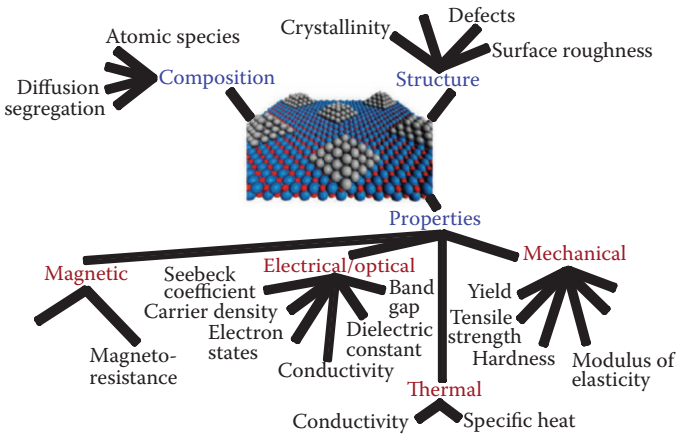


FIGURE 6.1 Characterizing integrated circuits, microelectromechanical systems, and nanoelectromechanical systems.

analysis (IV). In Table 6.1, we group metrology techniques in these four categories and list examples of typical tools.

Several metrology tools are used for different types of metrology. For example, scanning electron microscopy (SEM) can be used for dimensional metrology [e.g., to measure the CD of an integrated circuit (IC)] as well as in structural metrology (e.g., to measure defects in an IC). If the SEM comes equipped with EDAX, which it often does, chemical composition information can be collected as well. Similarly, SPMs may image MEMS, measure their size, or establish important material properties.

The measurement tools in Table 6.1, when used in IC, MEMS, or NEMS manufacturing, are further described as being stand-alone or integrated with the manufacturing line.[1] A stand-alone tool is a separate piece of equipment that is either off-line, at-line, or in-line. Off-line is an instrument that is available only outside the fab (usually involves destructive or contaminating testing). An at-line piece of equipment is one that is available in the fab itself, and in-line tools are used during production to measure, for example, processed Si wafers. An integrated tool is one that is combined with a manufacturing line and is either on-line or in situ. An on-line tool is available at the process workstation to measure patterned wafers but is not able to measure during wafer processing. In the sections below, we select some of these techniques for further scrutiny. In Table 6.2 additional characteristics of the most important metrology techniques from Table 6.1 are listed, and in Figure 6.2 we compare this most important set of analytical tools in terms of detection limit, sampling depth, and spot size.

Tools to See and Size: Dimensional Metrology

Introduction

An important category in metrology equipment comprises seeing and sizing tools, for example, to image an etched profile in a MEMS structure, to measure film thickness, or to determine a critical dimension (CD) on an IC. From Table 6.1, seeing tools range from optical microscopy (OM), to scanning electron microscopy (SEM) and scanning proximal probe microscopes (SPMs), including scanning tunneling microscopy (STM), atomic force microscopy (AFM), and scanning near-field optical microscopy (NSOM or SNOM). Film thickness is accessible, for example, through ellipsometry, photoacoustic measurements, x-ray fluorescence (XRF), reflection spectroscopy, and interferometry. CD in ICs and MEMS and NEMS are typically measured using a scanning electron microscope (CD-SEM) or an atomic force microscope (CD-AFM) with subnanometer resolution and accuracy. In the sub-100-nm IC generations, CDs are becoming harder and harder to control. With 65-nm features, for example, one typically has an error budget of only ±5 nm and current CD-SEM metrology techniques offer about ±2 nm precision. An AFM, in theory, should provide better resolution, but unfortunately AFM scanning is slow and suffers from probe tip wear. Another entry in CD measurements is the use of scatterometry that involves measurements on IC wafers by diffraction from grating test patterns. Scatterometry is a faster and more economic way to characterize periodic patterns on a wafer to relieve/bypass the heavy workload of CD-SEM and CD-AFM in a typical fab. Scatterometry involves flood exposure of a field, the measurement of changes in the polarization state of the reflected light, and calculation of IC-structure characteristics via real-time or library-based regression software algorithms. With visible light, scatterometry affords a resolution, R, of about 0.15 nm and can deliver multiple measurements, including

Table 6.1 Metrology Equipment for Integrated Circuits, Microelectromechanical Systems, and Nanoelectromechanical Systems

	Examples	Abbreviation
Dimensional metrology (e.g., film thickness, CD) (I)		
	High-resolution microscopy (R = diffraction limit, ~150 nm)	HM
	Phase contrast microscope	PCM
	Scanning tunneling microscope	STM
	Scanning electron microscope (lateral resolution, R = 35 Å)	SEM
	Ellipsometer for thickness of transparent films	
	Scanning proximal probe microscope	SPM
	Stylus profilometer for sample surface roughness	
	Reflection spectroscopy for thickness of opaque films	
	X-ray fluorescence for film thickness	XRF
	Photoacoustic technology for film thickness	
	Scatterometer for analysis of periodic structures. CD measurement	
	Vernier alignment scales (on-chip)	
Metrology for material properties (II)		
Mechanical properties		
	Micro/nanoindentation with scanning proximal probe	SPM
	Residual stress measurements with interferometry	
	Vibrational analysis with laser Doppler	
Electrical properties		
	Capacitance	
	Resistance with a four-point probe	
	Electron energy loss spectroscopy	EELS
	Carrier type with hot probe	
	Dopant concentration with thermal wave system	
Optical properties		
	Ellipsometer	
	Photoelectron spectroscopy	XPS, UPS
	Ultraviolet/visible spectroscopy	UV/VIS
	Scanning near-field optical microscopy	NSOM/SNOM
Magnetic properties		
	Secondary electron microscopy with polarization analysis	SEMPA
	Scanning magnetic force microscopes	SFM
Metrology for structural properties (III)		
	Scanning electron microscope, R = 35 Å (lateral)	SEM
	Transmission electron microscope, R = 2 Å	TEM
	Scanning probe microscopies	SPMs such as STM and AFM
	High-resolution microscopy (R = diffraction limit, ~150 nm)	HM
	Field ion microscope	FIM
	Phase contrast microscope	PCM
	X-ray microscope (Synchrotron radiation and x-ray optics; R = ~1 Å)	
	Low energy electron emission microscopy (low energy e-beam; R = 20 nm)	LEEM
	Focused ion beam	FIB
	X-ray diffraction spectroscopy	XRD
	Scanning x-ray photoelectron microscopy	SXPEM

Table 6.1 Metrology Equipment for Integrated Circuits, Microelectromechanical Systems, and Nanoelectromechanical Systems (*Continued*)

	Examples	Abbreviation
Metrology for chemical composition (IV)		
	Auger electron spectroscopy	AES
	Energy dispersive analysis of x-rays	EDAX or EDS
	X-ray photoelectron spectroscopy	XPS
	Photoemission electron microscopy: lateral R = ~100 nm	PEEM
	Secondary ion mass spectroscopy	SIMS
	Mass spectrometry	MS
	Rutherford back scattering	RBS
	X-ray fluorescence	XRF
	Nuclear magnetic resonance spectroscopy	NMR
	Photoluminescence spectroscopy	PL
	Cathodoluminescence spectroscopy	CL
	Raman spectroscopy	RAMAN
	Infrared spectroscopy	IR, FTIR
	Nuclear magnetic resonance spectroscopy	NMR
	Photoluminescence spectroscopy	PL
	Neutron activation analysis	NAA
	Ultraviolet/visible spectroscopy	UV/VIS

R, resolution.

CDs, feature shapes (i.e., CD profiles), overlay, and film thickness on patterned structures, including stacked films. Because scatterometry tools are diffraction based, the target must be a periodic structure that consists of a single feature repeated with a well-defined pitch. The technique cannot measure isolated single lines or more complex three-dimensional (3D) shapes.

There is a need not only for sizing metrology with faster throughput but also for more equipment that provides 3D information, not only in ICs but particularly for the newer areas of MEMS and NEMS, which add a more prominent *z*-dimension to parts. These three-dimensional parts require metrology tools that can measure in three dimensions. Most available tools are two-dimensional (2D) in nature, and measurements are made on the top or bottom of the microcomponent and use an algorithm to provide a pseudo-3D image (2.5D in reality), as often no sidewall information is available.

We cannot detail all tools that are available for dimensional metrology and limit ourselves to a review of light microscopes (including dark-field, light-field, phase contrast, and fluorescence), scanning confocal and two-photon microscopy, white light interferometry, scanning electron microscopy (SEM), NSOM, STMs, AFMs, and ellipsometry.

Light Microscopy

Since the resolution of the human eye is limited to about 100 micrometers, microscopes have always been the engines for new discoveries. Optical microscopy (OM) is usually what one thinks of when studying microscopy; this is also where we are starting our discussion. It was in 1590 that the Dutch spectacle-makers Hans Janssen and his son Zacharias Janssen invented a compound microscope with a magnification of ×3 to ×9. Robert Hooke, in 1665, published *Micrographia*, a collection of biological micrographs and he coined the word *cell* for the structures he observed in cork bark. In 1674 Anton van Leeuwenhoek improved on a simple microscope for viewing biological specimens and discovered bacteria (from tooth scrapings), protozoans (from pond water—he called them *animalcules*), sperm (his own, we presume), and blood cells (magnification of ×270). In Table 6.3, a brief history of microscopy is provided. The table includes light microscopy as well as electron microscopy, proximal probe microscopy, and near field superlenses.

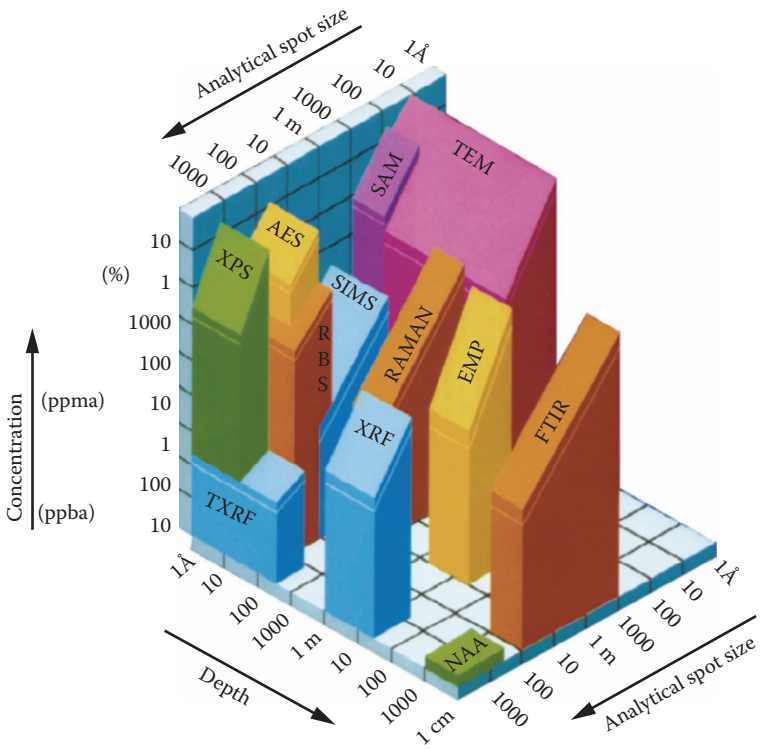
Table 6.2 Metrology for Integrated Circuits (ICs), Microelectromechanical Systems (MEMS), and Nanoelectromechanical Systems (NEMS)

	Spot Size or Probe Area	Resolution/ Magnification	Thickness Range	Applications	Remarks (n = Refractive Index)
Reflection spectroscopy (visible: 400–900 nm)	3.5–50 μm		40 Å to 5 μm	Nonabsorbing films.	Limit measures below 100 Å, n measures only above 700 Å. Fast.
Reflection spectroscopy (UV: 200–400 nm)	10 μm		25 Å to 2 μm	Transparent very thin films on metals.	Limit measures below 50 Å, n measures only above 100 Å.
Ellipsometry (632.8 nm standard), 790, 830, 1,300, and 1,500 nm (optional)	12–100 μm	1–2 μm	10–1,000 Å	Nonabsorbing films.	Automated, highest accuracy. Thickness and n measurements.
Optical microscopes		Magnification 10–1,000	0.3–0 5 μm	Universal tool for all dimensions > 0.29 μm.	Poor depth of field.
Scanning electron microscope (SEM)	Spots down to 10–30 Å	Lateral: –35 Å (standard SEM); –9 Å (field-emission SEM); –150 Å (low-voltage SEM); magnification is 20–150,000		SEM pictures, CD measures below 0.8 μm.	Good for surface topography. Large depth of focus. Needs vacuum. Can be destructive.
Transmission electron microscope (TEM)	10 Å	1–2 Å (resolving power 10^4 times better than optical microscope). Magnification is 500–500,000	10 Å to 1 μm	Atomic structure defect analysis nanoprobe.	Requires thin sample preparation.
Scanning acoustic microscopy	1.3 mm (50 MH_z)	Lateral: 10 nm to 2 mm Depth: 10 nm to 5 mm	Signal pene trates 1λ	Inspection inside a ceramic packaged IC. Interface inspection, wide range of samples.	Nondestructive. Subsurface information. Needs liquid medium.
Stylus profilometer	Tip size limits bandwidth	Resolution of better than 1 Å and lateral resolution of ~0.03 mm		Measures sharp steps from a few angstroms to several micrometers. Materials and process R&D, low volume QC.	Measures surface profile using a contact stylus. May damage soft samples.
Nomarski microscope (differential interference contrast)	0.3 μm			Separation of images of different layers. Sample roughness inspection.	Pseudo-three-dimensional image.
Confocal laser scanning microscope (CLSM)	0.3 μm	0.3 μm	<1 μm	Cross-sectioning. Independent layer imaging.	Nondestructive. No vacuum. Diffraction limit resolution.
Atomic force microscope (AFM)	Radius of curvature of probe tip: 50–100 Å	Atomic imaging (one atom)	<1 Å	Surface roughness. Tribology.	Slow scan. Probe tip wear. Non-destructive.

TABLE 6.2 Metrology for Integrated Circuits (ICs), Microelectromechanical Systems (MEMS), and Nanoelectromechanical Systems (NEMS) (*Continued*)

	Spot Size or Probe Area	Resolution/ Magnification	Thickness Range	Applications	Remarks (n = Refractive Index)
Total integrated scattering		With visible light 0.15 μm is possible		Lithography control, etch structure monitoring and metal grain size determination.	Nondestructive, rapid quantitative process monitor.
Scanning tunneling microscope	50–100 Å	Lateral: 0.1–1 nm Depth: 0.1 nm		Atomic resolution. measurement of chemical properties.	Sample must be conduct. Slow scan. Probe wear. Cannot measure surfaces larger than a few μm^2.
Optical interference microscope		Lateral: 0.2–2 μm Depth: 400 μm to 2 cm		Cross-sectioning.	Lateral resolution limited by diffraction.
Near-field scanning optical microscope	0.02–4 μm Depth: 10 nm	Lateral: 1–200 nm		General imaging. Lithography.	No vacuum.

See also http://www.mwrn.com/product/product.htmt



AES	Auger electron spectroscopy	SIMS	Secondary ion mass spectrometry
EMP	Electron microprobe	SAM	Scanning auger microscopy
FTIR	Fourier transform infrared	TEM	Transmission electron microscopy
NAA	Neutron activation analysis	TXRF	Total reflection x-ray fluorescence
RAMAN	Raman spectroscopy	XPS	X-ray photoelectron spectroscopy
RBS	Rutherford backscattering spectrometry	XRF	X-ray fluorescence

FIGURE 6.2 Comparison of analytical techniques: detection limits, sampling depth, and spot size. (Courtesy of Dr. Shaffner, National Institute of Standards and Technology.)

Table 6.3 Seeing Better and Better

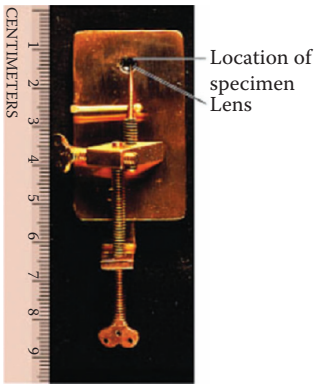
Year	Researcher	Discovery/Instrument
1590	Hans and Zacharias Janssen	First compound microscope, magnification × 3 and × 9
1590–1609	Galileo	One of the earliest microscopists; came up with the term *microscope*
1660	Marcello Malpighi	Father of embryology and early histology-observed blood capillaries
1665	Robert Hooke	Publishes the book *Micrographia* with the study of thin slices of cork
1674	Anton van Leeuwenhook	Microscope with ×275 magnification 
1827	Robert Brown	While examining pollen grains and the spores of mosses suspended in water under a microscope, Brown observed minute particles within vacuoles in the pollen grains executing a continuous jittery motion.
1860s	Ernst Abbe (1840-1905)	The Abbe sine condition: intrinsic diffraction limit
1926	Jean Baptiste Perrin	After Albert Einstein published his theoretical explanation of Brownian motion in terms of atoms (1905), Perrin did the experimental work to test Einstein's predictions, thereby settling the century-long dispute about John Dalton's atomic theory.
1931	Ernst Ruska and Max Knoll	The first electron microscope [a transmission electron microscope (TEM)]
1935	Frits Zernike	The phase contrast microscope
1936	Erwin Müller	The field emission microscope (FEM)
1951	Erwin Müller	The field ion microscope (FIM). Müller published his historic first FIM paper describing the significant improvement in contrast and resolution brought about by imaging with positive (hydrogen) ions compared with imaging by FEM and presenting the first evidence that atomic resolution was achieved.

Table 6.3 Seeing Better and Better (*Continued*)

Year	Researcher	Discovery/Instrument
1957	Marvin Minsky	Confocal microscope
1967	Erwin Müller (the same Müller)	The first atom probe, also known as the *atom probe field ion microscope,* incorporates FIM capability to give an atomic map of the specimen surface, by means of which the user selects atoms for chemical identification by time-of-flight mass spectrometry.
1981	G. Binnig, H. Rohrer	The scanning tunneling microscope (STM)
1984	Pohl	Near-field optical microscope
1986	Binnig, Quate, Gerber	The atomic force microscope (AFM)
1988	A. Cerezo, T. Godfrey, G. Smith	Added a position-sensitive detector to the atom probe → 3D
1988	Kingo Itaya	The electrochemical scanning tunneling microscope
1991	M. Nonnenmacher, M. P. O'Boyle, and H. K. Wickramasinghe	Kelvin probe force microscope is the Kelvin mode of scanning probe microscopy; it measures contact potential difference between the probe and the sample.
2003	Zhang	Silver near-field superlens

The first lenses were made in about 1000 AD, and spectacles came about around 1200 AD. From Table 6.3, it is obvious that optical microscopes have come a long way since van Leeuwenhoek's animalcules (1674), Brown's observation of moving pollen grains (Brownian motion) (1827), and Jean Perrin's demonstration of Einstein's proof that atoms really do exist (1926 Nobel Prize for Physics) by observing the distribution of particles undergoing Brownian motion. Ernst Ruska and Max Knoll invented the electron microscope in 1931, interestingly; at first they were working with the electron as a particle concept. But as de Broglie showed, the wavelength of an electron is 100,000 times smaller than the smallest wavelength of light, small enough to allow for 2.2-Å resolution. That type of resolution was only achieved forty years after Ruska and Knoll's breakthrough, and by the 1970s, researchers were able to visualize individual proteins and DNA molecules.

Microscopy helps to describe the size and structure of samples. This information is often critical for understanding the mechanisms of forming the sample or of the observed performance of the material. A very important question to ask regarding a microscope is about its ultimate resolution: what is the smallest feature size that it can resolve? Resolution is usually different when comparing lateral resolution (resolution in the plane of the sample) and vertical resolution. One also needs to know the dynamic range of the instrument. Techniques that have a high resolution often are not able to image large objects or feature sizes that vary over several orders of magnitude (they have a small field of view).

Roughly speaking, the lateral resolution of an optical microscope is limited to Abbe's limit, about 1/2 of the wavelength of the light being diffracted through the optical lensing elements of the microscope [i.e., Abbe's formula: $R = \lambda/2NA$, with $NA = n \sin \theta_{max}$ (see Volume I, Equations 5.6 and 5.3)]. In theory, the R using visible light (shortest wavelength is 400 nm) and the best numerical aperture (NA) (1.4) is about 200 nm, but in practice it is challenging to get good optical resolution down to the 1-μm range. Based on Abbe's formula, the lateral R of a microscope can be improved by working at shorter wavelengths, λ, by using a higher aperture lens (Figure 6.3) and

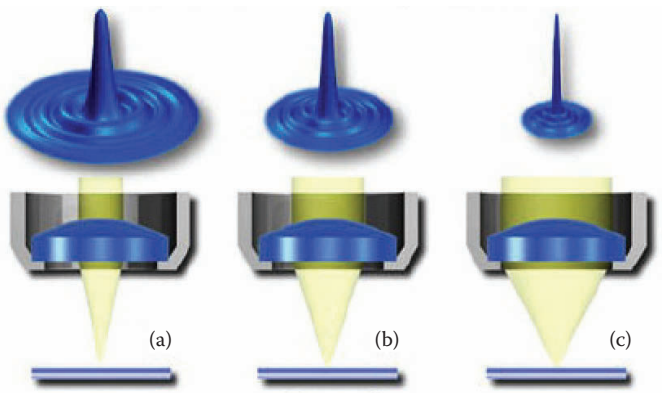


FIGURE 6.3 Increasing numerical aperture (NA) of the lens from (a) to (c). As NA increases, the acceptance angle increases and the Airy disc decreases.

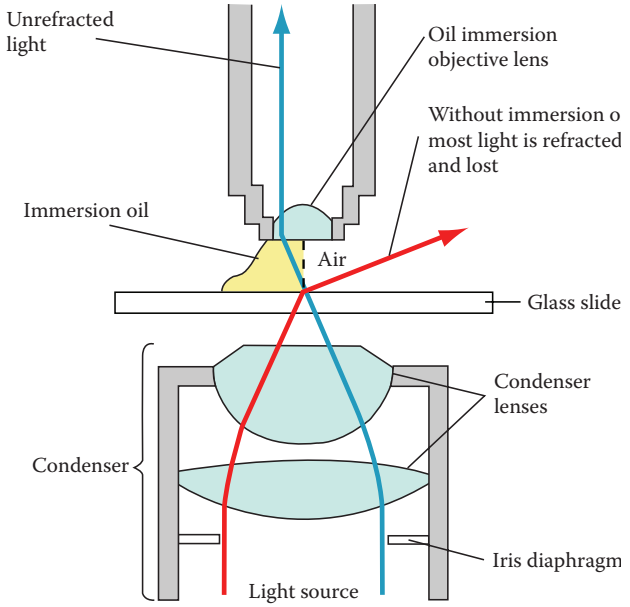


FIGURE 6.4 Oil immersion microscopy: increasing the refractive index of the medium (*n*).

in a medium with a higher index of refraction (*n*) (Figure 6.4). A larger NA means that the lens has a larger acceptance angle, so it can collect higher orders of diffracted light enhancing the resolution of the image. Working with oil between sample and objective lens, in so-called *oil immersion microscopy*, ensures a better capture of all diffracted wavelengths. The refractive index n (1.0 in air, 1.33 in water, and 1.55 in immersion oil) is the light-bending ability of a medium. The light coming from a sample under the microscope may bend in air so much that it misses the small, high-magnification lens. Immersion oil is used to keep light from bending too far.

The Abbe imaging resolution is a theoretical limit. Light passing through a circular aperture interferes with itself, creating ring-shaped diffraction patterns, known as *Airy patterns*, that blur the image (Volume I, Figures 5.13 and 5.14), and the Rayleigh criterion, which takes these Airy patterns into account, yields a minimum spatial resolution as follows:

$$R = \frac{0.61\,\lambda}{NA} \qquad \text{(Volume I, Equation 5.13)}$$

Objective lenses in a microscope are designed to correct for aberrations: achromats are the lowest level of correction, and apochromats, are the highest and most sophisticated. Eyepieces can have a reticle (graticle) etched inside to impose a scale on an image for measurement purposes. Light microscopy often is divided up into bright-field, dark-field, phase contrast, and fluorescence microscopy. In Figure 6.5 we illustrate bright-field (Figure 6.5a), dark-field (Figure 6.5b), and phase-contrast microscopy (Figure 6.5c). In bright-field microscopy, dark objects are

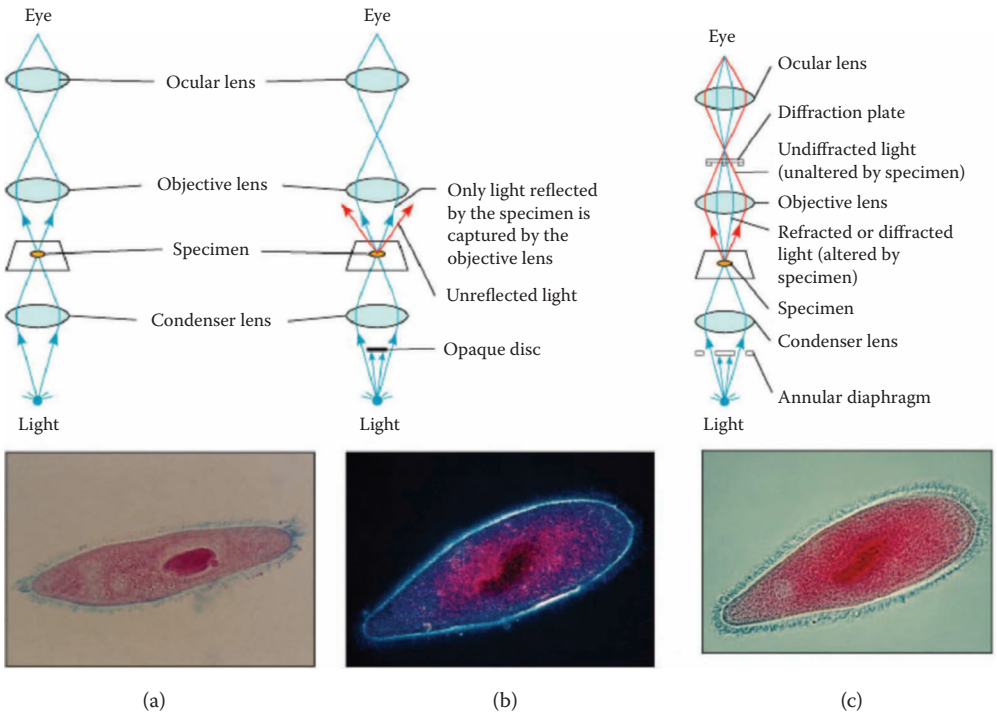


FIGURE 6.5 Comparison of (a) bright-field, (b) dark-field, and (c) phase-microscopy of a protozoa.

made visible against a bright background, and in dark-field microscopy, light objects are made visible against a dark background. In bright-field illumination, light reflected off the sample does not enter the objective lens, whereas in dark field it does.

The human eye does not see phase changes in a light beam directly. This is unfortunate, as phase changes carry a large amount of additional information about a sample. To make a phase change visible, one needs to combine the light passing through the sample with a reference so that the resulting interference reveals the phase information of the sample. In phase-contrast microscopy, one uses an annular transparent phase plate* so that the change in phase of light that passes through the sample is increased by shifting back ¼ wavelength relative to the background, while background light is shifted forward ¼ wavelength. When the waves recombine they produce a ½-wavelength shift, resulting in excellent contrast. The phase contrast microscope converts these phase changes to changes in brightness that the eye can see. A differential interference contrast (DIC) microscope, also known as a *Nomarski microscope*, is similar to a phase contrast microscope in that imaging again produces amplitude difference from phase differences in the specimen. The Nomarski microscope uses polarized light and emphasis differences in the refractive index of transparent samples. This gives relief to a sample corresponding to the variation of optical density. In Figure 6.6, we show a standard light microscope (Olympus BH2).

In Table 6.4, we compare the resolution of various types of optical microscopes. For a discussion of NSOM, see below.

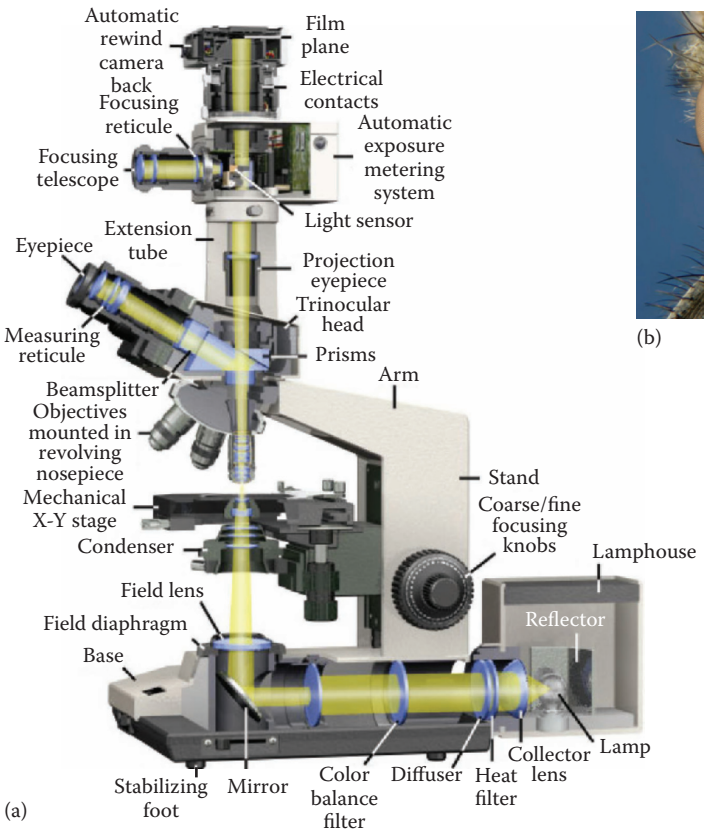
Vibration is a fundamental problem in all microscopy experiments. Buildings vibrate because of motors, air conditioners, passing vehicles, etc. Vibrations blur images since the excitation, detection, and sample regions of the microscopes may move with respect to each other during the course of the experiment. Most buildings have vibrational resonances in the 1–100 Hz range and feature amplitudes in the micrometer range. The effect of vibrations is mitigated by constructing the microscope on stages connected to each other either very rigidly or very loosely. The rigid connections permit the transmission of only high-frequency vibrations, while the loose connections permit only low-frequency vibrations. Compare this with the implementation of high-pass and low-pass filters in optics or electronics. Often, this filtering is achieved by building the microscope to be as small and rigid as possible and then mounting the entire system on pneumatically supported legs that couple only loosely to the building. Sometimes an additional stage using soft springs is included. While these procedures reduce mechanical vibration, acoustical vibrations may still couple to the microscope and circumvent these vibration isolation stages. Including a surrounding of acoustic-absorbing material is often necessary for the best microscopy work.

Light Microscopy Summary

Optical microscopes have the great advantages of being fast, in situ,† and nondestructive. Test parts usually do not have to be modified [e.g., coated with a conductive material, as in scanning electron microscopy (SEM)] from their original form/condition. Quantitation in light microscopy may be achieved with a reticle etched into one of the eyepieces. Images can be registered on film or, better, recorded electronically to a charge-coupled device (CCD) camera. Optical microscopes tend to be repeatable for features as small as 0.25 μm. Optical metrology hardware has a resolution that is diffraction limited, and it has a poor ability to produce images with clear intensity changes to accurately detect edges. Often locating the edge of a part is quite difficult, as location varies with lighting condition, noise, and assumptions made in the edge position algorithm.[2] Other significant errors of optical techniques typically stem from interference, resonance, shadowing, secondary reflections, and lens distortions. Perhaps the most important limitation of optical

* Rings in a phase plate can include attenuating layers (absorption but no phase shift), phase-shifting layers (phase shift but no absorption), or any combination of the two.

† In situ experiment: an experiment performed on a sample while it is still located in its native environment. With an in situ measurement, there is less risk of altering the sample's true properties. Ex situ experiment: an experiment performed on a sample after it has been removed from the location wherein it was formed. A wider range of experimental techniques are available.



(b)

FIGURE 6.6 (a) A standard light microscope (Olympus BH2 research microscope). (b) First place winner, Nikon's Small World 2005 competition: Charles B. Krebs, muscoid fly (house fly) (× 6.25), reflected light.

microscopes for MEMS inspection is the inability to acquire true three-dimensional data. Some optical microscopes are integrated with software that uses image-processing techniques to determine the *z*-height at which the image is taking place. The current state-of-the-art software uses a projected grid (a Ronchi grid) to determine the height at which the microscope is focused in one region of the image. If the region selected has multiple focus points (i.e., the region selected is not all on one plane), the algorithm assigns the average value for the *z*-height. Further edge detection algorithms are run to extract *x* and *y* data from the microscope image. This technique, in theory, produces three-dimensional data from an image; however, the algorithms used after finding the *z*-height in one location of the image assume that all of the data are on the same plane. Thus, the data acquired from vision systems such as these can be characterized as 2.5D data sets.[2]

TABLE 6.4 Resolution of Various Types of Optical Microscopes

Objective (Magnification/ Numerical Aperture)	Illumination	Resolution (μm)
×10/0.25	White light	1.34
×20/0.40	White light	0.84
×100/0.95	White light	0.35
×100/0.95	488 nm, confocal	0.22
×100/1.35 (with glycerin immersion)	248 nm (UV)	<0.16
Optical fiber (scanning near-field optical microscopy; values down to 10 nm have been reported)	488 nm	0.05

A good Web site for an introduction to optical microscopes can be found at http://www.olympusmicro.com/primer/opticalmicroscopy.html. Also visit http://micro.magnet.fsu.edu/index.html and go through the tutorials—they are excellent!

Fluorescence Microscopy

In BIOMEMS, fluorescence microscopy is a most important tool by which to study cells, DNA arrays, tissue, etc. It is used to study properties of samples

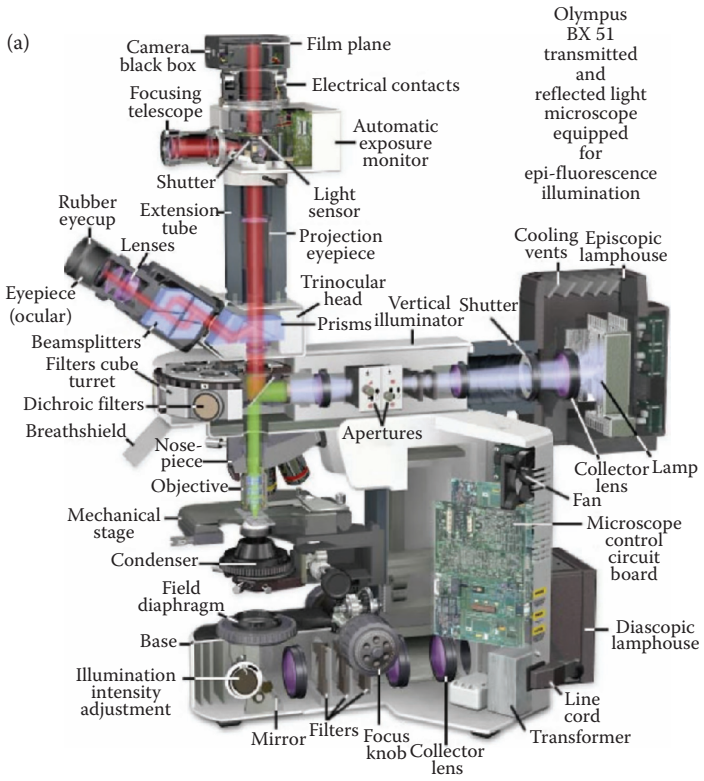


FIGURE 6.7 (a) Typical epifluorescence microscope. (b) Olympus Fluoview 300 LSM.

using the phenomena of fluorescence and phosphorescence (see Volume I, Chapter 7). Samples are labeled with dyes that emit visible light when bombarded with shorter ultraviolet (UV) rays (see Volume I, Figure 7.93). A fluorescence microscope is an optical microscope with an ultraviolet radiation source and a filter that protects the viewer's eye. The filter that the microscope uses is a special dichroic mirror (or more precisely, a dichromatic mirror). This mirror reflects light shorter than a certain wavelength and passes light longer than that wavelength. Thus the observer sees only the emitted red light from the fluorescent dye, rather than seeing scattered purple light. Most fluorescence microscopes are epifluorescence, i.e., excitation and observation of the fluorescence is from above (epi). In reflected light fluorescence, the intensity of the image, I, varies directly as the fourth power of the numerical aperture (NA) of the objective in use, as well as inversely as the square of the magnification, M:

$$I \propto \frac{(NA)^4}{(M)^2} \tag{6.1}$$

e.g., a ×40 objective with an NA of 1 will yield images more than a five times brighter than a ×40 objective with a NA of 0.65. Electronic sensors (e.g., a CCD) are typically used to measure the intensity, as they give a much higher sensitivity than film. To avoid fading of the fluorophore labels, a phenomenon known as *photobleaching*, the UV light intensity and duration need to be carefully chosen. In Figure 6.7 we show a typical fluorescence microscope.

Confocal Microscopy

Normally, in fluorescence microscopy, the sample is completely illuminated by the excitation light, so that the whole sample is fluorescing at the same time. The highest intensity of the excitation light is at the focal point of the lens, but nonetheless, the other parts of the sample do get some of this light and they do fluoresce as well.* This contributes to a background haze in the resulting fluorescence image. To get rid of this haze and make for a sharper image, one resorts to confocal microscopy. In such a microscope, patented in 1957 by Marvin Minsky (see inset in Figure 6.9), one places a small pinhole in a conjugate optical plane in front of the detector to eliminate out-

* From introductory optics, we remember the formula $1/s + 1/s' = 1/f$ for locating the image formed by a lens: points do not need to be at the focal point of the lens in order for the lens to form an image.

of-focus light (Figure 6.9). Because the focal point of the objective lens of the microscope forms an image where the pinhole is, these two points are known as *conjugate points*. The pinhole is conjugate to the focal point of the lens; thus, it is a confocal pinhole (or alternatively, the sample plane and the pinhole/screen are conjugate planes). Only light from the plane of focus will pass through the pinhole, so that the image quality is much better than that of wide-field images. Since only one point is illuminated at a time, 2D or 3D imaging requires that a focused laser beam be scanned across the field of view and the image be reconstructed electronically. This type of microscopy is called *confocal laser scanning microscopy* (CLSM), as it combines a confocal microscope with a scanning system in order to image an entire specimen. Although scanning can be performed in several different ways, it is most often done by moving the beam, which alleviates focus problems caused by objective lens scanning and is faster than specimen scanning. With a scanning system added, the system has the ability to scan multiple times on different imaging planes, resulting in a three-dimensional data set.

In confocal microscopy, the background rejection is thus substantially improved and the method allows for optical sectioning; moreover, the lateral and axial extent of the point spread function (see Volume I, Figure 5.12) is reduced by about 30% compared with that in a wide-field microscope. As a consequence, focal plane lateral resolution is improved by a factor of 1.5 over ordinary microscopy for incoherent imaging (Rayleigh limit or $R = 0.61\ \lambda/NA$; Volume I, Equation 5.13), as follows:

$$R_{xy} = 0.4\frac{\lambda_{emission}}{NA} \tag{6.2}$$

The vertical depth resolution of a confocal microscope enables optical sectioning and is poorer than the lateral resolution:

$$R_z = 1.4\frac{\lambda_{emission}n}{NA^2} \tag{6.3}$$

For typical objectives ($NA \leq 1.4$) and visible light ($\lambda \geq 400$ nm), the best lateral resolution is $R_{xy} = 0.13\ \mu m$ and the vertical resolution is $R_z = 0.43\ \mu m$, so that the depth discrimination is not comparable to the focal plane resolution.

Work has been done to use CLSM to perform 3D analysis of microstructures such as micro end mills and hot embossing tools, both with overall dimensions on the order of 1 mm^3. One of the most important advantages of confocal microscopy found is the ability of this microscope to measure steep slopes, up to almost 90° on a part with minimal surface roughness. This measurement requires a high-resolution, high-NA objective, which has a limited lateral measuring field unsuitable for measuring the entire sample. Because of this, a stitching procedure is used to combine scans taken with several objectives.

Two-Photon Microscopy

Using two-photon confocal microscopy further enhances the resolution of fluorescence microscopy. Multiphoton fluorescence microscopy is a relatively novel imaging technique that relies on the quasisimultaneous absorption of two or more photons (of either the same or different energies) by a molecule (see also quantum lithography in Volume II, Chapter 2). Two-photon microscopy was pioneered by Winfried Denk in the laboratory of Watt Webb at Cornell University.[4] He combined the idea of two-photon absorption with the use of a laser scanner.

The use of two-photon absorption has several important advantages. First, the probability of two-photon absorption increases quadratically with illumination intensity, essentially limiting excitation to a tiny focal volume (see Figure 6.8). This limits automatically out-of-focus background excitation and provides innate confocality, since only fluorophores in the plane of the focused beam are excited. One obvious benefit of this is to obviate the need for the confocal aperture in front of the detector used in one-photon confocal microscopy, maximizing the amount of light detected. Second, the use of longer wavelength excitation light slows the rate of photobleaching of the fluorescent labels on the sample, allowing the use of higher magnification objectives. Third, longer wavelength light is scattered less as it passes through a sample such as biological tissue, resulting in a greater penetration depth into the sample. These factors increase both the resolution and duration of measurement available with this technique. Fourth, the fact that two-photon excitation microscopy does not require a pinhole to obtain

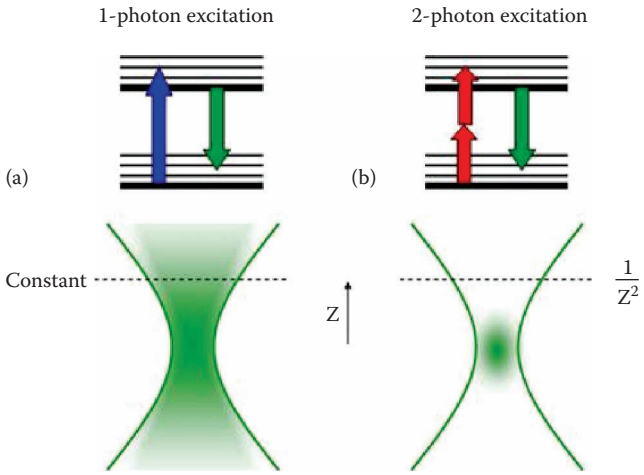


FIGURE 6.8 One-photon (a) and two-photon (b) excitation.

three-dimensional resolution allows more flexibility in detection geometry. Two-photon microscopy is thus a novel form of laser-based scanning microscopy that allows three-dimensional imaging and avoids many of the problems inherent in confocal microscopy.

In Figure 6.9 we compare confocal microscopy with laser scanning two-photon microscopy. Resolution limits of quantum lithography were analyzed in Volume II, Chapter 2. The results derived there directly apply to two-photon microscopy discussed here.

Interferometry

Another powerful method for MEMS part inspection is scanning white light interferometry (SWLI), also called *noncontact profilometry*. Interferometry, a method to determine a point in space on the basis of phase differences between a reflected probe beam and an isolated reference beam, has long been used by various industries as one of the most accurate ways to profile a surface. In Figure 6.10 and the example below, we illustrate how a typical interferometer can be used to measure distances in a number of wavelengths or to measure wavelengths. A Michelson interferometer, the particular two-path interferometer shown here, is constructed using a beam-splitter, i.e., a partially silvered mirror that transmits and reflects half of the incident intensity. This is useful because the reflected and transmitted beams are coherent, i.e., they have a definite common phase, even when the incident light has a time-varying random phase. This interferometer allows a

Inset

Prof. Marvin Minsky:

This is what I remember about inventing the confocal scanning microscope in 1955. When I read the following account, the plot seems more coherent now than it ever did in those times of the past. Perhaps, though, those activities which seemed to me the most spontaneous were actually those which unconsciously were managed the most methodically. The story actually begins in childhood ...

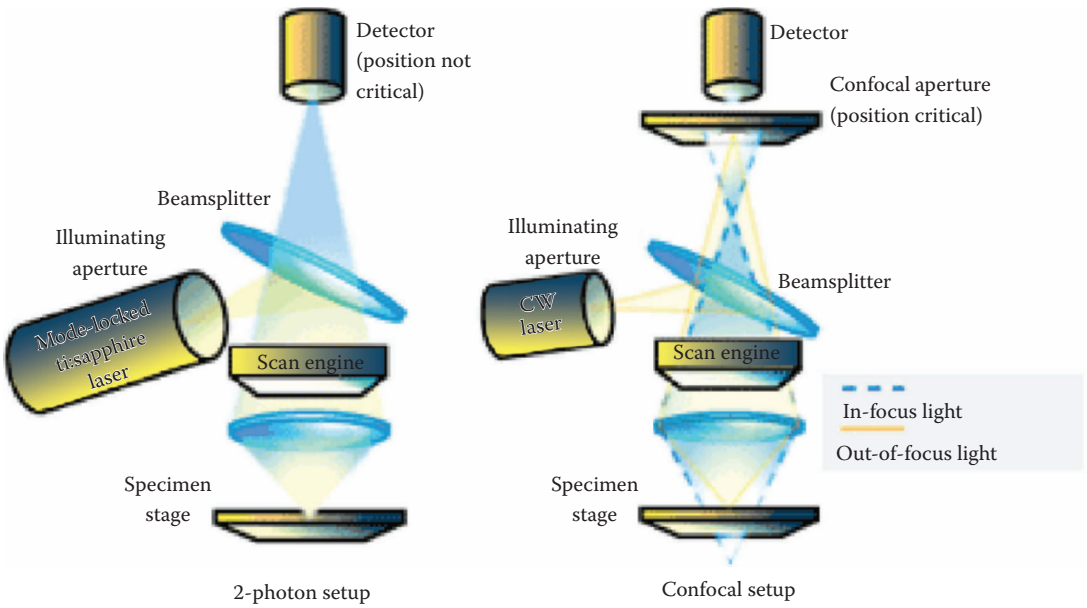


FIGURE 6.9 Comparison of a confocal microscope with a two-photon microscope. Inset, Marvin Minsky remembering the confocal microscope.

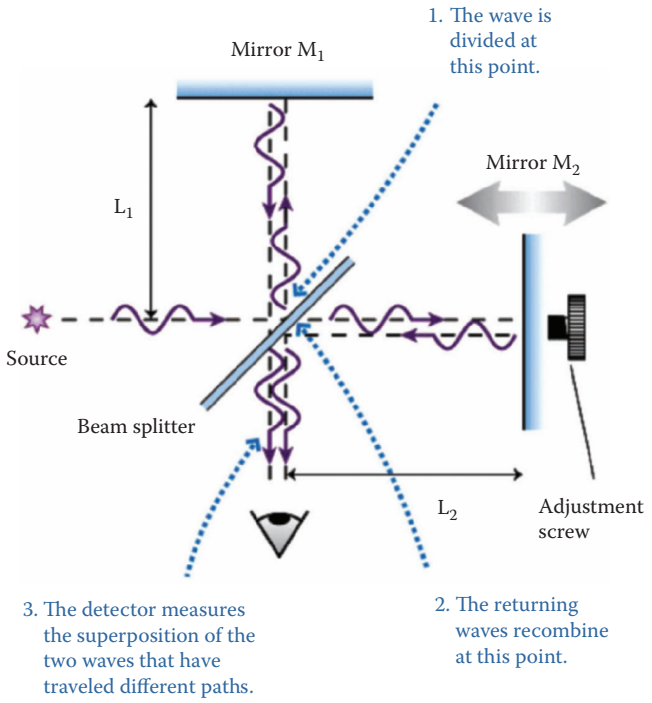


FIGURE 6.10 A Michelson interferometer. John G. Cramer, Professor of Physics, University of Washington.

beam of light to be sent along two paths, reflected, and recombined so that it can interfere. Reflection from a surface produces a "phase-flip" of 180° in the reflected light. However, when light is reflected at an angle other than 180°, the phase change in the reflected light is the same as the reflection angle. Thus, light reflected at 90° has a 90° change in phase. If the interferometer is set for an interference maximum, a new maximum will occur each time one of the arms is moved by λ/2.

Example 6.1 (Dr. J. G. Cramer): Measuring Wavelengths

Question: An experimenter uses the Michelson interferometer shown in Figure 6.10 to measure one of the wavelengths of light emitted by electrically excited neon atoms. She slowly moves the mirror M2 and uses a photo-detector and a computer to determine that 10,000 new bright central spots have appeared. She then determines with a micrometer that the mirror has moved 3.164 mm. What is the wavelength of the light?

Answer: $$\lambda = \frac{2\Delta L_2}{\Delta m} = \frac{2(0.003164\ \text{m})}{(10{,}000)} = 632.8\ \text{nm}$$

If the wavelength is known, then of course a distance can be determined.

In SWLI, recombined light is directed to image plane of a CCD camera. Points on a surface that are separated from the lens by an integer number of wavelengths are bright and those a half-integer apart are dark. Interference is strong only when the reflected light is in focus i.e., the sample-lens distance is at the focal position. The sample-lens distance is scanned around the focal length and ach pixel of the CCD will show a strong intensity change when the lens reaches the focal position corresponding to each point on the surface. High-resolution position information comes from a linear variable differential transformer (LVDT) connected to the lens scanning drive.

SWLI has the ability to quickly measure step heights changes and deflections (see also mechanical property measurements, below): it is possible to resolve step height changes greater than λ/4. Using white light increases the dynamic range of the measurement and allows for higher resolution by comparing data from multiple wavelengths. In addition, when integrated with an image processing system, SWLI can provide lateral dimensions. By using image processing, line widths, diameters, and feature positions can be measured as well. However, the lateral resolution of commercially available systems is poor, except when equipped with high-power objectives that severely limit the field of view. Moreover, these tools are limited in their ability to measure sloped surfaces. Indeed, interference can occur only if light is reflected back into the objective lens. If the surface angle is inclined beyond the acceptance angle of the lens, no interference is observed. The largest slope that can be identified is typically around 30° with a × 100 objective. As the objective power decreases, the identifiable slopes also decrease.

In practice, white light interferometers have subnanometer resolution in the scanning direction, at best submicrometer resolution in the lateral directions, and can be used on a multitude of parts with different surface finishes. A typical instrument, from Wyko, is shown in Figure 6.11. The equipment shown measures surface features with 0.45–11.8-μm lateral resolution and an approximately 0.10-nm vertical resolution. Although initially developed mostly for surface characterization, such as surface roughness, SWLI is currently used to make a wide variety of other measurements on MEMS parts.

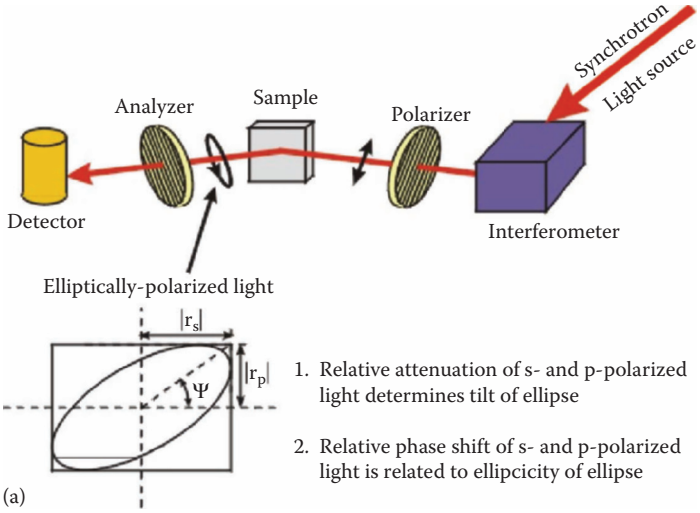
FIGURE 6.11 Optical Wyko white-light profilometers with dynamic microelectromechanical system capability. This capability enables both in-plane and out-of-plane resonant frequency measurements.

Ellipsometry

Introduction

Fresnel derived the equations that determine the reflection/transmission coefficients of light at interfaces in the early nineteenth century, and Paul Drude invented ellipsometry, based on these equations, in 1887. Ellipsometry is a noncontact optical technique used for analysis and metrology and became very important in the 1960s, as it was demonstrated to have the sensitivity required to measure nanometer-scale thin films used in microelectronics.

Ellipsometry allows one to establish very accurately and with high reproducibility the complex dielectric function of a given material. In Volume I, Chapter 3, we saw how the dielectric function is dependent on both frequency (ω) and wave-vector $\mathbf{k}$: $\varepsilon(\omega, \mathbf{k})$. In ellipsometry, one determines the change in polarization state of light reflected from a sample. This polarization change is represented as an amplitude ratio, Ψ, and a phase difference, Δ. Since ellipsometry is measuring the ratio (or difference) of two values (rather than the absolute value of either), it is very robust, accurate, and reproducible. For instance, it is relatively insensitive to scatter and fluctuations and requires no standard sample or reference beam. The measured response depends on optical properties and thickness of individual materials. Transparent films from subnanometers up to several micrometers can be measured, and the surface upon which the film is measured can be a semiconductor, dielectric or metal. The film can be transparent or absorbing on a transparent or opaque substrate as long as there is enough contrast between substrate and film. It is a model-based technique that, besides being used to determine film thickness and optical constants, can be employed to characterize composition, crystallinity, roughness, and doping level—in short, all material properties that can change the optical characteristics. In Figure 6.12 we illustrate a typical ellipsometer setup.



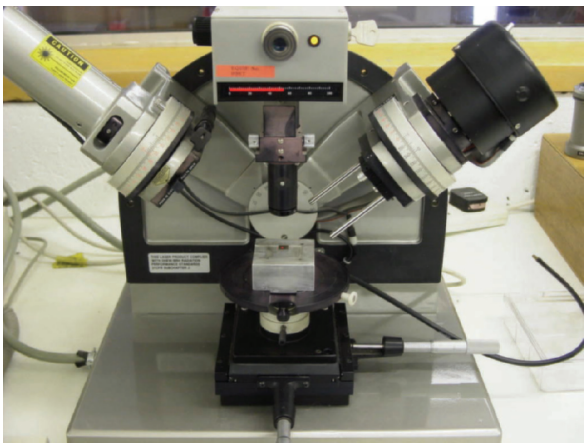


FIGURE 6.12 Schematic (a) and photograph (b) of an ellipsometer. The ellipsometer measures polarization changes in terms of two parameters: Δ (sometimes cosΔ) and Ψ.

From the schematic in Figure 6.12a, we see that a first polarizer linearly polarizes incident light with finite field components in the directions parallel (*p*-component) and perpendicular (*s*-component) to the plane of incidence of the light. The plane of incidence of light is the plane containing the surface that is normal to the sample and the incident direction of light. The wavelength of the measuring polarized light can range from the ultraviolet to the far-infrared. (In Figure 6.12a, a synchrotron light source is used in order to obtain a high-enough intensity of far-infrared light.) Upon reflection, the *s*- and *p*-components of the light undergo different attenuations and phase shifts, resulting in elliptically polarized reflected light, giving the technique its name. The ellipse of polarization of the reflected light is measured with a second polarizer (called the *analyzer*), as illustrated in Figure 6.12a.

Theory

When light interacts with a material, the Maxwell equations remain satisfied by imposing the proper boundary conditions at the interface (see Volume I, Table 5.2). At an interface a portion of light reflects, and the remainder transmits at the refracted angle (see Volume I, Figure 5.36). It is relatively easy to measure reflectance, R, of a material, and from this the more difficult-to-access complex refractive index, n_c, can be derived. The values for the refractive index, n, and extinction coefficient, κ, which make up the complex refractive index n_c, are not independent but are linked by the Kramers-Kronig relations. Either of these equations, plus a knowledge of $R(\omega)$ at all frequencies, permits one to disentangle the separate values of $n(\omega)$ and $\kappa(\omega)$.

With reference to Volume I, Figure 5.38, we obtained the four amplitude coefficients (the Fresnel equations) as follows:

$$r(\omega)_{||} = \left(\frac{E_m^r}{E_m^i}\right)_{\parallel} = \left[\frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}\right]_{\parallel}$$

(Volume I, Equation 5.140)

$$r(\omega)_{\perp} = \left(\frac{E_m^r}{E_m^i}\right)_{\perp} = \left[\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}\right]_{\perp}$$

(Volume I, Equation 5.141)

$$t(\omega)_{||} = \left(\frac{E_m^t}{E_m^i}\right)_{\parallel} = \left[\frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i + \theta_t)\cos(\theta_i - \theta_t)}\right]_{\parallel}$$

(Volume I, Equation 5.142)

$$t(\omega)_{\perp} = \left(\frac{E_m^t}{E_m^i}\right)_{\perp} = \left[\frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i + \theta_t)}\right]_{\perp}$$

(Volume I, Equation 5.143)

where $r(\omega)$ is the dimensionless reflectivity coefficient defined as the ratio of the reflected electric field to the incident electric field, and $t(\omega)$ is the dimensionless transmission coefficient.

Reflection and transmission, we saw, are related by the following equation.

$$t(\omega) = 1 + r(\omega)$$

(Volume I, Equation 5.144)

A complex $r(\omega)$ [or $t(\omega)$] means that a phase shift is introduced at the interface upon reflection (or transmission). The link between reflectivity at normal incidence and the complex refractive index was calculated in Volume I, Chapter 5 as follows:

$$r(\omega) = \frac{E_m^r}{E_m^i} = \left(\frac{n + i\kappa - 1}{n + i\kappa + 1}\right) = \frac{n_c - 1}{n_c + 1}$$

(Volume I, Equation 5.147)

The quantity that is actually measured in experiments is not reflectivity, r, but reflectance, R, defined as the ratio of the reflected intensity to the incident intensity, as follows:

$$R(\omega) = \frac{\left|E_m^r\right|^2}{\left|E_m^i\right|^2} = \left|\frac{n_c - 1}{n_c + 1}\right|^2 = \frac{(n-1)^2 + \kappa^2}{(n+1)^2 + \kappa^2}$$

(Volume I, Equation 5.148)

The phase, $\theta(\omega)$, of the reflected wave, it can be shown, may be calculated from the measured reflectance, $R(\omega)$, if known over all frequencies (Kramers-Kronig). With both $R(\omega)$ and $\theta(\omega)$ known, we can use Volume I, Equation 5.147 to calculate $n(\omega)$ and $\kappa(\omega)$; as we saw earlier, $n(\omega)$ and $\kappa(\omega)$ are related to the complex dielectric function, ε_r, as follows:

$$n_c(\omega) = n(\omega) + i\kappa(\omega) = \sqrt{\varepsilon_r'(\omega) + i\varepsilon_r''(\omega)}$$

(Volume I, Equation 5.115)

Plugging the values for $n(\omega)$ and $\kappa(\omega)$ into Equation 5.115, we thus obtain the complex dielectric function.

We now need to explain how the Fresnel equations are linked to the experimental parameters obtained from ellipsometry. The reflectivity coefficients $r(\omega)_{||}$ and $r(\omega)_{\perp}$ in what follows are symbolized more simply as r_p (p for parallel) and r_s (s for senkrecht: German for normal or perpendicular). Using Snell's law (Volume I, Equation 5.19), the expressions for r_s and r_p (Volume I, Equations 5.140 and 5.141) can be rewritten as follows:

$$r_s = \left(\frac{E_m^r}{E_m^i}\right)_{\perp} = \left[\frac{n_t\cos\theta_i - n_t\cos\theta_t}{n_i\cos\theta_t + n_t\cos\theta_i}\right]_{\perp} \tag{6.4}$$

$$r_p = \left(\frac{E_m^r}{E_m^i}\right)_{\parallel} = \left[\frac{n_t\cos\theta_i - n_i\cos\theta_t}{n_i\cos\theta_t + n_t\cos\theta_i}\right]_{\parallel} \tag{6.5}$$

In these expressions we now see the explicit appearance of the refractive index of the two media (n_i and n_t; see Volume I, Figure 5.38). Thin-film and multilayer structures involve multiple interfaces, with Fresnel reflection and transmission coefficients applicable at each. The superposition of multiple waves introduces interference that depends on the relative phase of each light wave. To track the relative phase of each light component so as to correctly determine the phase of the overall reflected or transmitted beam, a film phase thickness is introduced as follows (see Figure 6.13).

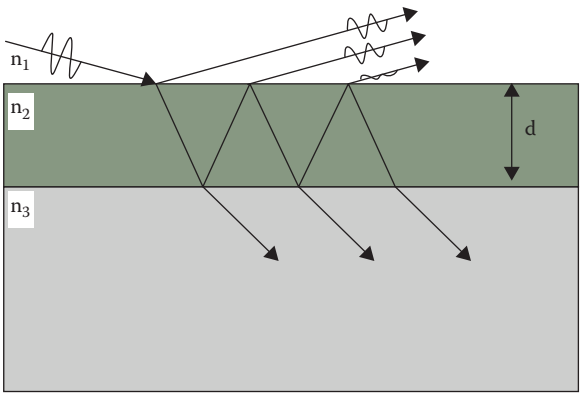


FIGURE 6.13 Light reflects and refracts at each interface, which leads to multiple beams in a thin film. Interference between beams depends on relative phase and amplitude of the electric fields. Fresnel: reflection and transmission can be used to calculate the response from each contributing beam.

$$\beta = 2\pi\left(\frac{d}{\lambda}\right)n_2\cos\theta_2 \tag{6.6}$$

The reflectivity coefficients r_p and r_s for this case are given as the following.

$$r_p = \frac{r_{p(1,2)} + r_{p(2,3)}\exp(-i2\beta)}{1 + r_{p(1,2)}r_{p(2,3)}\exp(-i2\beta)} \tag{6.7}$$

$$r_s = \frac{r_{s(1,2)} + r_{s(2,3)}\exp(-i2\beta)}{1 + r_{s(1,2)}r_{s(2,3)}\exp(-i2\beta)} \tag{6.8}$$

Ellipsometry determines how p- and s-components of the light change upon reflection or transmission in relation to each other. A known polarization is reflected or transmitted from the sample, and the output polarization is measured. The ellipsometer measures this polarization change of the light reflected from the sample in terms of two parameters: Δ (sometimes $\cos\Delta$) and Ψ (see Figure 6.12a). These values are related to the ratio of the Fresnel amplitude reflection coefficients for p- and s-polarized light as follows:

$$\frac{r(\omega)_{\parallel}}{r(\omega)_{\perp}} = \frac{r_p}{r_s} = \tan\Psi\exp(i\Delta) \tag{6.9}$$

This ratio is complex, with $\tan\Psi$ measuring the relative attenuation of s- and p-polarized light (which determines the tilt of the ellipse in Figure 6.12a) and Δ measuring the phase difference between p- and s-polarized reflected light (related to the ellipticity). In the case of an ideal bulk material, the result in Equation 6.9 can be directly converted to give the complex dielectric constant and complex refractive index as:

$$\varepsilon_r = n_c^2 = \sin^2(\phi)\left[1 + \tan^2(\phi)\left(\frac{1 - \frac{r_p}{r_s}}{1 + \frac{r_p}{r_s}}\right)\right] \tag{6.10}$$

This equation is an oversimplification implying that there are no surface layers of any type. A realistic bulk material comes with a surface oxide and at least some roughness, and Equation 6.10 would include these as part of the bulk optical constants. As a consequence, ellipsometry is used as an indirect method/model based technique, i.e., in general the measured Ψ and Δ cannot be converted directly into the optical constants of the sample. From Ψ, Δ, and an optical model, the desired information can be

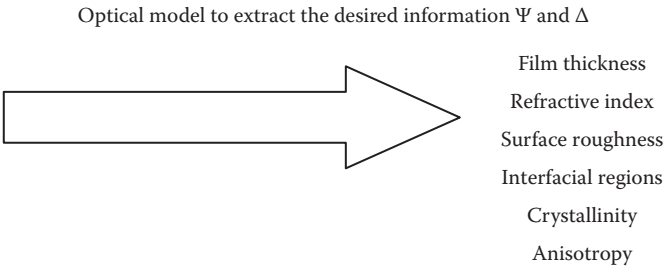


FIGURE 6.14 From Ψ, Δ, and an optical model the desired information can be extracted.

extracted, as illustrated in Figure 6.14. After a sample is measured, a model is constructed and is used to calculate the predicted response from the Fresnel's equations, which describe each material layer in the sample with its own thickness and optical constants. If these values are not all known, an estimate is made for a preliminary calculation. The calculated values are then compared with the experimental data, and any unknown material property can be varied to improve the match between experiment and calculation. Finding the best match between the model and the experiment is typically achieved through regression. An estimator, such as the mean squared error (MSE), is used to quantify the difference between curves. The unknown parameters are allowed to vary until the minimum MSE is reached.

Ellipsometers may be single-wavelength or spectroscopic and may be have single or multiple angles of incidence. They may also operate with a rotating element or as nulling ellipsometer, and they may be single-point or imaging. Spectroscopic ellipsometer allows the accurate characterization of a range of properties, including the layer thickness, optical constants, composition, crystallinity, anisotropy, and uniformity. Thickness determinations ranging from a few angstroms to tens of micrometers are possible for single layers and complex multilayer stacks.

Ellipsometry measures the ratio of two values, so it is highly accurate and reproducible, it does not need a reference sample, and it is not very susceptible to light source fluctuation. Since it measures phase, it is highly sensitive to the presence of thin films (down to submonolayer coverage). The sampling region ranges from 30 μm to 10 mm, and the depth of sampling, which depends on the absorption of the light in the sample, is typically 100 Å or more. For further reading we refer to the *Handbook of Ellipsometry*.[5]

Scanning Electron Microscopy

In ICs, MEMS, and NEMS, scanning electron microscopes (SEMs) like optical microscopes, are fabrication line workhorses. Scanning electron microscopy is one of the primary tools used for analysis of MEMS and NEMS devices, one that was used long before the development of scanning probe techniques. Ernst Ruska, the inventor of the electron microscope (1931),* received his Nobel in 1986, at age of 80, and shared the prize with Gerd Binnig and Heinrich Rohrer for the scanning tunneling microscope (STM) (see below), invented in 1981.

In an SEM, the surface or microvolume to be examined is irradiated with a finely focused electron beam; the beam may be swept in a raster† across the surface of the specimen, which is contained in a vacuum, to form images or can be static to obtain an analysis at just one position. The electrons are produced in an electron gun that accelerates them at the operating voltage. The operating voltage ranges from 0 to between 30 and 60 kV, depending upon the type of SEM.

Scanning electron microscopy requires a vacuum system so that the electrons can make the trip from the source in the electron gun, e.g., a thermal field emitter to the sample and to the detector (an electron multiplier). Insulating samples have traditionally required pretreatment in the form of coating with a conductive film, typically gold or carbon, before SEM examination. Nonconductive samples are subject to a buildup of electrons on their surface. This buildup of electrons, or charging, eventually causes scattering of the incoming electron beam, which interferes with imaging and analysis. Furthermore, samples that contain substantial water or other materials that volatilize in high vacuum also present challenges for SEM examination. These samples require controlled drying to allow the SEM chamber to reach high vacuum and to prevent deformation of the sample at the SEM vacuum. Coating with gold or

* There are two major types of electron microscopes: SEMs and TEMs. In the former, electrons scattered from a thick analyzed sample; in the latter, electrons transmitted through a very thin specimen are utilized. The first electron microscopes that were developed were actually TEMs (see "Transmission Electron Microscopy").

† The beam is scanned over the specimen in a series of lines and frames called a *raster,* just like the (much weaker) electron beam in an ordinary television.

carbon and drying are not needed if an environmental scanning electron microscope (ESEM) is used (see below). The focusing of electrons in an SEM is achieved magnetically instead of with a polished and shaped lens. Electron beams (probes) of sizes down to ~6 nm are attainable with conventional thermionic emission sources, although smaller probes, ~2 nm, can be achieved using field emission sources (for a listing of different electron sources, see Volume II, Figure 2.17). The SEM also does not have lenses to magnify the image as in the optical microscope; instead, magnification results from the ratio of the area scanned on the specimen to the area of the television screen [cathode ray tube (CRT)] (× 25 to × 250,000 or more). Increasing the magnification in an SEM is therefore achieved quite simply by scanning the electron beam over a smaller area of the specimen. The raster movement is accomplished by means of small coils of wire carrying the controlling current (the scan coils), and magnification is therefore controlled by the current supplied to the x, y scanning coils and not by objective lens power. A schematic drawing of an electron microscope is shown, together with a photograph of a commercial machine in Figure 6.15.

We remember that the resolution, R, of a light microscope can be calculated as $0.61\lambda/\mathrm{NA}$ (Raleigh criterion; Volume I, Equation 5.13). In this expression, $\mathrm{NA} = n\sin\theta_{max} = \frac{D}{2f} = \frac{1}{2F\#}$, with $F\#$ being the effective F-number, which equals f/D, or focal length over lens diameter. The assumption here is a large $F\#$, which implies $\theta \ll 1$ and $\mathrm{NA} = \sin\theta \sim \tan\theta \sim \theta$; this is called the *paraxial approximation* (see Volume I, Chapter 5). A good light microscope objective has an f/D (effective F-number) of about 2, so with λ = ~500 nm, a typical optical microscope R is about 1 μm. This resolution can be much improved by using a scanning electron microscope. The matter wave of an electron (accelerated at 1,000,000 V) is 0.001 nm (see de Broglie wavelength; Volume I, Equation 3.96) versus 450 nm for a light microscope; in other words, electron waves are 100,000 times shorter than the waves of visible light. So electrons have tremendous power to resolve minute structures because the resolving power is a function of wavelength. Moreover this wavelength depends on the applied voltage and increasing the voltage leads to faster electrons with yet shorter de Broglie

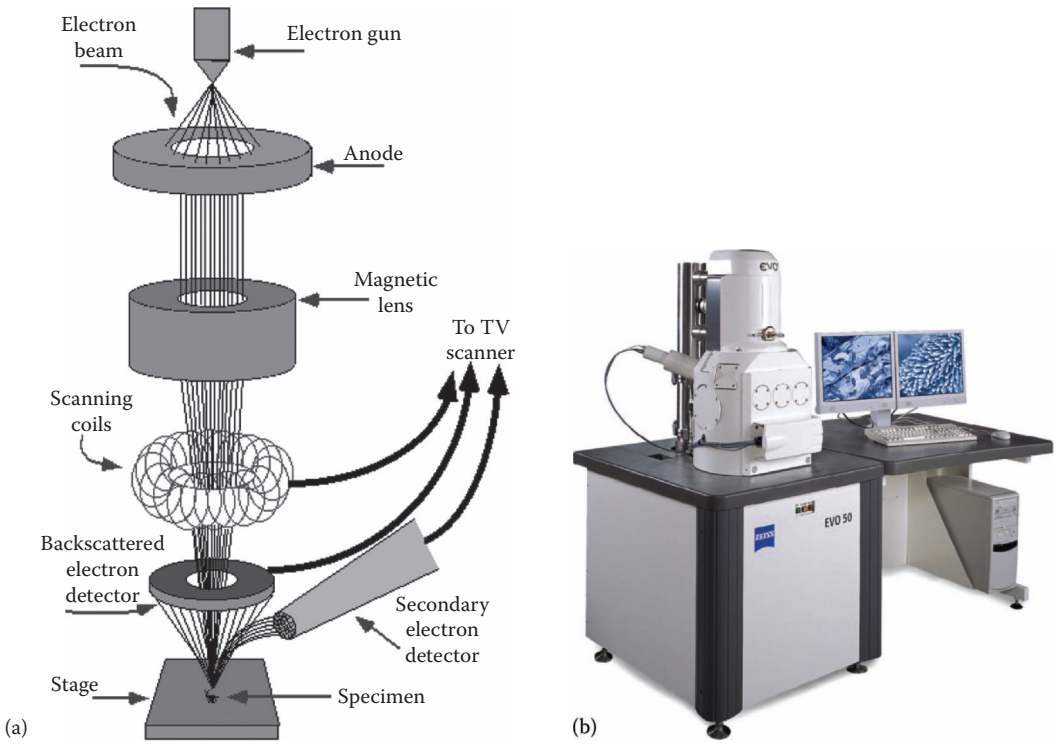


FIGURE 6.15 (a) Schematic of scanning electron microscope (SEM). (b) Carl Zeiss SEM. Scanning electron microscopy provides detailed three-dimensional view.

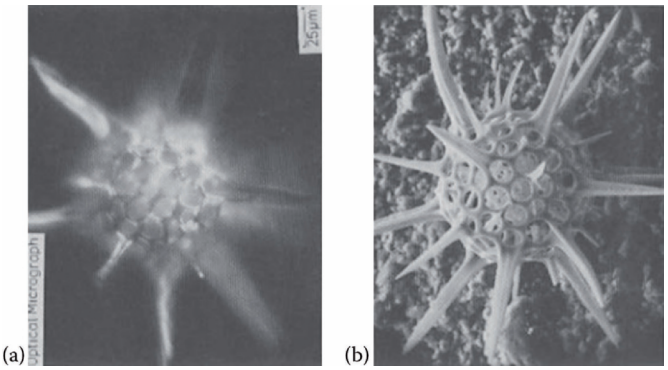


FIGURE 6.16 *Trochodiscus longispinus* in optical microscopy (a) and scanning electron microscopy (b). Note improved depth of field and resolving capability of the scanning electron microscopy experiment. (Taken from Goldstein, J., D. Newbury, D. Joy, C. Lyman, P. Echlin, E. Lifshin, L. Sawyer, and J. Michael. 2003. *Scanning electron microscopy and x-ray microanalysis.* New York: Kluwer Academic/Plenum Publishers.[6])

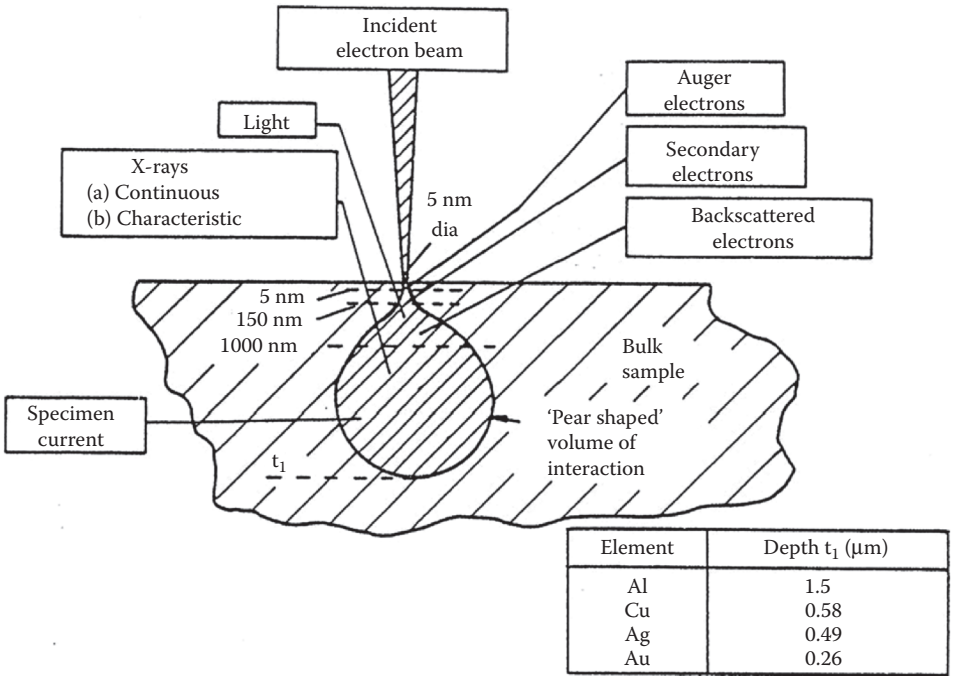
wavelengths. The effective F-number (also *F#* or *f*/*D*) for an electron microscope, on the other hand, is about 100 (quite large). The much shorter wavelength despite the poorer *f*/*D* gives an SEM a resolution 1000-fold better than that of the best light microscope. This improved resolution is dramatically illustrated in Figure 6.16, where we compare an image from an optical microscope (OM) with that of a SEM of the same *Trochodiscus longispinus*.

As the rastered focused electron beam irradiates the sample, secondary electrons are ejected from the sample. Secondary electrons are specimen electrons that obtain energy by inelastic collisions with beam electrons. They are defined as electrons emitted from the specimen with energy less than 50 eV and are predominantly produced by the interactions between energetic beam electrons and weakly bonded conduction-band electrons in metals or the valence electrons of insulators and semiconductors. With a flux amplitude that depends on the nature of the material and on the angle that the feature's surface makes with the incident beam, these secondary electrons are counted and used to create an image of the sample. The accuracy of the images captured is highly dependent on machine capability and the specific part being examined. An electron multiplier is used to detect and amplify the secondary electron intensity as a function of beam position to create the SEM image (see more on SEM detectors on page 308). The output generated from the electron detector is displayed on a CRT rastered in synchronization with the electron beam. The final result is a two-dimensional image on a screen. Since no coordinate data are directly output from the SEM, performing any analysis other than line width measurements directly with the SEM software becomes difficult. Thus, SEMs are ideal for visualizing MEMS parts but are inadequate tools for quantitative analysis.[2]

The type of particles produced from the interaction of an electron beam with a solid sample is not limited to secondary electrons but may include backscattered electrons (BSEs), elastically scattered electrons, characteristic x-rays, and other photons of various energies. One of the main features of the SEM is that, in principle, any radiation emitted from the specimen or any measurable change in the specimen can be used to provide the signal to modulate the CRT and thus provide the contrast.

From Figure 6.17, it is obvious that several processes occur when electrons bombard a solid surface and that different emission mechanisms arise from different depths within the sample. The two basic types of electron scattering events that take place in an SEM are elastic scattering, which causes backscattering of electrons without loss or little loss of kinetic energy, and inelastic scattering, which results in transfer of energy from the beam electrons to the sample. The inelastic scattering leads to the generation of secondary electrons but also to Auger electrons, x-rays (continuous and characteristic), electron-hole pairs in conductors and insulators, long wavelength electromagnetic radiation or cathodoluminescence [visible, UV, and infrared (IR) light], lattice vibrations (phonons), and electron oscillations in metals (plasmons, specimen current). The penetration depth of an electron into a solid depends on the beam energy, which in turn depends on the accelerating voltage, usually between 30 and 60 kV in a typical SEM and on the atomic number, *Z*, of the bombarded material. The probability of an incident electron being scattered varies, with larger atoms (large *Z*) having a higher probability of producing an elastic collision, so low-*Z* materials have a greater depth of penetration.

When a 20 KV e-beam is used, say for Ni (Z = 28), Auger electrons originate from very close to



Element	Depth t_1 (μm)
Al	1.5
Cu	0.58
Ag	0.49
Au	0.26

FIGURE 6.17 Interaction mechanism of electrons with a solid. A bulk specimen showing the spread of the beam and the activated volume, giving rise to backscattered, secondary, and Auger electrons (i), characteristic x-rays (ii), and electron current (iii). Typical values given for the depth of electron penetration and the excited volume for a 30-keV, 100-nm-diameter electron probe.

the surface (10–30 Å), secondary electrons from 100 Å deep, backscattering electrons from 1–2 μm deep, and x-rays from 5 μm deep (the depth to which the incoming electrons may penetrate). Cathodoluminescence (CL), which arises from the recombination of electron-hole pairs generated by the incident beam, consists of visible photons (light). These can escape the sample only when emitted close to the surface, making cathodoluminescence very surface specific.

The interaction volume of the e-beam with the sample (pear shape shown in Figure 6.17) can be made visible in photoresist materials such as PMMA. Molecular bond damage (scission) during electron bombardment renders these materials sensitive to etching in a suitable solvent revealing the exposed volume. This phenomenon is the basis for e-beam lithography as detailed in Volume II, Chapter 2. The interaction volume as a function of beam energy can be estimated from Monte Carlo calculations (see also Volume II, Figure 2.15).

In Figure 6.18 the relative intensity versus abundance of the various emission phenomena under e-beam radiation are indicated. Whereas surface topography is obtained when low energy secondary electrons are collected, atomic number or orientation information is obtained from higher-energy backscattered electrons and elastically scattered electrons. The latter may be used to differentiate between surface roughness, porosity, granular deposits, stress-related gross microcracks (often used in conjunction with microsectioning), and the observation of grain boundaries in unetched samples.

Elastically scattered electrons have the same energy as the incident electron beam and retain orientation and phase information. These electrons are used, for example, in diffraction imaging experiments, such as low-energy electron diffraction (LEED). In surface

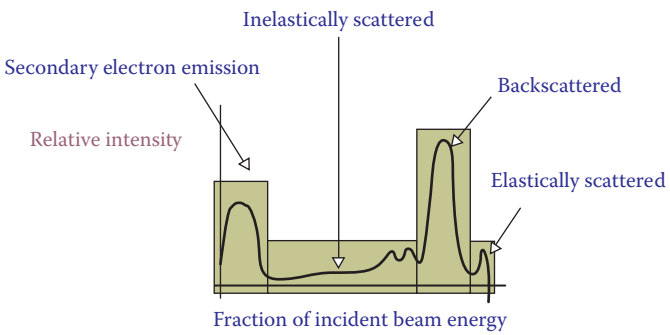


FIGURE 6.18 Relative intensity versus abundance of various emission phenomena under e-beam radiation.

science, well-ordered single-crystal surfaces are often investigated this way. Electrons penetrate (energy 50–300 eV) only a few monolayers into the sample; therefore, they are surface sensitive (for comparison, x-rays are more bulk sensitive). The observed diffraction pattern is the reciprocal lattice of the geometric surface lattice. Well-ordered surfaces exhibit sharp bright spots and low background intensity. The presence of surface defects and crystallographic imperfection results in broadening and weakening of the spots and increased background.

The backscattered electrons (BSEs) have an energy of about 80% of the incident beam. They are emitted more or less specularly. These electrons, just like secondary electrons, can be used to form topographic SEM images, but they have a higher energy than secondary electrons and also yield compositional information. Larger atoms (large atomic number, Z) have a higher probability of producing an elastic collision because of their greater cross-sectional area. Consequently, the number of BSEs reaching a BSE detector is proportional to the mean atomic number of the sample. Thus, a "brighter" BSE intensity correlates with greater average Z in the sample, and "dark" areas have lower average Z. BSE images are very helpful for obtaining compositional maps of a sample and for quickly distinguishing different phases. BSE images are limited to a grayscale range, though, because they only record one variable, average Z (a combination of all the elements in a sample). Thus, they do not convey as much information as can be obtained by elemental composition mapping using EDS* x-ray detectors (see "EDAX (EDX or EDS)").

SEM electron detectors can be based on different principles and oriented and operated differently to preferentially detect either secondary or backscattered electrons. A scintillator-photomultiplier system known as the *Everhart-Thorny* (E-T) *detector* detects the secondary electron signal, which is much more dependent on the sample topography than other electrons emitted from the sample. Topographic images obtained with secondary electrons actually look remarkably like images of solid objects viewed with light. A typical E-T detector consists of a Faraday cage in front of a scintillator, which in turn is coupled to a light pipe leading to a photomultiplier tube. The Faraday cage is kept at a positive potential of a few hundred volts so as to efficiently collect most of the secondary electrons emitted from the sample (remember they have an energy less than 50 eV). The scintillator typically has a positive voltage of several kilovolts applied to its surface, so that the electrons that pass the Faraday cage are accelerated onto the scintillator. When the electrons strike the scintillator they produce light, which is directed to the photomultiplier by a light pipe.

Using BSEs to construct topographical images either relies on a simplified scintillator–light pipe–photomultiplier (E-T), as used for secondary electrons, or uses solid state detectors. This stripped-down version of the E-T detector consists of scintillator coupled to a photomultiplier by a light pipe. Because there is no collection potential, the collection efficiency of these detectors is directly related to the scintillator size and its proximity to the sample. The simplest solid-state detector is a p-n junction with electrodes on the front and back of the sample. When hit by electrons, the holes tend to migrate to one electrode while the electrons migrate to the other, producing a current, the total of which is dependent on the electron flux and the electron energy. Response time of the detector is, unfortunately, slow but can be improved by putting a potential across the diode but this comes at the expense of increased noise in the signal. Often a set of solid-state detectors for simultaneous collection of backscattered electrons in different directions is used.

Detectors above the sample collect electrons scattered as a function of sample composition, whereas detectors placed to the side collect electrons scattered as a function of surface topography. BSE images can be obtained nearly instantly, depending on scan rate, and at any magnification within the instrument range.

The composition of samples is sensitive to the backscattered coefficient (η), which varies monotonically with Z. It has been shown that the coefficient can be expressed as follows.

$$\eta = -0.254 + 0.016Z - 1.86\times10^{-4}Z^2 + 8.3\times10^{-7}Z^3 \qquad (6.11)$$

* EDAX (EDX or EDS): Energy dispersive analysis of x-rays (EDAX) is also called *EDX* and, more and more frequently, EDS. See "EDAX (EDX or EDS)".

The magnitude of the compositional or atomic number contrast, C, from two phases of backscattered coefficients, η_1 and η_2, is readily calculated as follows.

$$C = \frac{\eta_1 - \eta_2}{\eta_1} \tag{6.12}$$

This method is not recommended for phases with similar atomic number, as the resolution may be poor.

The inelastically scattered electrons have interacted strongly with the substrate electrons, and in the process they have lost all information of their initial direction and phase. These include Auger electrons, photoelectrons, and e-h pair recombination processes. These have mainly spectroscopic value but can also be used for imaging (see SAM, "Auger Electron Spectroscopy and Scanning Auger Microscopy").

In general, a high vacuum is required in SEM, as the sample environment and the samples have to be clean, dry, and electrically conductive. An environmental SEM (ESEM) offers high-resolution secondary electron imaging in a gaseous environment of practical any composition, at pressures as high as 50 Torr and temperatures as high as 1500°C. In this case wet, oily, dirty, nonconductive samples may be examined in their natural states without modification or preparation. An immediate concern is that the gas molecules will scatter the electrons and degrade the beam. To avoid this, instead of using a single pressure-limiting aperture as used in conventional SEM, the ESEM uses multiple pressure limiting apertures to separate the sample chamber from the vacuum. This leaves the column still at very high vacuum, but the sample chamber may sustain pressures as high as 50 Torr.

An ESEM uses a special type of detector that can function in a nonvacuum environment. The detector uses the principle of gas ionization. A positive potential of a few hundred volts is applied to the detector, which attracts the secondary electrons emitted by the sample. As the electrons accelerate in the detector field, they collide with gas molecules, creating additional electrons, amplifying the original secondary electron signal, and generating more positive ions. The detector collects the amplified secondary electron signal and passes it on to an electron amplifier.

In nonconductive samples, the positive ions created in the gas ionization process are attracted to the sample surface, and they effectively suppress charging artifacts.

A short review of environmental scanning electron microscopy by Krisada Kimseng and Marcel Meissel can be found at http://www.calce.umd.edu/TSFA/ESEM.pdf.

Scanning Proximal Microscopes

Introduction

A scanning proximal microscope (SPM) is a mechanical imaging instrument in which a small (<1 μm) probe is scanned over a sample surface. All SPMs owe their ultimate existence to the scanning tunneling microscope (STM), invented by IBM research scientists Gerd Binnig and Heinrich Rohrer in the early 1980s.

By monitoring the motion of a tiny probe tip, the surface topography and/or images of surface physical properties are measured with an SPM. Because close proximity or contact between the specimen and probe is a general requirement, SPMs need a feedback system that precisely controls the physical separation between probe and sample. Several methods are available to monitor the z-position of the probe tip at all times:

- Interferometric measurement of the tip amplitude
- Electron tunneling (limited to conductive specimens)
- Detection of the light emitted through the probe tip
- Constant force (atomic force feedback) is the most common method; it can be further subdivided as follows:
  - Diffraction of a separate light source by the tip
  - Mechanical sensor attached to the tip (e.g., a quartz tuning fork)
- Capacitance detection

The two most commonly used mechanisms of tip positioning are optical methods that monitor the tip vibration amplitude (usually interferometric), and nonoptical tuning fork techniques.

In addition to the z-axis feed-back control, an x-y-z scanner (usually driven piezoelectrically) controls the movement of the probe over the sample

surface. The *x-y-z* scanner moves either the sample or the proximal probe. With the sample attached to the scanner, the sample moves under the fixed probe tip in a raster pattern, and an image is generated from the interaction of the tip and the sample. The size of the area that can be imaged depends on the maximum displacement of the scanner used. The typical size scale of features measured ranges from the atomic level (less than 1 nm) to more than 100 μm. A computer monitors the probe position and gathers data from the feedback system to control both the scanning of the tip (or sample) and the separation of the tip from the sample surface. The information generated by sensing the interaction between the probe and sample surface is collected and recorded by the computer point-by-point during the raster movement, and the computer renders the data into two-dimensional data sets (lines). The interactions between probe and sample used for imaging in scanning probe microscopes include electron tunneling, magnetic forces, electrical forces, electrochemical interactions, mechanical interactions, capacitance, ion conductance, Hall coefficient, thermal properties, and optical properties. SPMs offer an alternative to noncontact techniques and are characterized by their potential for very high resolution (subangstrom), limited only by the probe size and the property probed.

As illustrated in Table 6.5, SPMs are used in a wide range of applications. The three most used SPMs are the near-field scanning optical microscope (NSOM), the STM, and the atomic force microscope (AFM). In this section, we analyze the potential of these three SPMs for geometrical metrology. The utility of other specialized SPMs in property metrology is surveyed below (see "Tools to Measure Materials Property: Property Metrology").

NSOM or SNOM

To attain resolution beyond Abbe's diffraction limit of 200 nm, the textbook knowledge used to be that a complicated and costly SEM (0.1 nm resolution) or STM was required. Although these instruments do have a better resolution, they sacrifice many of the advantages of an optical microscope, such as nondestructiveness, low cost, high speed, reliability, versatility, accessibility, and ease of use. Thus, it would be very attractive to miniaturize a light microscope and have it operate below Abbe's diffraction limit. This objective is met by combining a scanning proximal probe technique with optical microscopy (OM) in so-called *scanning near-field OM*, a technique developed in the mid-1980s. We encountered NSOM first in Volume I, Chapter 5; here we give some additional background supplementing the information presented there.

The idea of using a subwavelength aperture to improve optical resolution was first proposed by E. H. Synge[7] in a letter to Einstein in 1928 and again, independently, in 1956 by John A. O'Keefe.[8] These ideas were realized only much later. First results were

Table 6.5 Scanning Proximal Microscopes and Their Applications

Scanning Proximal Microscopes	Examples of Applications
Scanning tunneling microscopy	Topography, density of states
Atomic force microscopy	Topography, force measurement
Lateral force microscopy	Friction
Magnetic force microscopy	Magnetism
Electrostatic force microscopy	Charge distribution
Scanning near-field optical microscopy	Optical properties
Scanning capacitance microscopy	Dielectric constant, doping
Scanning thermal microscopy	Temperature
Ballistic electron emission microscopy	Interface structure
Spin-polarized scanning tunneling microscopy	Spin structure
Kelvin probe force microscope is the Kelvin mode of scanning probe microscopy	Measures contact potential difference between the probe and the sample
Scanning electrochemical microscopy	Electrochemistry
Photon emission scanning tunneling microscopy	Chemical identification

achieved in 1972, by Ash and Nicholls,[9] in the microwave region of the spectrum, and in 1984, in the visible, by D. W. Pohl et al.[10] During image acquisition in NSOM, the point light source must be scanned over the surface, without touching it, and the optical signal from the surface must be collected and detected. To realize a constant probe-sample distance, a feedback mechanism that is unrelated to the NSOM/SNOM signal is utilized. Harootunian et al.[11] were the first to demonstrate SNOM microscopy in fluorescence mode. This result led eventually to single-molecule detection experiments by several groups.

During image acquisition in NSOM, a point light source is scanned very close to the surface, without touching it, and the optical signal from the surface is detected. In order to achieve an optical resolution better than Abbe's diffraction limit (the resolution limit of conventional OM), the probe tip must be brought within the near-field region, with the illumination source closer to the specimen than the wavelength of the illuminating radiation. For NSOM, the separation distance between probe and specimen surface is typically on the order of a few nanometers. Radiation near the source is highly collimated within the near-field region, but after propagation of a few wavelengths' distance from the specimen, the radiation experiences significant diffraction and enters the far-field regime. To make a point light source, laser light is fed to a microscopic aperture via an optical fiber. The aperture can be a tapered fiber with its shaft coated with a metal, a microfabricated probe, or a tapered pipette (see example aperture SNOM implementations in Volume I, Chapter 5). The aperture-based resolution is limited to about $\lambda/10$. With careful probe design and sample preparation a resolution of ~30 nm has been achieved.

One challenge for aperture based SNOM is the small aperture resulting in a very low light throughput. Scattering SNOM or apertureless SNOM, developed in the 1990s, is an alternative that avoids this low light intensity problem. Apertureless near-field optics, as we saw in Volume I, Chapter 5, uses a sharp, metallic subwavelength tip instead of an optical point source. The sharp metal tip acts as a nanoscopic light scatterer when placed in near-field distance to the sample. Oblique incident light is scattered at the tip, and the resulting strongly enhanced local plasmonic field is used for excitation of the sample. The resulting optical response signal is collected via an objective in far field and can be used for imaging and chemical analysis. The tip radius in this case is the only factor determining resolution. The best resolution attained today is 10 nm, and it is expected that this can be reduced to atomic level resolution. A significant problem with the scattering SNOM approach is that the light scattered by the tip carrying the information on tip-sample interaction is easily overwhelmed by background light. Therefore the method is particularly suitable for light-matter interaction processes where the light detected has a different wavelength from the excitation light (e.g., Raman scattering, fluorescence, second harmonic generation). Consider tip-enhanced Raman spectroscopy as an example. In tip-enhanced enhanced Raman spectroscopy (TERS), a silver-coated tip is scanned across the illuminated area, acting as a Raman signal enhancer. As we saw in Volume I, Chapter 5, Raman signals of molecules deposited on silver or in close proximity to silver nanoparticles can be increased by a factor of 10^{14}. It is believed that this enhancement results principally from the resonant enhancement of the surface plasmon mode localized at the apex of the sharp metallic tip. This technique represents an elegant combination of two Nobel Prize-winning discoveries: the Raman effect and the scanning tunneling microscope. TERS has created a new tool that achieves a long-awaited goal in surface science, namely the correlation of topographic and chemical information of the same sample region at nanometer scale. Although the enhancement processes allowing for TERS are still not fundamentally understood, the use of an illuminated scanning probe tip to greatly enhance Raman scattering from the sample underneath the tip is one of the most exciting developments in optical spectroscopy.

In the rest of this discussion, we focus on aperture-based SNOM only.

In an aperture-scanning NSOM instrument, the quantitative point-spread function in the near field can be approximated by a Gaussian profile whose 1/e intensity value is of the same order as the radius of the aperture at the tip of the NSOM probe. The mode of light propagation is primarily evanescent (and parallel to the specimen surface) when the

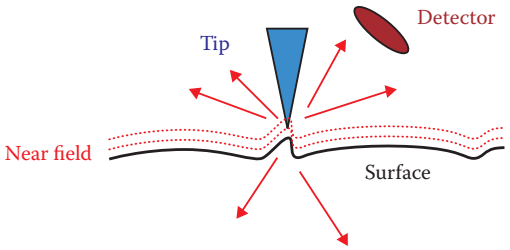


FIGURE 6.19 Light source is held in near-field to the surface.

radius of the illuminating source is less than one-third of the imaging light wavelength as shown in Figure 6.19. The NSOM configuration illustrated here positions the objective in the far field, in the conventional manner, for collection of the image-forming optical signal.

A schematic illustrating the control and information flow of an inverted optical microscope-based NSOM system is presented in Figure 6.20a. A laser source is coupled to a fiber optic probe for near-field specimen illumination. In the NSOM setup shown, the probe tip movements are controlled through an optical feedback loop incorporating a second laser that is focused on the tip. Two-dimensional data sets gathered by the NSOM instrument are registered and displayed as a 3D reconstruction on a computer monitor. Figure 6.20b is a photograph of an actual near-field scanning optical microscope set-up.

To improve signal-to-noise ratios for the feedback signal, the NSOM tip is oscillated at its resonance frequency. This allows for more sensitive lock-in detection techniques, eliminating positional detection problems associated with low-frequency noise and drift. As the oscillating tip approaches the sample surface, forces between the tip and specimen change both the peak resonance and the quality factor. These changes can be monitored by several techniques, classifiable in two broad groups: 1) shear-force mode and 2) tapping mode. In the shear-force mode, to control the tip-specimen gap during imaging, one utilizes lateral oscillation shear forces generated between a straight optical fiber and the sample (Figure 6.20a). The fiber is drawn to a sharp point and coated on the sides with an opaque material (usually Al). A small (~50-nm) hole is formed in the coating at the apex of the fiber and the evanescent field associated with this aperture is used to image the sample. In the optical feedback approach, illustrated in Figure 6.20a, a second laser is tightly focused as close to the end of the NSOM probe as possible. With the laser feedback established, the probe is then vibrated, at a known frequency, utilizing a dither piezo. A split photodiode collects the laser light, and the difference between the signals from each side of the detector is determined. The main problem associated with such an optical type of feedback mechanism is that the light source, which is used to detect the tip vibration frequency, phase, and amplitude, becomes a

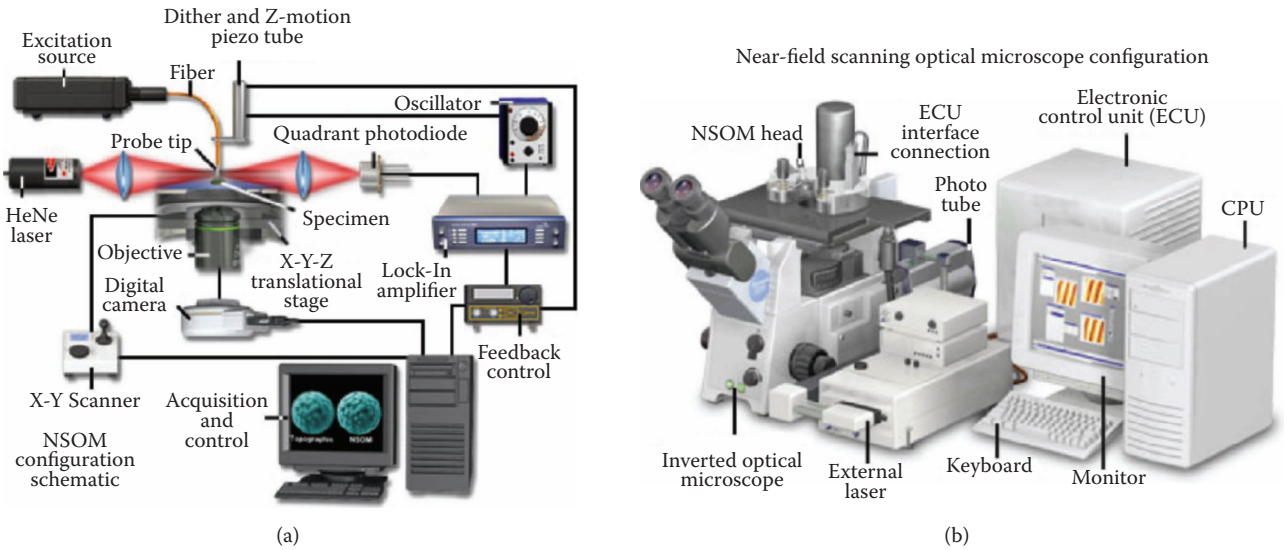


FIGURE 6.20 A schematic (a) illustrating the control and information flow of an inverted optical microscope-based scanning near-field optical microscopy system. (b) Actual photo of an Olympus NSOM system (http://www.olympusmicro.com/primer/techniques/nearfield/ nearfieldintro.html).

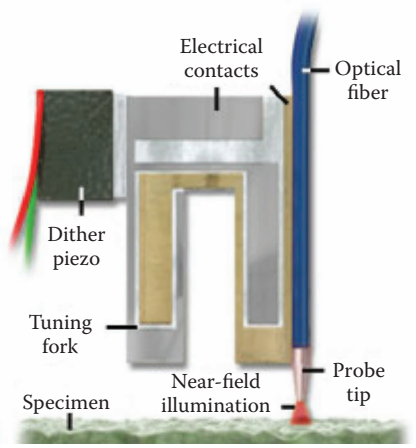


FIGURE 6.21 Tuning fork feedback (http://www.olympusmicro.com/primer/techniques/nearfield/nearfieldintro.html).

potential source of stray photons that can interfere with the detection of the NSOM signal. To realize a constant probe sample distance, a feedback mechanism that is unrelated to the NSOM signal is obviously preferred. This is the primary reason why the nonoptical feedback method based on the tuning-fork technique, illustrated in Figure 6.21, has become increasingly popular. Tuning forks are incorporated into the NSOM to serve as inexpensive and simple, nonoptical excitation and detection devices. When quartz tuning forks are utilized for regulation in a feedback loop, their very high mechanical quality factor, Q (as high as approximately 10,000), and corresponding high gain provide the system with high sensitivity to small forces, typically on the order of piconewtons. The fiber is typically fixed to one leg of a tuning fork, which is caused to oscillate. Interaction between the tip and the sample (due to shear forces—hence the name shear-force mode) leads to damping of this oscillation, and the height is adjusted to maintain a constant oscillation frequency.

In the tapping mode, to generate the feedback signal, one relies on atomic forces modulating the oscillation of a cantilever probe. In this case, a bent optical fiber, effectively functioning as an AFM cantilever probe, may be used as the light source. The nonoptical feedback method based on the tuning-fork technique, illustrated in Figure 6.21, is also applicable here. It is also possible to use an AFM-type cantilever probe, in which the cantilever has an aperture at its apex; a lens focuses light into the aperture and an AFM-type feedback mechanism is used.

For the best image quality, shear-force feedback techniques are restricted to specimens that have low surface relief, and longer scan times are required compared with operation in tapping mode. Moreover, the interactions that underlie shear force imaging are not very well understood. However, the straight probes employed in shear-force feedback techniques are easier to fabricate, are lower cost, and have a higher light throughput than their bent probe counterparts.

Previously developed high-resolution scanning techniques, such as scanning electron microscopy (SEM), transmission electron microscopy (TEM), scanning tunneling microscopy (STM), and atomic force microscopy (AFM), do not benefit from the wide array of contrast mechanisms available to an OM such as SNOM and, in most cases, are limited to the study of specimen surfaces only. Aside from the available contrast-enhancing techniques of staining, fluorescence, polarization, phase contrast, and differential interference contrast, optical methods have inherent spectroscopic and temporal resolution capabilities.

Advantages and disadvantages of SNOM and a list of applications are provided in Volume I, Chapter 5, "NSOM Applications."

Scanning Tunneling Microscope

The scanning tunneling microscope (STM) was the first and remains the most famous of the scanning probe microscope techniques. As detailed in Volume I, Chapter 3, scanning tunneling microscopy measures topography of surface electronic states using a tunneling current that is dependent on the separation between the probe tip and a conductive sample surface (metals or semiconductors). The violation of classical physics in tunneling is allowed by the quantum mechanical uncertainty principle. On the basis of this principle, a particle can violate classical physics by tunneling inside and even through an energy barrier, ΔE, for a short time, $\Delta t \sim \hbar/\Delta E$ (see Volume I, Equation 3.107). The exponential decay of the wave function inside a rectangular energy barrier that stretches from $x = 0$ to $x = L$ and has a height of V_0 is given as follows:

$$\psi(x) = Ae^{-\alpha x}$$

(Volume I, Equation 3.185)

with $\alpha^2 = \frac{2m(V_0 - E)}{\hbar^2}$ (Volume I, Equation 3.165). If the barrier is narrow enough (L is small), there will be a finite probability, P, of finding the particle on the other side of the barrier, as follows:

$$P = |\psi(x)|^2 = A^2 e^{-2\alpha L} \quad \text{(Volume I, Equation 3.186)}$$

where A is a function of the energy, E, of the particle and the barrier height, V_0. The probability of finding an electron on the other side of a barrier of width L can be probed with a fine needle tip from a STM. The tunneling current, picked up by the sharp needle point of an STM, based on Volume I, Equation 3.186, is given by the following equation:

$$I = f_w(E) A^2 e^{-2\alpha L} \quad \text{(Volume I, Equation 3.187)}$$

where $f_w(E)$ is the Fermi-Dirac function, which contains a weighted joint local density of electronic states (DOS) in the solid surface that is being probed and those states in the needle point (see Figure 6.22). Tunneling occurs between states of the same energy and the electron's energy does not change during the tunneling event. From Volume I, Equation 3.187, when L changes by 1 Å, the current changes by a factor of about 10. Obviously, the tunneling current is very sensitive to the gap distance. The size of the gap in STM experiments is kept on the order of a couple of angstroms. The DOS can vary as a function of position on the sample, which means that one can define a local density of states (LDOS). LDOS depends on both energy and on position. An intuitive way to think about LDOS is that it gives the density of electrons of a certain energy at that particular spatial location [LDOS (x, y, E)]. An STM measures and controls the current that flows between the tip and the sample (I), the bias voltage between the tip and the sample (V), the xy (in sample plane) position of the tip, and the z (perpendicular to sample plane) distance between the tip and sample. Using these variables, an STM is able to measure the LDOS of a material as function of position on the surface (controlled by where the tip is above the surface) and as a function of energy (controlled by the bias voltage between the tip and sample). The LDOS is proportional to the differential increase in tunneling current given a differential increase in bias voltage; in other words, one can measure the LDOS by measuring dI/dV. This was illustrated in Volume I, Figure 3.135, for a metallic and semiconducting carbon nanotube where singularities (sharp peaks) in the density of states function are observed directly in scanning tunneling microscopy images.

An STM may be viewed as an electron microscope that uses a single atom on a tip to attain atomic resolution. In a bit less lofty way, it may be described as a glorified phonograph needle, as illustrated in Figure 6.23. Resolution is determined by manufacturing a probe whose tip is of the same dimension we seek to resolve. If one can get a single atom at the tip, the vast majority of the current will run through it and thus give us atomic resolution. Tungsten is commonly used for the tip because one can use electrochemical-etching techniques to create very sharp tips like the one shown in Figure 6.24. This tip is brought to within a few angstroms of the surface that we wish to probe, and a small bias voltage (10 mV to 2 V) is applied between the tip and the sample while an electrical

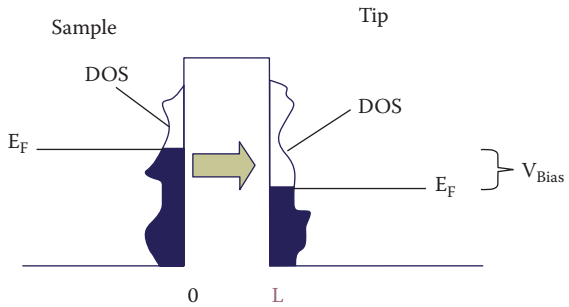


FIGURE 6.22 Density of state functions in sample and tip.

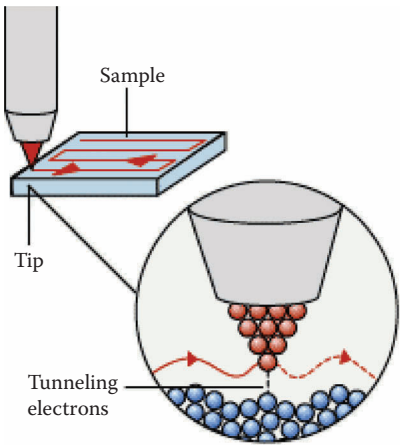


FIGURE 6.23 Scanning tunneling microscope.

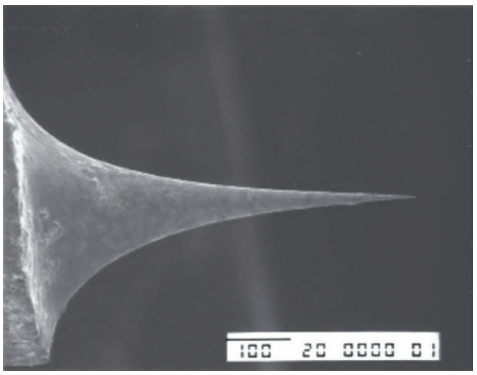


FIGURE 6.24 Tungsten scanning tunneling microscope tip; magnification ×150.

current flows by electrons tunneling across the gap between the two conductors. As we saw above, this tunneling current magnitude depends upon the gap voltage, the gap distance, and the local density of electron states (Volume I, Equation 3.187).

At its heart, a scanning probe microscope has a piezoelectric material (e.g., quartz) that expands or contracts slightly when a voltage is applied to it. These voltages can be carefully controlled and cause one end of the crystal to move with better precision than the width of a single atom. The needle or tip is brought very close to the surface and moved over it piezoelectrically and a tunneling current, resulting from the overlap of the tip's electron wave functions with the conducting surface is measured. This tunneling current is generated, even though there is a break in the circuit. Using a feedback servo loop, the tip slowly scans across the surface and is raised and lowered in order to keep the signal constant and maintain a distance of a couple of angstroms, where the tunneling current is at a measurable level.

An STM does not measure nuclear position directly; rather, it measures the electron density clouds on the surface of the sample. In some cases, the electron clouds represent the atom locations pretty well, but not always. The trace from an STM (Figure 6.25A) can be interpreted as a grid (Figure 6.25Ba) that can be shown as a grayscale picture (Figure 6.25Bb). The grayscale picture can be interpreted as a contour map (Figure 6.25Bc), which can then be averaged out to make it smooth (Figure 6.25Bd) and finally can be colored (Figure 6.25C). This explains the color of the Kanji character, made by manipulation of individual iron atoms on a copper surface with an STM, in Figure 6.26.

The vertical resolution of an STM is remarkably good because of the exponential dependence of the current on vertical height (L in Volume I, Equation 3.187). This means that 1-Å change in height corresponds to almost 1 order of magnitude difference in current. The noise floor is in the tens of femtometer

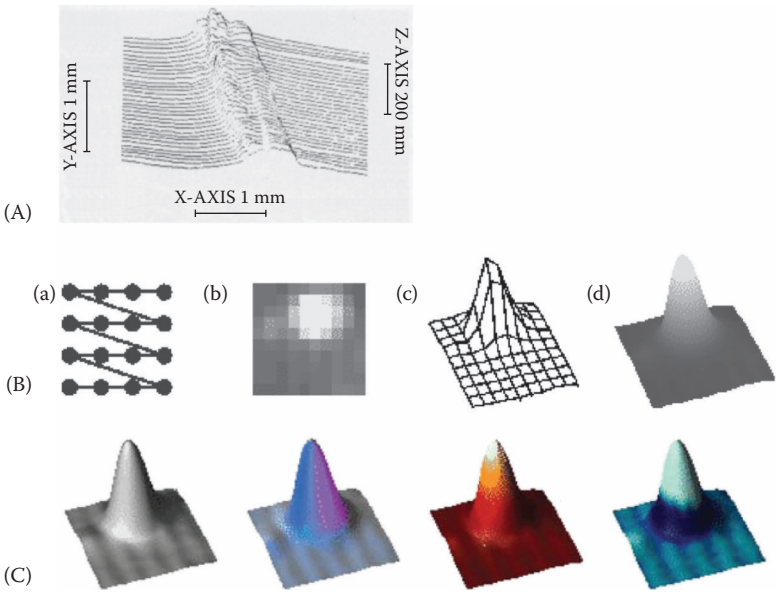


FIGURE 6.25 (A) Trace of a scanning tunneling microscope (STM). Shown is a trace of an individual turn mark on a diamond-turned aluminum substrate to be used for subsequent magnetic film deposition for a high-capacity hard disc drive. (B) The trace from an STM can be interpreted as a grid (a), which can be shown as a grayscale picture (b). The grayscale picture can be interpreted as a contour map (c) that can then be averaged out to make smooth (d) and finally colored (C).

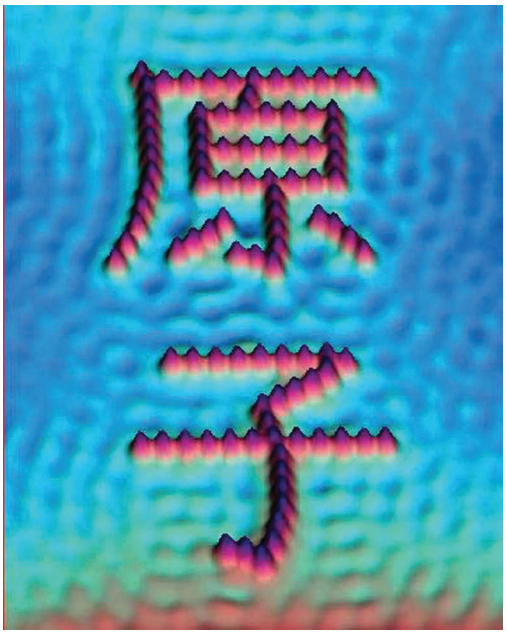


FIGURE 6.26 Kanji character for *atom*, by Lutz and Eigler (IBM Almaden). An example of manipulation/mechanosynthesis of iron atoms on a Cu surface with a scanning tunneling microscope. (Courtesy of D. Eigler.)

range. The lateral resolution depends on the tip radius as clarified in Figure 6.27. A tip with radius R = 1000 Å actually focuses 90% of the current in a circle of radius (Δx in Figure 6.27) of 45 Å. A 100-Å tip focuses 90% of the current in a circle of radius 14 Å. In practice, a 2-Å lateral resolution is routine. These numbers can be ascertained by rewriting $I = f_w(E)A^2e^{-2\alpha L}$ (Volume I, Equation 3.187) as:

$$I(\Delta x) = f_w(E)A^2e^{-2\alpha\frac{\Delta x^2}{2R}} \qquad (6.13)$$

There are two major modes of operation for an STM: constant current and constant height. In constant current mode one maintains a constant tunneling current by adjusting the height (separation between tip and sample). In constant height mode one, maintains the height and measures the current.

STMs can be used in solution, as there still is a vacuum gap, even in water. This mode of STM operation is called *scanning electrochemical microscopy*, which we encountered in Volume II, Chapter 4. To minimize faradaic processes the tip shank must be shielded with melted wax or plastic to expose only the last few nanometers of metal. The tunneling current to the exposed metal tip must be large compared with faradaic current to the tip shank.

STMs are limited to imaging parts with conductive surfaces, and electronic inhomogeneities can have significant effects on the acquired topographical images. Vibrations in the probing mechanism limits gap width stability, which, in turn, can affect the fidelity of the measurements. All SPMs are limited, in the same sense as white light interferometers, to a maximum measurable slope. When features with perpendicular sidewalls are scanned, the data typically exhibit a slope that is actually not present. The height of measurable features is also limited to the probe tip length, which is typically <10 μm in commercial systems. This limitation severely prohibits the inspection of high-aspect ratio parts with dimensions on the order of millimeters. Though these tools have extremely high resolution, it is unfeasible to collect scans that cover all of the surfaces of a part, given the limited scan range of the tools.

Gerd Binning and Heinrich Rohrer (Figure 6.28) invented the scanning tunneling microscope at IBM

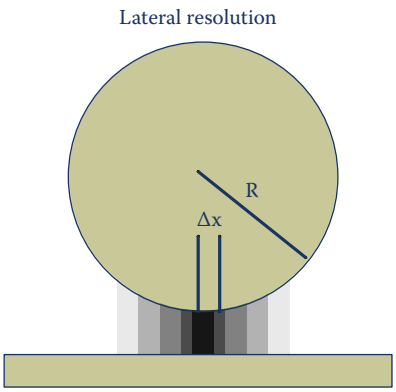


FIGURE 6.27 Lateral resolution of a scanning tunneling microscope.

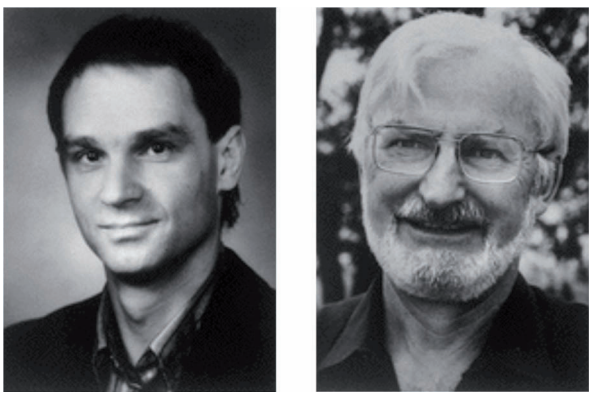


FIGURE 6.28 The inventors of the scanning tunneling microscope. Nobel laureates in Physics, 1986.

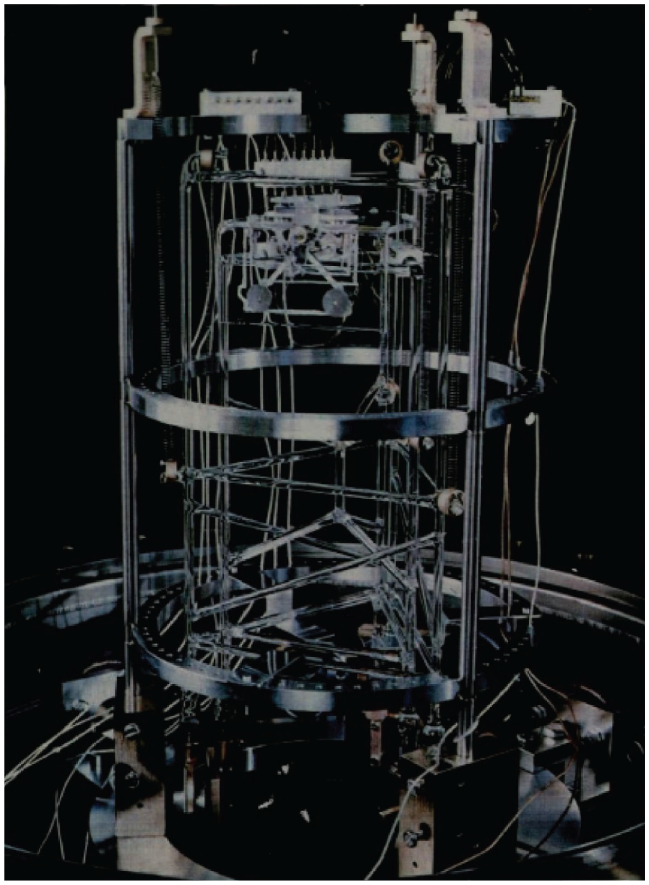


FIGURE 6.29 Binning and Rohrer's original scanning tunneling microscope.

Zürich in 1981. They shared the Nobel Prize for Physics in 1986 for their microscope (Figure 6.29).

Atomic Force Microscopy

Scanning force microscopy, also called *atomic force microscopy* (AFM), constitutes a major extension of the STM technique. The technique, first described in 1986, depends on forces (repulsive or attractive) between atoms of a sample and a probe (Figure 6.30).

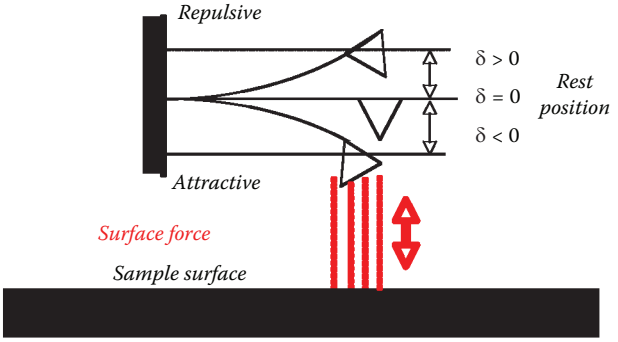


FIGURE 6.30 An atomic force microscope measures surface forces.

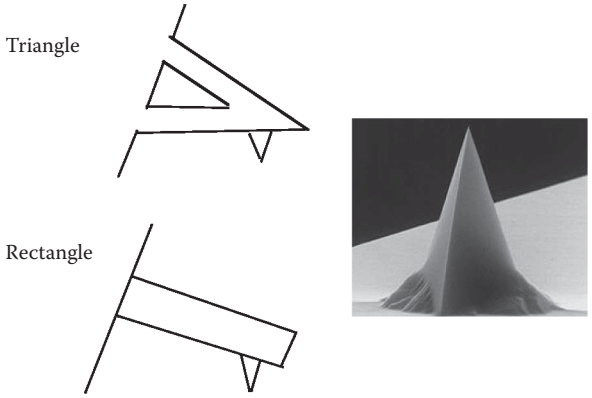


FIGURE 6.31 Cantilevers are either triangular or rectangular. Material can be silicon on silicon nitride (plastic in development). Probes shapes can be pyramidal or conical and may be sharpened.

No current flows between sample and AFM tip, so the sample need not be conducting.

In a typical AFM, a cantilever with a tip (the probe; see Figure 6.31) is oscillated above the sample surface. The oscillation amplitude is less than10 nm, and in the so-called *noncontact mode*, the tip does not touch the sample. Constant oscillation amplitude is maintained and surface forces, which extend from 1 to 10 nm above the surface, decrease the resonant frequency of the cantilever. This in turn changes the amplitude of oscillation. Typical surface forces range from the piconewton to the nanonewton [quantum mechanical (covalent, metallic bonds): 1–3 nN; Coulomb (dipole, ionic): 0.1–5 nN; polarization induced (dipoles): 0.02–0.1 nN].

A cantilever must now be designed to convert these very small mechanical forces into measurable deflection, δ.

A microfabricated cantilever beam and probe tip, as shown in Figure 6.32, deflects in response to an applied force and can be represented as a linear elastic, Hookean spring:

$$F = -k\delta \tag{6.14}$$

where δ = displacement at the end of cantilever (in meters) (this is what we actually measure in a force spectroscopy experiment)
F = external force applied to cantilever (in newtons) (this we can calculate from δ)
k = cantilever spring constant

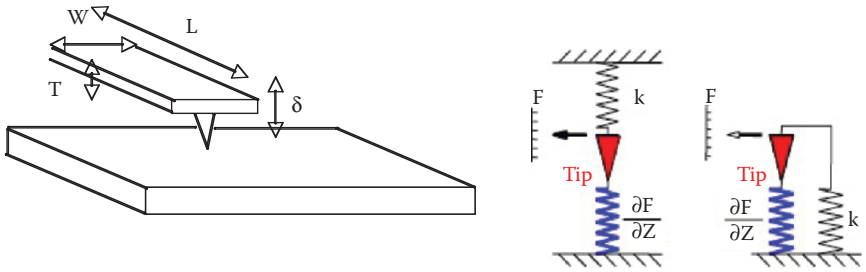


FIGURE 6.32 Atomic force microscope cantilever model. T = cantilever thickness (m), W = cantilever width (m), L = cantilever length (m), and δ = deflection of the end of the cantilever.

For the cantilever in Figure 6.32 the spring constant (in N/m) is given as follows:

$$k = \frac{EW}{4}\left(\frac{T}{L}\right)^3 \tag{6.15}$$

where T = cantilever thickness (in meters)
W = cantilever width (in meters)
L = cantilever length (in meters)
E = Young's (elastic) modulus of cantilever material (in pascals or N/m^2)

For Si, $E = 11 \times 10^{10}$ N/m^2.

We can calculate k independently, and typical values are 0.01–100 N/m. The resonance frequency f, of the cantilever is given as follows:

$$f = 0.162\sqrt{\frac{E}{\rho}}\frac{T}{L^2} \tag{6.16}$$

where ρ is the density of the cantilever material. Typical values for f are 7–500 kHz.

The interaction of surface forces with this cantilever can now be modeled as a spring in series with the cantilever spring, as pictured in Figure 6.32 or, remembering that $\omega = \sqrt{\frac{k}{m}}$:

$$\omega = \sqrt{\frac{k - \frac{\partial F}{\partial z}}{m}} \tag{6.17}$$

where $\partial F/\partial z$ is the force gradient (F'). For small amplitudes, a Taylor expansion of the force can be carried out, and a frequency shift proportional to the force gradient is found, as follows:

$$\frac{\Delta\omega}{\omega} = -\frac{1}{2k}\frac{\partial F}{\partial z} \tag{6.18}$$

In practice one needs to compromise between a cantilever with a small k and a large resonance frequency f. The resonance frequency (f) should indeed be higher than the data acquisition rate and the noise level (kHz range). To increase the resonance frequency, the cantilever should be stiff (large k), short and thin. On the other hand, to maximize sensitivity, a soft cantilever (small k) is preferred. But making k too small, one might have a problem of cantilevers crashing into the sample. With a k value of ~1 N/m, an AFM can measure forces of pN, and even fN is feasible. The force detection limit is the Brownian motion of the cantilever (thermal noise limited).

In Table 6.6 we list some example AFM cantilevers with their force sensitivity, Young's modulus, spring constant, resonant frequency, density and type of material, and shape. For an excellent discussion and references on cantilever design, consult http://www.asylumresearch.com/springconstant.asp.

Example 6.2

Exercise: In atomic force microscopy, an ultrasharp tip on the end of a silicon cantilever beam is used to probe a surface at the nanoscale. By how much is the beam deflected by thermal motion? The cantilever has a spring constant of ~10 N/m.

Answer: The energy required for deflection of the beam by a distance δ is $E = ½\ k\delta^2$. At a temperature of 300 K, the thermal energy is on the order of $kT = 6 \times 10^{-21}$ J.

This energy will cause an average deflection of the beam by $\delta = (2E/k)^{0.5} = 1 \times 10^{-7}$ m, or 100 nm.

A light lever is the most common deflection measurement (δ) sensor used in atomic force microscopes. In this approach, when the cantilever moves up and down, the position of a laser spot on a photodetector moves up and down. As a

Table 6.6 Example Atomic Force Microscope Cantilevers

Force Sensitivity (N)	k (N/m)	ω	ρ (g/cm^{-3})	Module Young E (10^{10} N/m^2)	Dimensions (mm)	Material
~10^{-8}	5000	22 kHz	8.9	22	4 × 0.25 diameter	Ni wire
~10^{-11}	2.5	33 MHz	2.3	11	2 × 0.5 × 0.1 μm	Si bar
~10^{-14}	0.004	8 kHz	3.1	32	0.2 × 0.036 × 0.3 μm	Si_3N_4 (V-shape)
10^{-12}	0.4	14 kHz	2.2	7	0.25 × 5 μm diameter	Quartz fiber

position-sensitive photodetector, a four-quadrant photodiode (current in each quadrant changes with light intensity) is used, as illustrated in Figure 6.33.

In an alternative approach a piezoresistive sensor is embedded in the cantilever. This approach was discussed in Volume I, Example 4.1.

There are three main AFM techniques in use: contact, noncontact, and tapping mode. In the "repulsive" contact AFM, the tip is in perpetual contact with the sample as it slides across the surface. The tip is attached to the end of a cantilever with a low spring constant, *k*, while the feedback keeps the vertical deflection of the cantilever constant. In contact mode, the AFM measures hard-sphere repulsion forces between the tip and sample. Forces range from nano- to micronewtons in ambient conditions and even lower (0.1 nN or less) in liquids. This method provides the highest resolution but may damage the surface. It is the only AFM technique that can provide atomic-resolution images.

In noncontact mode, the AFM derives topographic images from measurements of the van der Waals attractive forces. The tip does not touch the sample; instead, the cantilever with the tip oscillates (amplitude <10 nm) above the surface. The resonant frequency of the cantilever is decreased by the attractive surface force, which extends from 1 to 10 nm above the surface (see Volume I, Chapter 7, "Calculating Interaction Forces between Surfaces and Particles"). The noncontact mode has the lowest resolution of the three main AFM approaches.

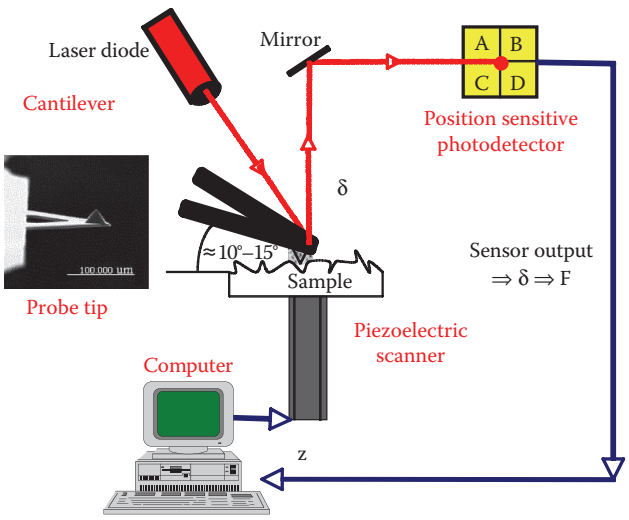


FIGURE 6.33 A light lever atomic force microscope setup.

Tapping mode AFM, the most commonly used of all AFM modes, is a patented technique (Veeco Instruments) that maps topography by lightly tapping the surface with an oscillating probe tip. This eliminates the shear forces potentially damaging to soft samples. The AFM tip taps the sample surface during the closest point of approach in each oscillation cycle, the amplitude of which is kept unchanged by the feedback loop. In this intermittent tip contact, force is measured with improved resolution over noncontact AFM. Surface forces again cause an oscillation frequency shift, but the tap also causes the cantilever/tip to lose energy, giving additional information about the surface. Another major advantage of tapping mode is related to limitations that can arise because of the thin layer of liquid that forms on most sample surfaces in an ambient imaging environment, i.e., in air or some other gas. The amplitude of the cantilever oscillation in tapping mode is typically on the order of 20–100 of nanometers, which ensures that the tip does not get stuck in this liquid layer. The amplitude used in noncontact AFM is much smaller (<10 nm; see above), and as a result, the noncontact tip often gets stuck in the liquid layer. In general, tapping mode is much more effective than noncontact AFM for imaging larger scan sizes that may include large variations in sample topography. Tapping mode can be performed in gases, liquids, and some vacuum environments.

The presence of a feedback loop is one of the subtler differences between AFMs and older

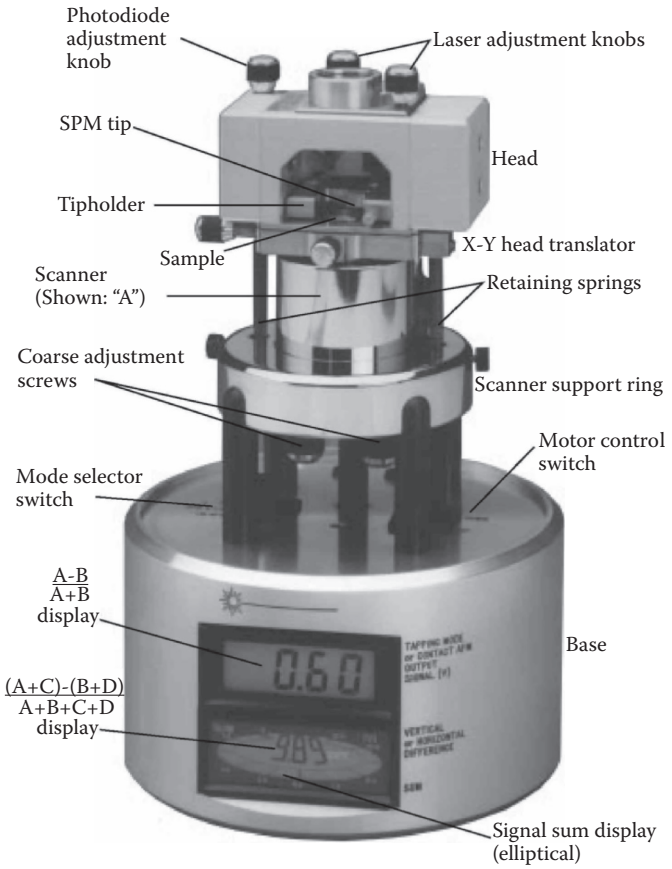


FIGURE 6.34 Veeco Instruments atomic force microscope.

stylus-based instruments such as record players and stylus profilometers (see Figure 6.40). The AFM not only measures the force on the sample but also regulates it, allowing acquisition of images at very low forces. One point of interest: the faster the feedback loop can correct deviations of the cantilever deflection, the faster the AFM can acquire images; therefore, a well-constructed feedback loop is essential to microscope performance. AFM feedback loops tend to have a bandwidth of about 10 kHz, resulting in image acquisition times of about 1 min.

In Figure 6.34 we show a photograph of a typical AFM.

Tools to Measure Materials Properties: Property Metrology

Introduction

While most MEMS metrology techniques focus on dimensional metrology and some on dynamic metrology, few systems exist for measuring other properties of microsystems, such as electrical, magnetic, thermal, and mechanical characteristics. In the category of material property measuring equipment, we start with a review of scanning proximal probe microscopes (SPMs) for determination of material properties, such as magnetic, thermal, and mechanical behavior. We then consider tools used to measure conductivity and carrier type of a material. MEMS devices are often mechanically dynamic by nature. Therefore, there is also a need to measure dynamic motion of MEMS devices. We cover two mechanical property metrology techniques: using interferometry with a stroboscopic LED to freeze device motion and laser Doppler vibrometry to measure contactless velocity and displacement of vibrating structures.

Material Property Measuring Scanning Proximal Probe Microscopes

The scanning proximal microscopes introduced above also can be utilized for a variety of material property metrology tasks (see Table 6.5). We briefly describe a number of the best-known techniques.

Magnetic force microscopy (MFM) is an imaging mode derived from atomic force microscopy (AFM) used in the noncontact mode that maps magnetic force gradients above a sample surface with a resolution of 10–25 nm. An image acquired with a MFM probe tip coated with a magnetized material (e.g., CoCr or NiFe) contains information about both the topography and the magnetic properties of a surface. Which effect dominates depends upon the distance of the tip from the surface, because the magnetic force persists for greater tip-to-sample separations than the van der Waals force. If the tip is close to the surface, in the region where standard noncontact AFM is operated, the image will be predominantly topographic. As the separation between tip and the sample increases, magnetic effects become apparent. While scanning in the latter mode, it is the magnetic field's dependence on tip-sample separation that induces changes in the cantilever's amplitude, resonance frequency, or phase. MFM can be used to image both naturally occurring and deliberately written magnetic domains. A MFM typically

measures magnetic properties and topography of a surface in a two-pass fashion. The first pass is a standard AFM trace that maps out the surface topography by gently tapping the tip along the surface. A second pass then samples the magnetic stray field by scanning at a constant height above the surface (see Figure 6.35).

Indentation (hardness) testing is very common for bulk materials where the direct relationship between bulk hardness and yield strength is well known. It can be measured by pressing a hard, specially shaped point into the surface and observing indentation. This type of measurement is of little use for measuring thin films below 5×10^4 Å. Consequently, very little is known about the hardness of thin films. Nanoindenting and scratching are new AFM-based methods to characterize material mechanical properties on a very small scale. Features less than 100 nm across, as well as thin films less than 5 nm thick, can be evaluated. Test methods include indentation for comparative and quantitative hardness determination and scratching for evaluation of wear resistance and thin-film adhesion. During nanoindentation, a hard tip of known geometry makes an impression in a sample, with a penetration depth and force dynamically measured at resolutions of nanometers and micronewtons, respectively. The resulting force-depth curve can be analyzed to give bulk properties such as modulus and hardness, as well as to study deformation mechanisms. Empirical relations have correlated hardness with Young's modulus and with uniaxial strength of thin films. Hardness calculations must include both plastic and long-distance elastic deformation. If the indentation is deeper than 10% of the film, corrections for elastic hardness contribution of the substrate must also be included.[12]

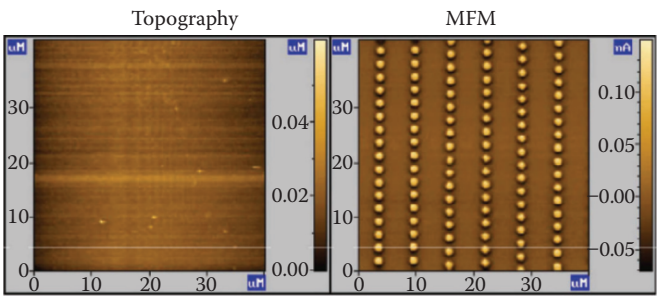


FIGURE 6.35 Magnetic force microscopy of a magneto-optical disc. (Veeco Instruments.)

Since in this new specialized instrument, load and displacement data are collected while the indentation is being introduced to the thin film; this eliminates the errors associated with later measurement of indentation size and provides continuous monitoring of load/displacement data similar to a standard tensile test. Load resolution may be 0.25 μN, displacement resolution 0.2–0.4 nm, and *x-y* sample position accuracy 0.5 μm. Diamond and sapphire are the primary materials of nanoindenting tips, but other hard materials also can be used, such as quartz, silicon, tungsten, steel, tungsten carbide, and almost any other hard metal or ceramic.

Bushan provides an excellent introduction to this field in the *Handbook of Micro/Nanotribology*.[13] In Figure 6.36 an indentation and scratch on polyurethane film is shown (17-μm scan).

In electrostatic force microscopy (EFM) a grounded tip first acquires the surface topography using the tapping mode. A voltage between the tip and the sample is applied in the second scan to collect electrostatic data (50–100 nm above the surface). EFM measures electric field gradient and distribution above the sample surface. EFM is used to monitor continuity and electric field patterns on samples such as semiconductor devices and composite conductors, as well as for basic research on electric fields on the microscopic scale. For thermal microscopy, there are quite a few options, as listed in Table 6.7, but scanning thermal microscopy (SThM) provides by far the best spatial resolution.

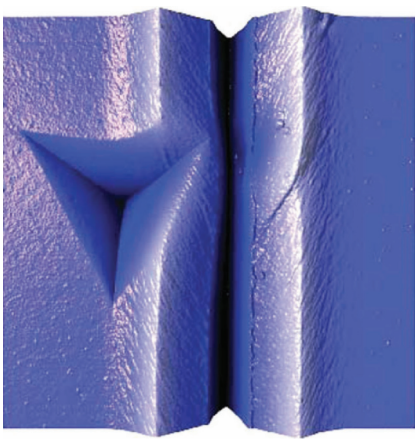


FIGURE 6.36 Indentation and scratch on polyurethane film (17-μm scan). (Asylum Research, http://www.asylumresearch.com/Products/NanoI/NanoIndenter.pdf.)

Table 6.7 Thermal Microscopy for Micro-Nano Devices

Techniques	Spatial Resolution
Infrared thermometry	1–10 μm*
Laser surface reflectance	1 μm*
Raman spectroscopy	1 μm*
Liquid crystals	1 μm*
Near-field optical thermometry	<1 μm
Scanning thermal microscopy	<100 nm

* Diffraction limit for far-field optics.

When an ultraminiaturized resistance is used as the scanning probe in an AFM, the image contrast reflects variations in thermal properties across the sample. The probe can be used as a thermometer or as a localized DC or AC heat source. When used as a thermometer, the image contrast represents variations in temperature across the surface. Hot spots, and also defects causing hot spots, in integrated circuits (ICs) and semiconductor devices may be detected this way. When used as a highly localized DC heat source, thermal conductivity/diffusivity variations across the surface can be mapped. In the AC heat source mode, thermal waves can be generated. Because of the evanescent nature of these waves, one can control the maximum depth of the sample that is imaged, according to the depth of penetration of the wave, which varies with the frequency chosen. SThM therefore looks below the surface in a controlled manner.

Conductivity and Carrier Type of a Material

In this section we cover how to go about measuring the conductivity and carrier type of thin films, such as thin-film resistors. In ICs and MEMS/NEMS devices, resistors are either thin-film metals or semiconductors. Metal resistors are deposited by evaporation, sputtering, or plating, whereas doping a semiconductor substrate creates semiconductor resistors. Metal resistors are patterned by wet etch, dry etch, or lift off and come in typical thicknesses of 0.1–10 μm and widths between 1 and 100 μm. Typical resistivity values range from 3 μΩcm for Al to 100 μΩcm for Nichrome. Metal films have nearly linear resistance-temperature behavior over wide temperature ranges, with a typical resistivity coefficient of 10^{-3}/°C. Issues that arise with metal resistors are morphological variations such as grain size changes and surface passivation. Semiconductor resistors, created by doping a semiconductor substrate, have a typical resistivity from 10^{-3} to 10^{7} Ωcm. The resistance-temperature behavior of semiconductor is much more dependent on temperature and impurities than that of metals and feature a typical resistivity coefficient per °C of −0.05. The biggest issues with semiconductor resistors are dopant concentration uniformity and deleterious effects of high-temperature processing.

In Volume I, Chapter 3, we saw that the Hall effect enables the measurement of the free-carrier type and concentration in semiconductors. Here we review two simpler techniques to determine these parameters, i.e., hot-point probe and four-point probe.

In Figure 6.37 we consider a very thin conductor film, of length l, width W, and thickness t, so that the current path area A = tW. The resistance, R, of this thin conductive layer is proportional to the resistivity, ρ, and inversely proportional to the thickness, t. It is convenient to define a sheet resistance, R_s, that is equal to ρ/t. Sheet resistance may be thought of as a material property for conductors that are essentially two-dimensional. For the resistance of a thin deposited conductor layer of length L, thickness t, and width W, we can write the following equation:

$$R = \frac{\rho}{t}\frac{L}{W} = R_s\frac{L}{W} \tag{6.19}$$

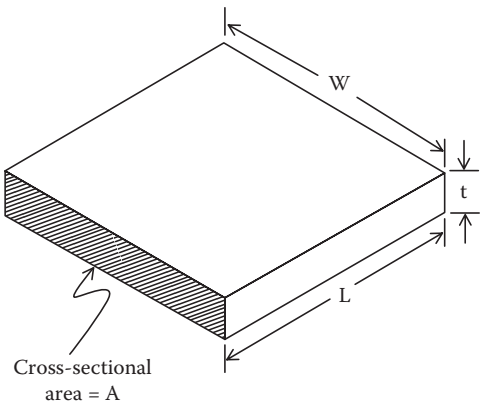


FIGURE 6.37 Sample geometry for a conductor thin film.

The ratio L/W is referred to as the number of squares (∼) and is dimensionless. The sheet resistance has the units of ohms, but it is convenient to refer to it as ohms per square (Ω/∼). The resistance of a rectangular thin layer is therefore the sheet resistance times the number of squares (L/W). The thickness and the number of squares, which can be accurately controlled with lithography, primarily determine the sheet resistance of a thin film.

The latter can be accurately controlled with lithography, whereas the sheet resistance primarily is determined by the film thickness carrier type and conductivity. Sheet resistance of a newly deposited thin film is typically measured with a four-point probe, an important tool in any IC or MEMS laboratory. For measuring resistance of a thin film, one could just measure the current that flows for a given applied voltage between two electrodes contacting the sample. However, in that case, contact resistances associated with the probes and current spreading problems around the probes can be significant and are not easily accounted for. Using four metal probes, as shown in Figure 6.38, allows one to force the current through the two outer probes using a constant current source and measure the voltage between the two inner probes. There still might be contact resistance and current spreading problems around the two outer probes, but since the voltage drop is measured between the two inner probes, problems with probe contacts are eliminated since no current flows through these inner contacts. The measured resistance (V/I) is converted to resistivity using the following relationship:

$$\rho = \frac{V}{I}.2\pi SF \tag{6.20}$$

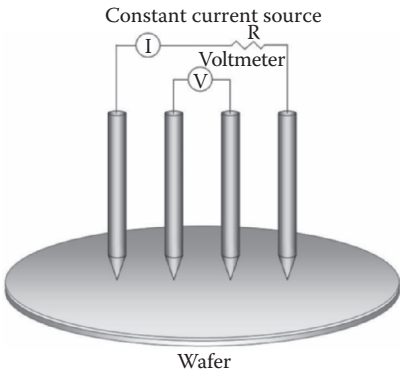


FIGURE 6.38 Four-point probe measurement arrangement.

with S the distance in centimeters between the metal probe needles and F a correction term. In the case the thickness, t, of the wafer or a measured film is much larger than S ($t >>>> S$), the correction factor F in Equation 6.18 equals 1. For thin films with $S >>> t$, Equation 6.20 must be replaced by the following:

$$\rho = \frac{\pi t}{\ln 2}\frac{V}{I} \tag{6.21}$$

and since $R_s = \rho/t$, we can calculate the sheet resistance as follows:

$$R_s = \frac{\pi}{\ln 2}\frac{V}{I} = 4.53\frac{V}{I} \tag{6.22}$$

The latter expression is most handy and easy to memorize.

The technique is more powerful than the sheet resistance method described above because it can determine the material type, carrier concentration, and carrier mobility separately. To determine whether a wafer is n- or p-type, a hot-point probe is a simple and reliable tool. The basic operation of this probe is illustrated in Figure 6.39. Two metal probes make ohmic contact with the semiconductor surface, and one probe is heated 25–100°C higher than the other. A voltmeter measures a potential difference across the probes, and the voltage polarity indicates whether the material is n- or p-type. In the case of an n-type sample, the majority carriers are electrons. At the hot probe, the thermal energy of the electrons is higher than at the cold probe, so the electrons will tend to diffuse away from the hot probe driven by the temperature gradient. As the electrons diffuse away from the hot probe, they leave behind a positively charged zone at the hot probe and the negatively charged mobile electrons tend to build up near the cold probe. This results in the hot probe becoming positive with respect to the cold probe. By a similar set of arguments, if the material were p-type, positively charged holes would be the majority carriers, and the polarity of the induced voltage would be reversed.

Dynamic Mechanical Properties

Interferometry with a stroboscopic LED and laser Doppler vibrometry can both be used to measure

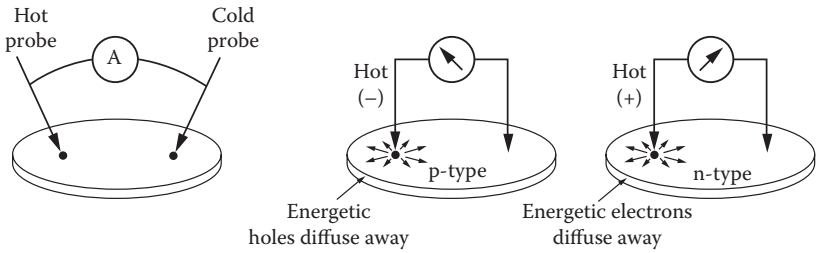


FIGURE 6.39 Using a hot-point probe to determine whether a semiconductor is n- or p-type.

contactless velocity and displacement of vibrating structures. A stroboscopic LED freezes device motion and enables an interferometer, as shown in Figure 6.11, to measure in-plane and out-of-plane motion at frequencies up to 10 MHz (this addition to an interferometer is called a *dynamic MEMS capability*).

A laser Doppler vibrometer is an interferometry system for accurately measuring velocity and displacement of vibrating structures again without contact. The system is based on the principle of the detection of the Doppler shift of coherent laser light scattered from a small area of the moving test structure. This system detects the Doppler shift of a coherent laser and then converts the frequency shift to velocity. As the laser light has a very high frequency, a direct demodulation of the light is impossible; therefore, an optical interferometer is used to mix the scattered light coherently with a reference beam. Such an arrangement can be based on a Michelson interferometer, as shown in Figure 6.10.

Structure Metrology

Introduction

Structural metrology measures crystallinity, defects, surface roughness, etc. In the structural metrology category, we review stylus profilometry for surface roughness and transmission electron microscopy (TEM) for crystallinity.

Stylus Profilometry

Surface texture of a sample is, in general, broken up into three components: roughness, waviness, and form (see also Volume II, Chapter 6, on mechanical energy-based removing). Roughness refers to the marks made on a sample surface by the machining tool itself, e.g., grooves from the tool or from each grinding granule on a grinding wheel. Waviness is the result of the distance between the tool and the workpiece changing during machining. This may be caused by vibration in the tool caused by sources outside the machining tool (e.g., a truck passing by) or by sources inside the tool (e.g., a worn spindle). Form errors are yet larger dimensional errors because the machine tool path is not always straight or it may be worn. This creates surface irregularities, but in a consistent manner, because the machine follow the same wrong path. All three surface finish components exist simultaneously and overlap one another. One often will want to look at each (roughness, waviness, and form) separately, so we make the assumption (a correct one, in most cases) that roughness has a shorter wavelength than waviness, which in turn has a shorter wavelength than form.

A stylus profilometer uses a hard tip (e.g., diamond) dragged over a surface with very light pressure and simultaneously measures form, dimension, texture, and step height. The first such devices for surface roughness measurement came about around 1930. The movement of the tip actuates a linear variable differential transducer (LVDT) that converts movement in an electrical signal. The tip of the profilometer is conical and has a finite rounded shape, which interacts with the sample being scanned. It may measure film thickness (step height) changes of 200 Å to 65 µm with a vertical resolution of about 1 nm. The vertical sensitivity is in the nanometer range, but steep edge profiles are distorted because of the shape of the tip, as indicated in Figure 6.40. The horizontal resolution depends on tip radius, so small-radius tips are better but are more likely to be damaged by mishandling.

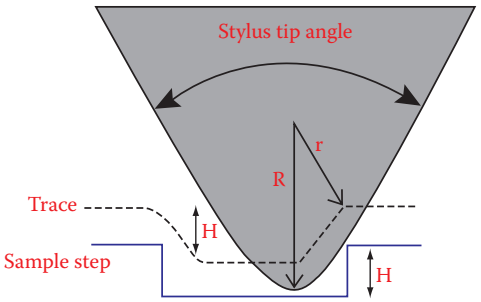


FIGURE 6.40 The stylus of a stylus profilometer.

As we saw earlier, an atomic force microscope (AFM) is a new instrument for measuring roughness of a sample surface. As for its design, an AFM has similarities to a conventional stylus profilometer; however, an AFM can image a sample surface and precisely measure up to nanometer size in three dimensions. Because of the sharper tip and small loading force, the lateral resolution in AFM is greatly improved in comparison with the conventional profilometer. In Table 6.8 we compare the major types of profilometers.

Transmission Electron Microscopy

Ernst Ruska and Max Knoll built the first transmission electron microscope (TEM) in 1931. As in the case of the first optical microscope, this was a very crude machine compared with its modern descendants. It did, however, put some important ideas about how to image with electrons to the test. For example, instead of using glass lenses to bend and focus light, electron microscopes use magnetic coils to do the same for electrons (H. Busch was the first to use a magnetic coil like a lens in 1926). Ruska used his first electron microscope to take pictures of gold and copper surfaces and produced images at about ×10 (about the same magnification as the 1595 Janssen microscope). The development of the electron microscope was much faster, though, than that of the optical microscope, and it was not long before electron microscopes far surpassed the resolution of optical microscopes (Ruska constructed a microscope that surpassed the 200-nm optical limit in 1933). Following the development

TABLE 6.8 Profilometers Compared

	White Light Interferometry and Phase Shift Technology	Stylus Contact	Atomic Force Microscope
Advantages	Excellent depth resolution, 0.1 nm; very good lateral resolution down, to 0.3 μm	Lowest-price technique for standard systems	Best lateral resolutions for detail imaging
	Use charge-coupled device camera to take roughness measurement in a few seconds	Depending on design can test steep angular surfaces	Can reveal surface structure in details
	Stitching technique allows coverage of large surfaces		
	Ideal for many microelectronic applications		
Disadvantages	Difficulty on high-angular surfaces or surface where line of interference pattern is difficult to see; chromatic aberration technique works better on these surfaces	Tip size often larger than 10 μm but can be down to 1 μm in higher-priced systems	Slow compared with white light systems
	More limited than white light chromatic aberration in the *z*-direction; stitching allows a few millimeters	Depending on size of probe and load applied, deformation may occur on softer materials, such as polymers or plastics	Requires more knowledge for use
		Normally two-dimensional instruments	Relatively small surfaces can be tested (100 μm by 100 μm)
		Slow, especially if three-dimensional imaging is performed	

Source: http://www.microphotonics.com/compare.htm.

of the first commercial available electron microscope by Ruska and the company Siemens in 1939, the electron microscope rapidly became an invaluable tool to science and engineering. Much of what we know about the nanoworld we owe indeed to the electron microscope. Although the technique of x-ray diffraction preceded the invention of the electron microscope and provided compelling evidence for atomic crystal structure, atoms themselves had never been seen until high-resolution TEMs came about.

A TEM maybe compared to a slide projector. In a slide projector, a beam of light passes through (transmits) a slide, and as the light passes through, the structures and objects on the slide modulate it. This transmitted beam is then projected onto the viewing screen, forming an enlarged image of the slide. TEMs work much the same way except that they shine a beam of electrons through a very thin specimen, and whatever electrons are transmitted are projected onto a phosphor screen or recorded on film or by a digital camera for the user to see. TEMs are used to characterize the microstructure of materials with very high spatial resolution. The electron gun, condenser lenses, and vacuum system in a TEM are similar to the ones in an SEM (see above), but in a TEM, a higher acceleration voltage of 120–200 kV is used (the operating voltage in an SEM ranges from 30 to 60 kV). In a scanning transmission electron microscope (STEM), the beam is rastered across the sample to form the image.

With a TEM, information about the morphology, crystal structure, defects, crystal phases, and composition is obtained by a combination of electron-optical imaging, electron diffraction, and small probe (20 Å) capabilities. The trade-off for this diverse range of structural information and high resolution is the challenge of producing very thin samples for electron transmission (Figure 6.41). A TEM sample must be 1000 Å or less in thickness in the area of interest. Sample preparation for TEM is more tedious and time consuming than for almost any other analytical technique.

There are three basic modes of TEM operation: the bright-field imaging mode, the dark-field imaging mode, and the diffraction mode. The way to choose which electrons form the image is by

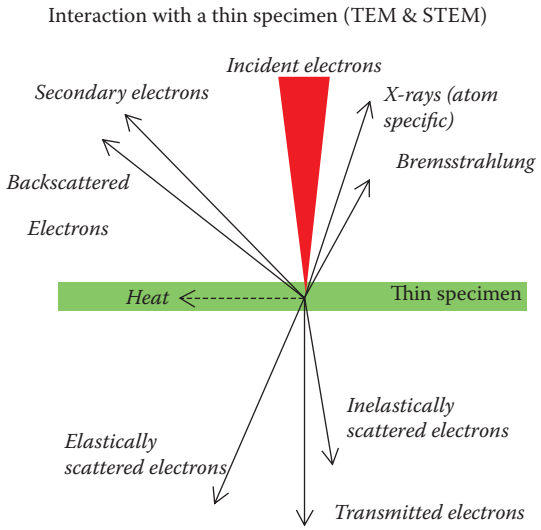


FIGURE 6.41 Electron path through a thin specimen (<1,000 Å) in transmission electron microscopy and scanning transmission electron microscopy.

inserting an aperture into the back focal plane of the objective lens.

In the bright-field (BF) mode of the TEM, an aperture is placed in the back focal plane of the objective lens, which allows only the direct beam to pass (Figure 6.42a). In this case, the image results from a weakening of the direct beam by its interaction with the sample. Therefore, mass-thickness and diffraction contrast contribute to image formation: thick areas, areas in which heavy atoms are enriched, and crystalline areas appear with dark contrast.

In dark-field (DF) images, the direct beam is blocked by the aperture while one or more diffracted beams are allowed to pass the objective aperture. Since diffracted beams have strongly interacted with the specimen, very useful information is present in DF images, e.g., about planar defects, stacking faults, or particle size (Figure 6.42b).

We can also think of the TEM as a diffraction camera. To obtain lattice images, a larger objective aperture is selected to allow many beams, including the direct beam, to pass (Figure 6.42c). The image is formed by the interference of the diffracted beams with the direct beam (phase contrast). If the point resolution of the microscope is sufficiently high and a suitable sample is oriented along a zone axis, then high-resolution TEM (HRTEM) images are obtained. In many cases, the atomic structure of a specimen can directly be investigated by HRTEM.

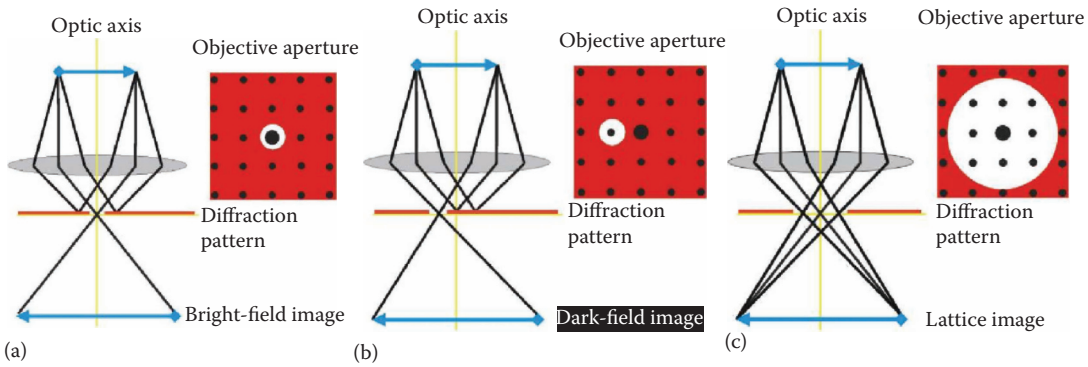


FIGURE 6.42 Bright-field image (a), dark-field image (b), and diffraction (c) in scanning transmission electron microscopy.

The objective lens forms a diffraction pattern in the back focal plane with electrons scattered by the sample and combines them to generate an image in the image plane (first intermediate image) as illustrated in Figure 6.43. Thus, diffraction pattern and image are simultaneously present in the TEM. In imaging mode (left panel in Figure 6.43), an objective aperture can be inserted in the back focal plane to select one or more beams that contribute to the final image (BF, DF, and HRTEM). In selected area electron diffraction (SAED), an aperture in the plane of the first intermediate image defines the region for which the diffraction is obtained (Figure 6.43, right panel). Which of them appears in the plane of the second intermediate image and is magnified by the projective lens on the viewing screen depends on the intermediate lens. Switching from real space (image) to reciprocal space (diffraction pattern) is easily achieved by changing the strength of the intermediate lens. By manipulating the magnetic lenses of the microscope, the diffraction pattern may thus be observed by projecting it onto the screen instead of the image (Figure 6.43).

An example of what a diffraction pattern obtained in this way may look like is shown in Figure 6.44. The center spot in this figure corresponds to the transmitted electron beam, and the other spots are diffracted portions of the initial electron beam. The location of the spots is governed by Bragg's law ($n\lambda = 2d \sin\theta$; see Volume I, Figure 2.17 and Equation 2.20).

Transmission and scanning transmission electron microscopy can produce images of atomic lattices with 2.3-Å point-to-point resolution. In Table 6.9, we compare scanning electron microscopy (SEM) and transmission electron microscopy (TEM) with optical microscopy (OM) and scanning proximal microscopy (SPM) (using both STMs and AFMs).

Compositional Metrology

Introduction

As examples of chemical composition metrology we discuss energy dispersive analysis of x-rays (EDAX) [also known as *energy dispersive x-ray spectroscopy* (EDX) or EDS], a standard procedure for identifying and quantifying elemental composition of specimen surfaces to shallow depths and Auger electron spectroscopy (AES) and scanning Auger

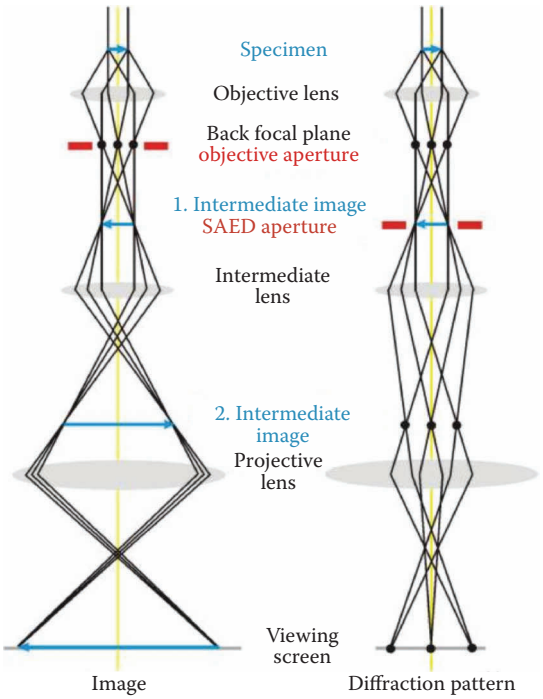


FIGURE 6.43 Switching from imaging to diffraction in scanning transmission electron microscopy. SAED = selected area electron diffraction.

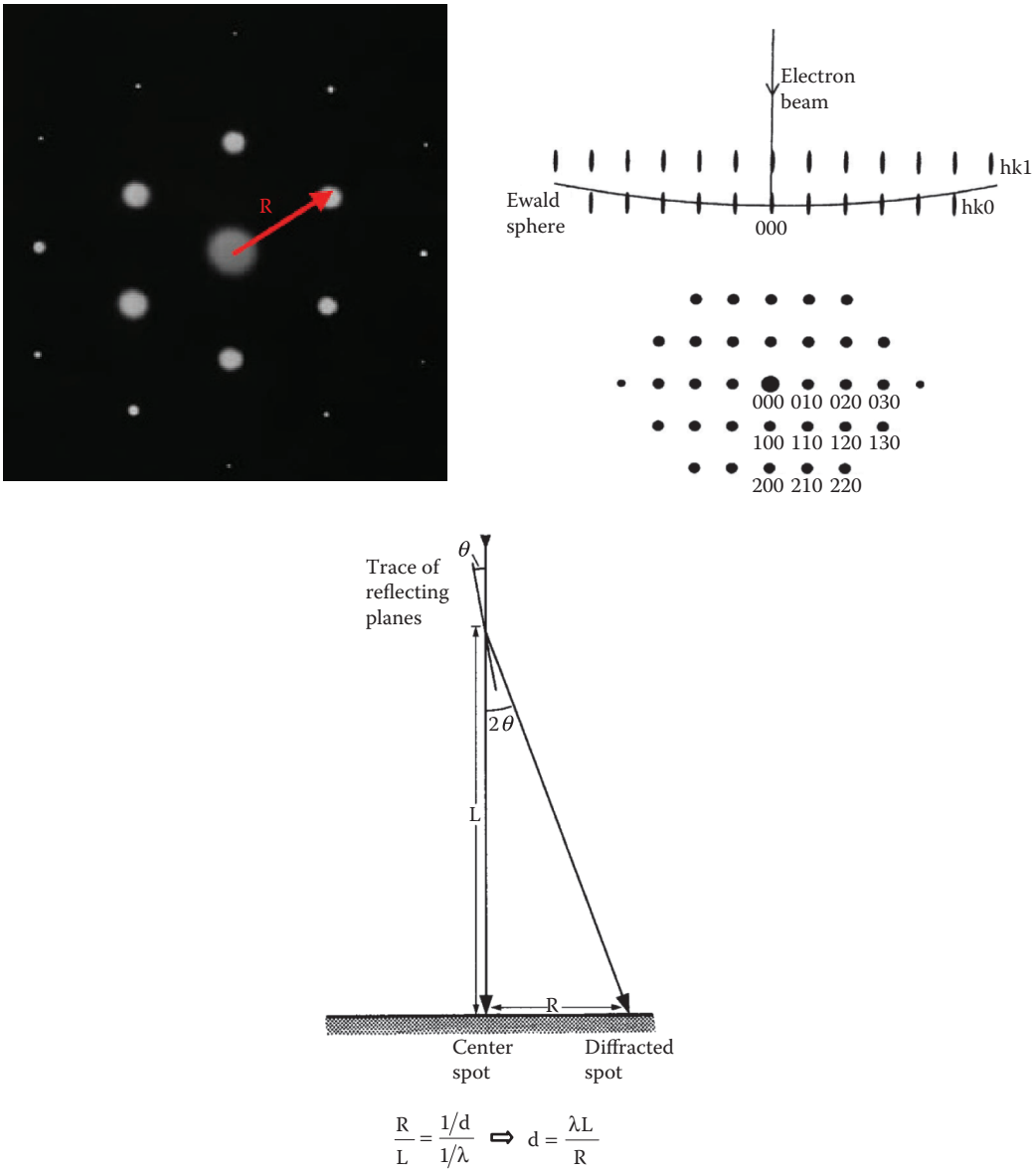


FIGURE 6.44 An example of what a diffraction pattern in a scanning transmission electron microscope may look like. The center spot corresponds to the transmitted electron beam, and the other spots are diffracted portions of the initial electron beam. The locations of the spots are governed by Bragg's law.

TABLE 6.9 The Use of SEMs and TEMs Compared with OM and Scanning Proximal Microscopy (Using Both STMs and AFMs)

	OM	SEM/TEM	Scanning Proximal Probe Microscopy
Environment	Air, liquid	Vacuum	Air, liquid, UHV
Depth of field	Small	Large	Medium
Lateral resolution	1 μm	1–5 nm: SEM 0.1 nm: TEM	2–10 nm: AFM 0.1 nm: STM
Vertical resolution	N/A	N/A	1.1 nm: AFM 0.2 0.01 nm: STM
Magnification	Up to 2,000	Up to 1,000,000	Up to 100,000,000
Sample	Not completely transparent	Unchargeable, vacuum compatible; thin film: TEM	Surface height <10 mm
Contrast	Absorption/reflection	Scattering/diffraction	Tunneling

microscopy (SAM), which can identify every atom from Li on up.

EDAX (EDX or EDS)

Energy dispersive analysis of x-rays (EDAX) is also called *EDX* and, more and more frequently, *EDS*. This compositional metrology technique is carried out in conjunction with scanning electron microscopy (SEM) and transmission electron microscopy (TEM). A modern trend in electron microscopy is indeed to fit x-ray analysis equipment as a bolt-on accessory to electron microscopes. Bombarding a specimen with electrons causes x-rays of characteristic wavelengths and energies to be emitted from the spot where the beam strikes the sample, and an EDS detector analyzes the x-ray photons by energy rather than wavelength. The EDS detector (sometimes called *EDX detector*) converts the energy of each individual x-ray into a voltage signal of proportional size. The detector involves a large single-crystal Si that either has been treated to approximate an ideal semiconductor or is of high-enough purity to truly be an intrinsic semiconductor. Front and back contacts on the Si piece are kept at several kilovolts potential relative to each other, and x-rays are converted into a charge by ionization of atoms in the crystal. To keep electronic noise at a minimum, the Si detector is cooled in liquid nitrogen. X-rays that pass through the front contact tend to dissipate their energy, creating electron-hole pairs in the intrinsic region. Next, this charge is converted to a voltage signal by a field effect transistor (FET) preamplifier, and the voltage signal is input into a pulse processor for measurement. Since each electron-hole pair has a characteristic creation energy, the total number of charge carriers created is proportional to the energy of the incident x-ray. A computer keeps track of the number of counts within each energy range, and the total collected x-ray spectrum can then be determined. Computer analysis of energy spectra makes it possible to measure accurately the nature and quantity of different elements in the material. This can then be used to make a chemical map of the surface. The technique is of little use to biologists because light elements such as carbon produce too weak an x-ray signal. However, it is of great value in materials science, particularly because an area as small as 1 μm^2 can be analyzed with precision (http://www.edax.com).

Typically, x-ray analysis is a bulk sensitive technique because of the small cross-section [10^{-6} $Å^2$ compared with 1 $Å^2$ for LEED (see above)], but grazing incidence (<1°) makes it surface sensitive (total reflection, because the refractive index of x-rays is slightly smaller than unity). The refracted wave becomes an evanescent wave traveling along the surface within a few 10 Å. This is called *gracing incidence x-ray diffraction* (GIXRD).

Auger Electron Spectroscopy and Scanning Auger Microscopy

Upon electron bombardment of metals, Pierre Auger, in 1925, observed (at first in a cloud chamber, and then on photographic plates) the occurrence of electron emissions with precisely determined energies. These electrons were later named Auger electrons and may serve to identify their parent atoms. It was J. J. Lander, who in 1953 suggested the idea of using Auger electron spectroscopy (AES) for surface analysis. AES was implemented as an analytic tool in 1967 (by Larry Harris), after increasing the sensitivity of the method by using differential spectra to discriminate the small Auger peaks in the electronic spectra. AES is now the most commonly used method by which to investigate surface composition.

The Auger process is a three-electron concerted process, as illustrated in Figure 6.45. An incident electron beam (2–10 keV) hits an atom and ejects a core electron, leaving a hole in a highly excited atom. The ionized atom relaxes by emitting either an x-ray photon (x-ray fluorescence) or by ejecting another electron (Auger electron). For lighter elements Auger emission is favored over x-ray fluorescence. In Auger spectroscopy the state of excitation is quenched by another electron falling into the hole left by the ejected core electron. The binding energy of the first electron is always very much greater than that of the second, so that there is still considerable excess energy in the system. This extra energy is shed by the ejection of a third electron from another shell. The residual kinetic energy of the ejected electron is measured and is a unique quantity determined by the binding energies of the three electrons and the work function of the materials involved.

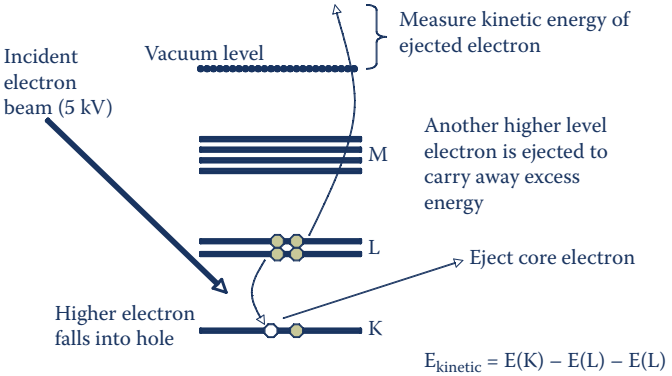


FIGURE 6.45 The Auger process.

An Auger process is identified by the three letters of the atomic shells that participate in the process. The transition in Figure 6.45, for example, is a KLL Auger transition.

In Figure 6.18 above, we showed the relative intensity versus abundance of the various emission phenomena from a sample under e-beam radiation in an SEM and a TEM. Some of the electrons, in the energy region between backscattered and secondary electrons, arise from the Auger process occurring in atoms near the surface. Many Auger electrons will fully lose their energy via inelastic collisions in the solid. Therefore, the analyzer will collect only those Auger electrons originating from the near surface region. The number of emitted electrons N(E) versus energy is redrawn in Figure 6.46, with emphasis on the Auger electron peaks. Electrons emitted in the 100–1000 eV range (higher than secondary, lower than backscattered) are energy analyzed, and a spectrum is obtained that is unique to every atom. Often, hemispherical analyzers are used because of its better geometric properties. Hence, it gives very specific atomic identity information rather than the general small/large *z*-guidance provided by backscattered electrons (BSEs) in an SEM. Auger spectroscopy can identify every atom from Li on up (since three electrons are involved in the Auger process, H and He do not produce Auger electrons). Auger spectra for all elements are compiled in an Auger atlas.

Typically, the relatively small Auger signals N(E) are superimposed on a large background. Therefore the spectra are usually taken in the derivative mode, by applying a modulation voltage on the analyzer and detecting with a lock-in amplifier. This is demonstrated in the AES example of Figure 6.47.

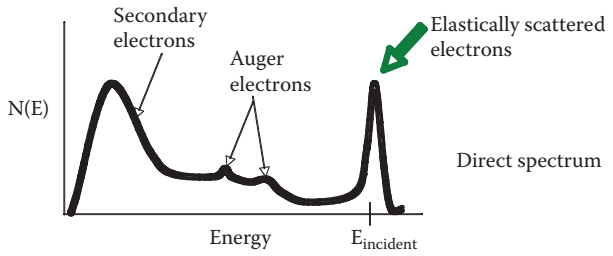


FIGURE 6.46 Auger peaks in the electron spectra of an scanning electron microscope or transmission electron microscope.

Since the electron beam can be can be focused to a spot of <1 nm, one can carry out SAM. In the excitation process, the backscattered electrons also find their way into neighboring regions of the sample and can excite electrons and induce an Auger process in regions of the sample that were not in the original excitation volume. The Auger signal then arises from an apparently larger region of the sample than the incident electron beam directly excites. This degrades the spatial resolution in the SAM process by a factor of 2–5 compared with the same SEM experiment. As a consequence, the penetration depth and scattering within the solid typically limits the SAM resolution to ~20 nm, but recent instruments claim a resolution of ~6 nm. By correlating a specific signal with the excitation beam's location, a chemical map of a surface can be produced, as illustrated in Figure 6.48. Although chemical shifts lead to changes in the Auger energies, AES is usually not used to get chemical binding information because of the three electrons involved.

Summary: Microscope Comparisons

In Figure 6.49 we compare the lateral resolutions of optical microscopes (OMs), scanning electron microscopes (SEMs), interference microscopes (IMs), and proximal probe techniques such as the use of scanning tunneling microscopes (STMs) and atomic force microscopes (AFMs).

OM is usually what one thinks of when studying microscopy. It is prevalent and widely available. Its resolution is limited by the diffraction of light, which is limited to about ½ of the wavelength of the light being diffracted through the optical lensing elements of the microscope. While this is several

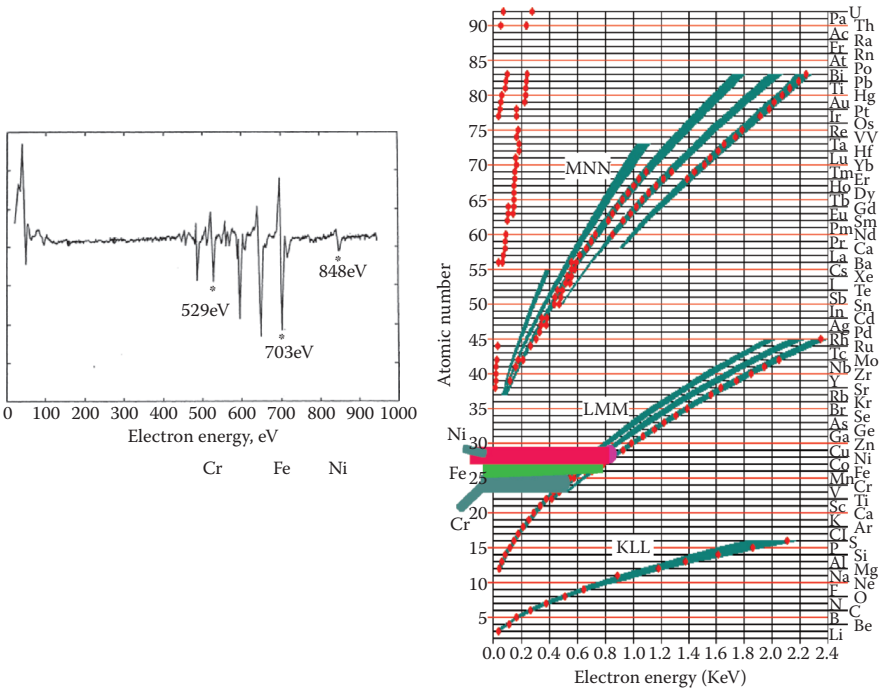


FIGURE 6.47 Example of an Auger electron spectroscopy (AES) spectrum. From the differential AES spectrum, Cr, Fe, and Ni have been identified.

hundred nanometers, in practice it is challenging to get good optical resolution down to the 1-μm range. Interference microscopy is a newer technique that uses optical wavelength light. Its lateral resolution is controlled exactly like that of optical microscopy, and it has the same limitations. As we will see below, its virtue comes from its vertical resolution and its wide dynamic range.

In an SEM, the resolution is limited by the diffraction of electron waves. Since these are in the angstrom range, SEMs with resolution of a few nanometers and a wide magnification range, from × 20 to × 650,000 are quite widely deployed. Recent reports have even described a new instrument able to observe atoms (a few angstroms), but these are not available for use yet. Probe microscopes such as STMs and AFMs have a resolution that is controlled by the sharpness of the probe tip. This can be in the range of a few nanometers, but the differing contrast mechanisms mean that an STM can easily achieve

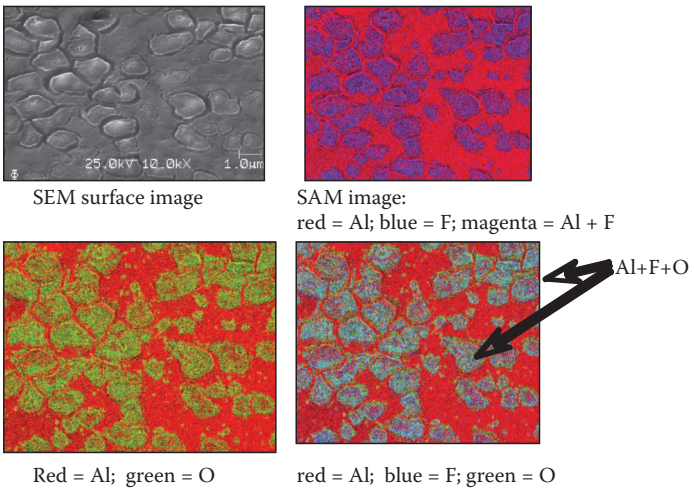


FIGURE 6.48 Scanning electron microscopy and Auger images of an aluminum oxide surface, in the absence and presence of fluorine contamination.

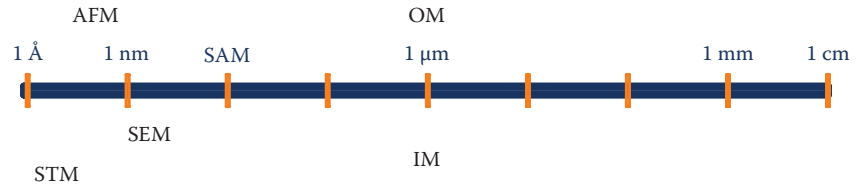


FIGURE 6.49 Lateral resolution of microscopies: optical microscopy (OM), scanning electron microscopy (SEM), scanning Auger microscopy (SAM), interference microscopy (IM), scanning tunneling microscopy (STM), and atomic force microscopy (AFM).

atomic resolution, whereas atomic resolution for an AFM is difficult to achieve.

AFM comes with lateral resolutions down to 30 Å and a vertical resolutions to 0.1 Å. STM has a lateral resolution of 1 Å and a vertical resolution of 0.1 Å. SAM is closely related to SEM, except that the scattering in the sample leads to a "smearing" of the incident beam and broadens the effective spot size. Its resolution general lags behind that of an SEM by a factor of 2–5.

Vertical resolutions for the types of microscopes listed above are quite different from their lateral resolutions, as is evident from Figure 6.50. OMs, SEMs, and SAMs do not give much vertical information at all. While it is rather straightforward to derive quantitative lateral data, these techniques do not provide for quantitation in the vertical direction. To counter this, it is possible to tilt samples and image their sides. However, in a specific image, the vertical information is very poor. Hence, these techniques have much better lateral resolution than vertical resolution. In contrast, probe microscopes and IMs have much better vertical resolution than lateral resolution. The IM, even though it is an optical technique, has a resolution limit approaching 1 nm. The AFM and the STM have vertical resolutions in the subatomic range of tens of picometers. The best instruments can resolve down to a few hundred femtometers.

We need to image large things as well as small. So another important point is, which instrument can provide the largest field of view? What is the range of features, largest to smallest, that can be observed? The resolution limits listed above do define the lower limits of the dynamic range of the microscopes reviewed, and the upper limit is controlled by the size of the field of view that can be imaged at once. OM is able to image from macroscopic features down to its resolution limit. The scanning electron techniques generally have a maximum field of view of a few millimeters. The proximal probe microscopes are quite limited in their maximum field of view. Also, the limitations are different for the lateral and vertical directions. Lateral limits range from a few micrometers to over 100 μm. By contrast, the range in the vertical direction is limited to around a few micrometers. These limits often are determining factors in assessing the applicability of a probe microscope technique for a particular application. IM has a very wide dynamic range. It is able to image features varying in vertical displacement of over a millimeter. The lateral range is set by its maximum field of view, which is around 5 mm. The dynamic ranges of the various

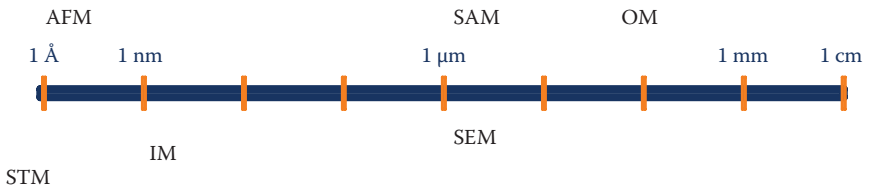


FIGURE 6.50 Vertical resolution of microscopies: optical microscopy (OM), scanning electron microscopy (SEM), scanning Auger microscopy (SAM), interference microscopy (IM), scanning tunneling microscopy (STM), and atomic force microscopy (AFM).

microscopes, considered here, are summarized in Figure 6.51.

MEMS Based Metrology: In Situ Stress-Measuring Techniques

Introduction

A stressed thin film will bend a thin substrate by a measurable degree. A tensile stress bends the surface and makes it concave, while a compressive stress renders the surface convex (see Volume II, Figure 7.58). The most common methods for measuring the stress in a thin film are based on this substrate bending principle. The deformation of a thin substrate due to stress is measured either by observing the displacement of the center of a circular disk or by using a thin cantilevered beam as a substrate and calculating the radius of curvature of the beam and hence the stress from the deflection of the free end. More sophisticated local stress measurements use analytical tools such as x-ray,[14] acoustics, Raman spectroscopy,[15] infrared spectroscopy,[16] and electron-diffraction techniques. Local stress does not necessarily mean the same as the stress measured by substrate bending techniques, since stress is defined microscopically, while deformations are induced mostly macroscopically. The relationship between macroscopic forces and displacements and internal differential deformation, therefore, must be modeled carefully. Local stress measurements may also be made using in situ surface micromachined structures such as strain gauges made directly out of the film to be tested.[17–22] The dimensions and structure of a thin suspended and pressurized micromachined membrane, as we learned above, may be measured by stylus profilometry, light microscopy, or interferometry. Intrinsic stress influences the frequency response of mobile microstructures (see Volume II, Equation 7.29), which can be measured by laser Doppler vibrometer or with an interferometer equipped with a stroboscope (see Figure 6.11: optical WYKO white-light profilometer with dynamic MEMS capability). Whereas residual stress can be determined from wafer curvature and dynamic deflection data, material structure of the film can be studied by x-ray diffraction and transmission electron microscopy (TEM). Krulevitch, among others, attempted to link the material structure of polysilicon and its residual strain.[23] Below, we will review stress-measuring techniques, starting with the more traditional ones and subsequently clarifying the problems and opportunities in stress measuring with in situ surface micromachined devices.

Disk Method: Biaxial Measurements of Mechanical Properties of Thin Films

For all practical purposes, only stresses in the x and y directions are of interest in determining overall thin-film stress, as a film under high stress can only expand or contract by bending the substrate and deforming it in a vertical direction. Vertical deformations will not induce stresses in a substrate because it freely moves in that direction. The latter condition enables us to obtain quite accurate stress values by measuring changes in bow or radius of curvature of a substrate. The residual stresses in thin films are large, and sensitive optical gauges may measure the associated substrate deflections.

The disc method most commonly used is based on a measurement of the deflection in the center of the disc substrate (say a Si wafer) before and after processing. Since any change in wafer shape is directly attributable to the stress in the deposited film, it

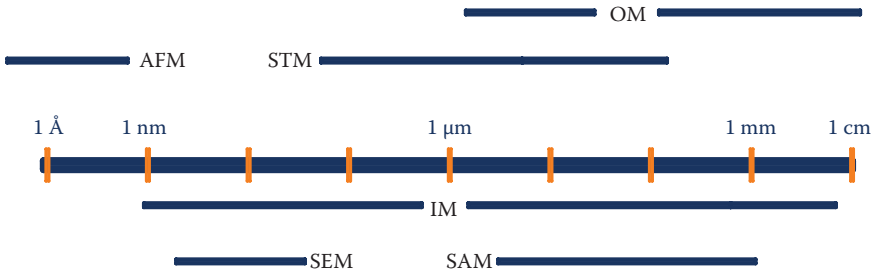


FIGURE 6.51 Dynamic range of microscopies: optical microscopy (OM), scanning electron microscopy (SEM), scanning Auger microscopy (SAM), interference microscopy (IM), scanning tunneling microscopy (STM), and atomic force microscopy (AFM).

is relatively straightforward to calculate stress by measuring these changes. Stress in films using this method is found through the Stoney equation,[24] relating film stress to substrate curvature, as follows:

$$\sigma = \frac{1}{R}\frac{E}{6(1-\nu)}\frac{T^2}{t} \qquad (6.23)$$

where R = measured radius of curvature of the bent substrate
E/(1 – ν) = biaxial modulus of the substrate
T = thickness of the substrate
t = thickness of the applied film[25]

The underlying assumptions include the following:

- The disc substrate is thin and has transversely isotropic elastic properties with respect to the film normal.
- The applied film thickness is much less than the substrate thickness.
- The film thickness is uniform.
- The temperature of the disk substrate/film system is uniform.
- The disc substrate/film system is mechanically free.
- The disc substrate without film has no bow.
- Stress is equibiaxial and homogeneous over the entire substrate.
- Film stress is constant through the film thickness.

For most films on Si, we assume that t ≤ T; for example, t/T measures ~10^{-3} for thin films on Si. The acceptability of the uniform thickness, homogeneous, and equibiaxial stress assumptions depend on the deposition process. Chemical vapor deposition (CVD) is a widely used process, as it produces relatively uniform films; however, sputter-deposited films can vary considerably over the substrate. In regard to the assumption of stress uniformity with film thickness, residual stress can vary considerably through the thickness of the film. Equation 6.23 gives only an average film stress in such cases. In cases where thin films are deposited onto anisotropic single-crystal substrates, the underlying assumption of a substrate with transversely isotropic elastic properties with respect to the film normal is not completely justified. Using single-crystal silicon substrates possessing moderately anisotropic properties such as <100> or <111> oriented wafers (Volume I, Equation 4.87) satisfies the transverse isotropy argument. Any curvature inherent in the substrate must be measured before film deposition and algebraically added to the final measured radius of curvature. To give an idea of the degree of curvature, 1 μm of thermal oxide may cause a 30-μm warp of a 4-inch silicon wafer, corresponding to a radius of curvature of 41.7 m.

The following companies offer practical disc method-based instruments to measure stress on wafers: ADE Corp. (Newton, MA; acquired by KLA-Tencor in 2006), Corning Tropel (Fairport, NY; http://www.corning.com/metrology), Ionic Systems (San Jose, CA; http://www.ionic.com), and KLA Tencor Instruments (Mountain View, CA; http://www.kla-tencor.com).[25] Figure 6.52a illustrates the sample output from KLA Tencor's optical stress analysis system. Figure 6.52b represents the measuring principle of Ionic Systems' optical stress analyzer. None of the above techniques satisfies the need for measuring stress in low modulus materials such as polyimide. For the latter applications, the suspended membrane approach (see below) is more suited.

Dog-Bone Method: Uniaxial Measurements of Mechanical Properties of Thin Films

Many problems associated with handling thin films in stress test equipment may be bypassed by applying micromachining techniques. One simple example of problems encountered with thin films is the measurement of uniaxial tension to establish the Young's modulus. This method, effective for macroscopic samples, proves problematic for small samples. The test formula is illustrated in Figure 6.53. The gauge length, L, of the bar-shaped structure in Figure 6.53a represents the region we allow to elongate, and the area, A (= $W \times H$), is the cross-section of the specimen. A stress, F/A, is applied and measured with a load cell; the strain, $\delta L/L$, is measured with an LVDT or another displacement transducer (a typical instrument used is the Instron 1123). The Young's modulus is then deduced from the following equation.

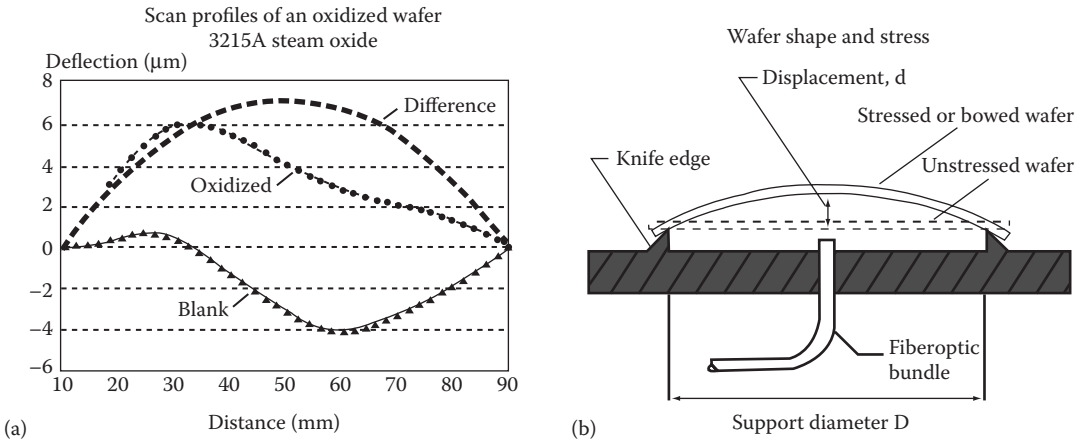


FIGURE 6.52 Curvature measurement for stress analysis. (a) Sample output from KLA Tencor's FLX 2908 stress analysis instrument, showing how stress is derived from changes in wafer curvature. (b) The reflected light technique, used by Ionic Systems to measure wafer curvature. (From Singer, P. 1992. Film stress and how to measure it. *Semicond Int* 15:54–58. With permission.[25])

$$E = \left(\frac{F}{A}\right)\left(\frac{L}{\delta L}\right) \tag{6.24}$$

The obvious problem, for small or large samples, is how to grip the sample without changing A. Under elongation, A will indeed contract by $(W + H)\delta\nu L$. In general, making a dog-bone shaped structure (also called an *Instron specimen*) solves that problem, as shown in Figure 6.53b. Still, the grips introduced in an Instron sample can produce end effects and uncertainties in determining L. Making Instron specimens in thin films is even more of a challenge, since the thin film needs to be removed from the surface, possibly changing the stress state, while the removal itself may modify the film.

As in the case of adhesion (see Volume II, Figure 7.56), some new techniques for testing stress in thin films, based on micromachining, are being explored. These microtechniques prove more advantageous than the whole wafer disc technique in that they are able to make local (in situ) measurements.

The fabrication of micro-Instron specimens of thin polyimide samples is illustrated in Figure 6.54A.[26] Polyimide is deposited on a p+ Si membrane in multiple coats. Each coat is prebaked at 130°C for 15 minutes. After reaching the desired thickness, the film is cured at 400°C in nitrogen for 1 h (Figure 6.54Aa). The polyimide is then covered with a 3,000-Å layer of evaporated aluminum (Figure 6.54Ab). The aluminum layer is patterned by wet etching (in phosphoric-acetic-nitric solution, referred to as the *PAN etch*) into the Instron specimen shape (Figure 6.54Ac). Dry etching transfers the pattern to the polyimide (Figure 6.54Ad). After removing the Al mask by wet etching, the p^+ support is removed by a wet isotropic etch (HNA) or an SF_6 plasma etch (Figure 6.54Ae), and finally the side silicon is removed along four pre-etched scribe lines, releasing the residual stress (Figure 6.54Af). The

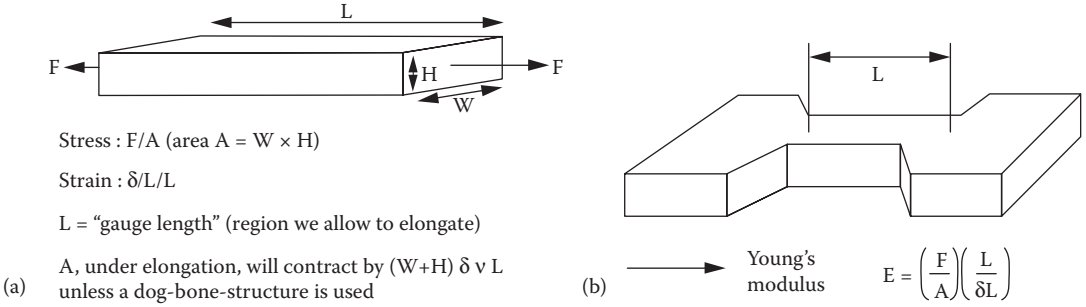
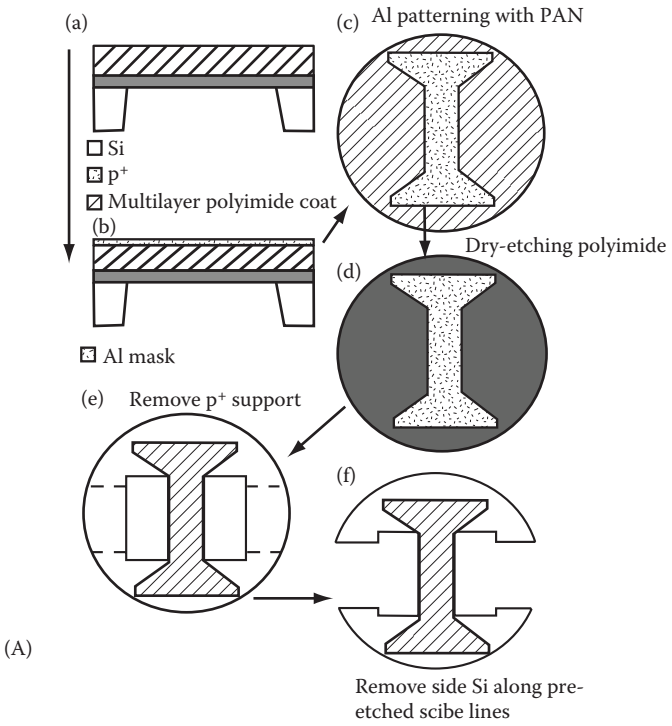


FIGURE 6.53 Measuring Young's modulus. (a) With a bar-shaped structure. (b) With a dog-bone; Instron specimen. (From Senturia, S. D., and R. T. Howe. 1990. Lecture notes. Boston: Massachusetts Institute of Technology.[26])



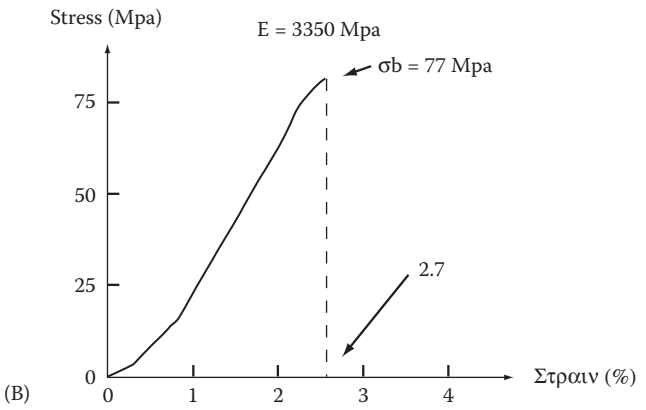


FIGURE 6.54 Uniaxial stress measurement. (A) Fabrication process of a dog-bone sample for measurement of uniaxial strain. (B) Stress versus strain for DuPont's 2525 polyimide. (Courtesy of Dr. F. Maseeh, IntelliSense.)

remaining silicon acts as supports for the grips of the Instron.[27] The resulting structure can be manipulated like any other macrosample without the need for removal of the film from its substrate. This technique enables the gathering of stress/strain data for a variety of commercially available polyimides.[28] A typical measurement result for DuPont's polyimide 2525, illustrated in Figure 6.54B, gives a break stress and strain of 77 MPa (σ_b) and 2.7% (ε_b), respectively, and 3350 MPa for the Young's modulus.

Figure 6.55a illustrates a micromachined test structure able to establish the strain and the ultimate stress of a thin film.[29] A suspended rectangular polymer membrane is patterned into an asymmetric structure before removing the thin supporting Si. Once released, the wide suspended strip (width, W_1) pulls on the thinner necks (total width, W_2), resulting in a deflection, δ, from its original mask position toward the right to its final position after release. The residual tensile stress in the film drives the deformation, δ, as shown in Figure 6.55a. By varying the geometry, it is possible to create structures exhibiting small strain in the thinner sections as opposed to others that exceed the ultimate strain of the film. For structures where the strain is small enough to be modeled with linear elastic behavior, the deflection, δ, can be related to the strain as follows:

$$\varepsilon = \frac{\sigma}{E} = \frac{\delta\left(\frac{W_1}{L_1} + \frac{W_2}{L_2}\right)}{W_1 - W_2} \qquad (6.25)$$

where the geometries are defined as illustrated in Figure 6.55a. Figure 6.55b displays a photograph of two released structures, one with thicker necks, the other with necks so thin that they fractured upon release of the film. On the basis of the residual tensile strain of the film and the geometry of the structures that failed, the ultimate strain of the particular polyimide used was determined to be 4.5%.

Using similar micromachined tensile test structures, Biebl et al.[30] measured the fracture strength of undoped and doped polysilicon and found 2.84 ± 0.09 GPa for undoped material and 2.11 ± 0.10 GPa in the case of phosphorus doping, 2.77 ± 0.08 GPa for boron doping, and 2.70 ± 0.09 GPa for arsenic doping. No statistically significant differences were observed between samples released using concentrated HF or buffered HF. However, a 17% decrease of the fracture stress was observed for a 100% increase in etching time. These data contrast with Greek et al.'s[31] in situ tensile strength test result of 768 MPa for an undoped polysilicon film, a mean tensile strength approximately one-tenth that of single-crystal Si (6 GPa).[32] We normally expect polycrystalline films to be stronger than single-crystal films (see Volume II, Chapter 7). Greek et al.[31] explain this discrepancy for polysilicon by pointing out that their polysilicon

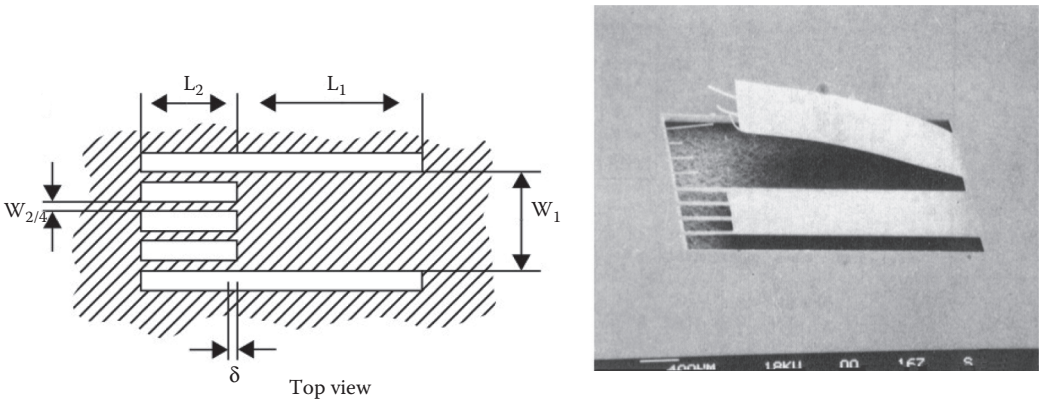


FIGURE 6.55 Ultimate strain. (a) Test structures for stress-to-modulus (strain) and ultimate stress measurements. (b) Two released structures, one of which has exceeded the ultimate strain of the film, resulting in fracture of the necks. (From Senturia, S. 1987. In *Proceedings: IEEE Micro Robots and Teleoperators Workshop* 3/1–5. Hyannis, MA. With permission.[29])

films have a very rough surface compared with single-crystal material, containing many locations of stress concentration where a fracture crack can be initiated.

Biaxial Measurements of Mechanical Properties of Thin Films: Suspended Membrane Methods

Disc stress-measuring techniques are not suitable for measuring stress in low modulus tensile materials such as polyimides, whereas suspended membranes are very convenient for this purpose. The same micromachined test structure used for adhesion testing, sometimes called the *blister test*, as shown in Volume II, Figure 7.56, can measure the tensile stress in low-modulus materials. This type of test structure ensues from shaping a silicon diaphragm by conventional anisotropic etching, followed by applying the coating, and finally removing the supporting silicon from the back with an SF_6 plasma.[29] By pressurizing one side of the membrane and measuring the deflection, one can extract both the residual stress and the biaxial modulus of the membrane. Pressure to the suspended film can be applied by a gas or by a point-load applicator.[33] The load-deflection curve at moderate deflections (strains less than 5%) answers to the following equation:

$$p = C_1 \frac{\sigma t d}{a^2} + C_2 \left(\frac{E}{1-\nu} \right) \frac{t d^3}{a^2} \qquad (6.26)$$

where

p = pressure differential across the film
d = center deflection
a = the initial radius
t = the membrane thickness
σ = the initial film stress

In the simplified Cabrera model for circular membranes, the constants C_1 and C_2 equal 4 and 8/3, respectively. For more rigorous solutions for both circular and rectangular membranes and references to other proposed models, consult Maseeh-Tehrani.[27] The relationship in Equation 6.26 can simultaneously determine σ and the biaxial modulus $E/1 - \nu$; plotting pa^2/dt versus $(d/a)^2$ should yield a straight line. The residual stress can be extracted from the intercept and the biaxial modulus from the slope of the least-squares best-fit line.[29] A typical result obtained via such measurements is represented in Figure 6.56. For the same DuPont polyimide 2525, measuring a Young's modulus of 3350 MPa in the uniaxial test (Figure 6.54B), the measurements give 5540 MPa for the biaxial modulus and 32 MPa for the residual stress. The residual stress-to-biaxial modulus ratio, also referred to as the residual biaxial strain, thus reaches 0.6%. The latter quantity must be compared with the ultimate strain when evaluating potential reliability problems associated with cracking of films. By loading the membranes to the elastic limit point, yield stress and strain can be determined as with the uniaxial test.

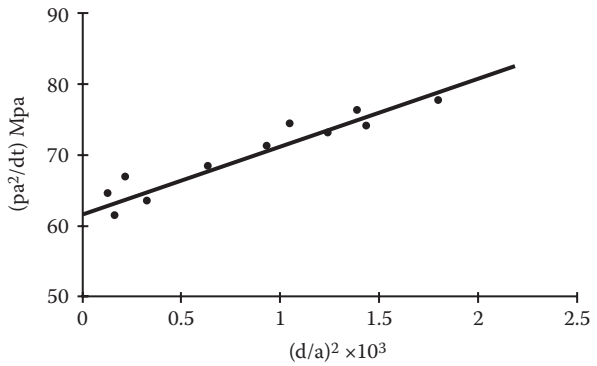


FIGURE 6.56 Load deflection data of a polyimide membrane (DuPont 2525). From the intercept, a residual stress of 32 MPa was calculated, and from the slope, a biaxial modulus of 5540 MPa was calculated. (From Senturia, S. 1987. In *Proceedings: IEEE Micro Robots and Teleoperators Workshop* 3/1–5. Hyannis, MA. With permission.[29])

Poisson Ratio for Thin Films

The Poisson ratio for thin films presents us with more difficulties to measure than the Young's modulus, as thin films tend to bend out of plane in response to in-plane shear. Maseeh and Senturia[28] combine uniaxial and biaxial measurements to calculate the in-plane Poisson ratio of polyimides. For example, for the DuPont polyimide 2525, they determined 3350 MPa for E and 5540 MPa for the biaxial modulus $[E/(1 - \nu)]$, leading to a Poisson ratio (ν) of 0.41 ± 0.1. The errors on both the biaxial and uniaxial measurements need to be reduced in order to develop more confidence in the extracted value of the Poisson ratio. At present, the precision on the Poisson ratio is limited to about 20%.

Other Surface Micromachined Structures to Gauge Intrinsic Stress

Introduction

Various other surface micromachined structures have been used to measure mechanical properties of thin films. We will give a short review here, but the interested reader might want to consult the original references for more details.

Clamped-Clamped Beams

Several groups have used rows of clamped-clamped beams (bridges) with incrementally increasing lengths to determine the critical buckling load and hence deduce the residual compressive stress in polysilicon films (Figure 6.57a).[34,35] The residual strain, $\varepsilon = \sigma/E$, is obtained from the critical length, L_c, at which buckling occurs (Euler's formula for elastic instability of struts):

$$\varepsilon = \frac{4\pi^2}{A}\frac{I}{L_c^2} \tag{6.27}$$

where A is the beam cross-sectional area and I the moment of inertia. As an example, with a maximum beam length of 500 µm and a film thickness of 1.0 µm, the buckling beam method can detect compressive stress as small as 0.5 MPa. This simple Euler approach does not take into consideration additional effects such as internal moments resulting from gradients in residual stress.

Ring Crossbar Structures

Tensile strain can be measured by a series of rings (Figure 6.57b) constrained to the substrate at two points on a diameter and spanned orthogonally by a clamped-clamped beam. After removal of the sacrificial layer, tensile strain in the ring places the spanning beam in compression; the critical buckling length of the beam can be related to the average strain.[37]

Vernier Gauges

Both clamped-clamped beams and ring structures need to be implemented in entire arrays of structures. They do not allow easy integration with active microstructures because of space constraints. As opposed to proof structures, one might use vernier gauges to measure the displacement of structures induced by residual strain.[36] The idea was first explored by Kim,[38] whose device consisted of two cantilever beams fixed at two opposite points. The end movement of the beams caused by the residual strain is measured by a vernier gauge. This method requires only one structure, but the best resolution for strain measurement reported is only 0.02% for 500-µm beams. Moreover, the vernier gauge device may indicate an erroneous strain when an out-of-plane strain gradient occurs.[36] Other types of direct strain measurement devices are the T- and H-shaped structures from Allen et al.[20] and

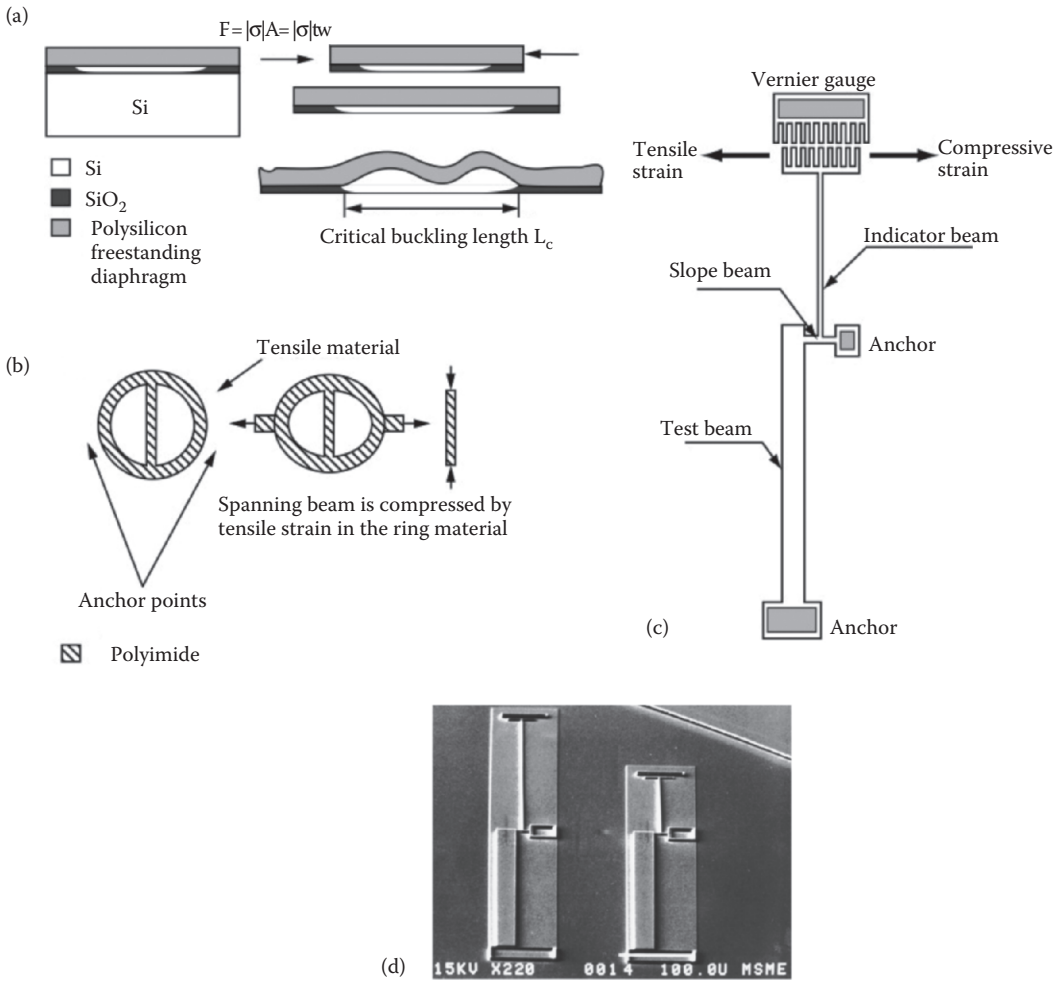


FIGURE 6.57 Some micromachined structures used for stress measurements. (a) Clamped-clamped beams: measuring the critical buckling length of clamped-clamped beams enables measurement of residual stress. (b) Crossbar rings: tensile stress can be measured by buckling induced in the crossbar of a ring structure. (c) A schematic of a strain gauge capable of measuring tensile or compressive stress. (d) Scanning electron microscope microphotograph of two strain gauges. (c and d from Lin, L. 1993. Selective encapsulations of MEMS: Micro channels, needles, resonators, and electromechanical filters. PhD thesis, University of California, Berkeley. With permission.[36])

Mehregany et al.[17] Optical measurement of the movement at the top of the T- or H-shape structures becomes possible only with very long beams (greater than 2.5 mm). They occupy large areas, and their complexity requires finite element methods to analyze their output. The same is true for the strain magnification structure by Goosen et al.[39] This structure measures strain by interconnecting two opposed beams such that the residual strain in the beams causes a third beam to rotate as a gauge needle. The rotation of the gauge needle quantifies the residual strain. A schematic of a micromachined strain gauge capable of measuring tensile or compressive residual stress, as shown in Figure 6.57c, was developed by Lin at University of California, Berkeley.[36] Figure 6.57d represents an SEM photograph of Lin's strain gauge. This gauge by far outranks the various in situ gauges explored. The strain gauge uses only one structure, can be fabricated in situ with active devices, determines tensile or compressive strain under optical microscopes, has a fine resolution of 0.001%, and resists the out-of-plane strain gradient. When the device is released in the sacrificial etch step, the test beam (length L_t) expands or contracts, depending on the sign of the residual stress in the film, causing the compliant slope beam (length L_s) to deflect into an S shape. The indicator beam (length L_i), attached to the deforming beam at its point of inflection, rotates through an angle θ, and the deflection, δ,

is read on the vernier scale. The residual strain is calculated as follows:

$$\varepsilon_f = \frac{2L_s\delta}{3L_iL_tC} \tag{6.28}$$

where C is a correction factor due to the presence of the indicator beam.[36] This equation was derived from simple beam theory relationships and assumes that no out-of-plane motion will occur. The accuracy of the strain gauge is greatly improved, because its output is independent of both the thickness of the deposited film and the cross-section of the microstructure. Krulevitch used these devices to measure residual stress in in situ phosphorus-doped polysilicon films,[23] while Lin tested LPCVD silicon-rich silicon nitride films with it.[36]

An improved micromachined indicator structure, inspired by Lin's work, was built by Ericson et al.[40] By reading an integrated nonius scale in an SEM or an optical microscope, internal stress was measured with a resolution better than 0.5 MPa.[40,41] Both thick (10 µm) and thin (2 µm) polysilicon films were characterized this way.

Lateral Resonators

Biebl et al.[42] extracted the Young's modulus of in situ phosphorus-doped polysilicon by measuring the mechanical response of polysilicon linear lateral comb-drive resonators (see Volume II, Figure 7.76). The results reveal a value of 130 ± 5 GPa for the Young's modulus of highly phosphorus-doped films deposited at 610°C with a phosphine-to-silane mole ratio of 1.0×10^{-2} and annealing at 1050°C. For a deposition at 560°C with a phosphine-to-silane ratio of 1.6×10^{-3}, a Young's modulus of 147 ± 6 GPa was extracted.

Stress Nonuniformity Measurement by Cantilever Beams and Cantilever Spirals

Introduction

The uniformity of stress through the depth of a film introduces a property that it is extremely important to control. Variations in the magnitude and direction of the stress in the vertical direction can cause cantilevered structures to curl toward or away from the substrate. Stress gradients present in the polysilicon film must thus be controlled to ensure predictable behavior of designed structures when released from the substrate. To determine the thickness variation in residual stress, noncontact surface profilometer measurements on an array of simple cantilever beams[43,44] or cantilever spirals can be used.[25]

Cantilever Beams

The deflections resulting from stress variation through the thickness of simple cantilever beams after their release from the substrate are shown in Figure 6.58A. The bending moment causing deflection of a cantilever beam follows out of prerelease residual stress and is given by the following equation:

$$M = \int_{-t/2}^{t/2} zb\sigma(z)dz \tag{6.29}$$

where $\sigma(z)$ represents the residual stress in the film as a function of thickness and b stands for cantilever width. Assuming a linear strain gradient Γ (physical dimensions 1/length) such that $\sigma(z) = E\Gamma z$, Equation 6.29 converts as follows:

$$\Gamma = \frac{12M}{Ebt^3} = \frac{M}{EI} \tag{6.30}$$

where the moment of inertia, I, for a rectangular cross-section is given by $I = bt^3/12$. The measured deflection z, i.e., the vertical deflection of the cantilever's endpoint, from beam theory for a cantilever with an applied end moment, is given as follows.

$$z = \frac{ML^2}{2EI} = \frac{\Gamma L^2}{2} \tag{6.31}$$

Figure 6.58B represents a topographical contour map of an array of polysilicon cantilevers. The cantilevers vary in length from 25 to 300 µm by 25-µm increments. Notice that the tip of the longest cantilever resides at a lower height (approximately 0.9 µm closer to the substrate) than the anchored support, indicating a downward-bending moment.[43] The gradients can be reduced or eliminated with a high-temperature anneal. With integrated electronics on the same chip, long high-temperature processing must be avoided. Therefore, stress gradients can limit the length of cantilevered structures used in surface-micromachined designs.

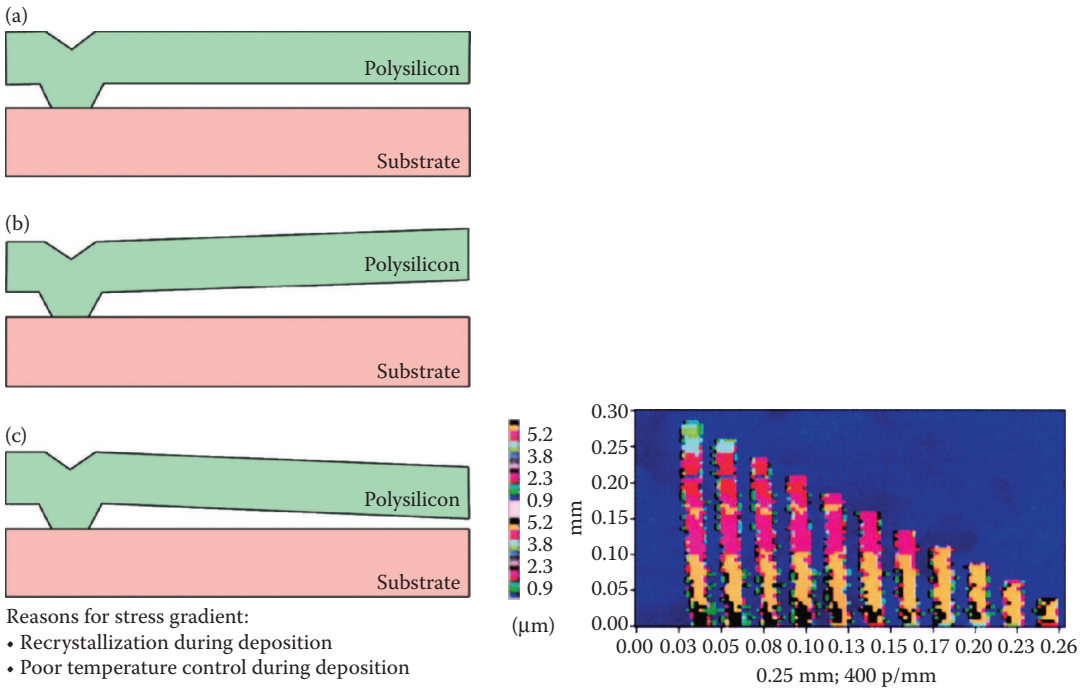


FIGURE 6.58 Microcantilever deflection for measuring stress nonuniformity. (A) Micro-cantilever deflection for measuring stress nonuniformity. (a) No gradient. (b) Higher tensile stress near the surface. (c) Lower tensile stress near the surface. (B) Topographical contour map of polysilicon cantilever array. (From Core, T. A., W. K. Tsang, and S. J. Sherman. 1993. Fabrication technology for an integrated surface-micromachined sensor. *Solid State Technol* 36:39–47. With permission.[43])

The use of silicon-on-insulator (SOI)-based micromachined cantilevers for measuring surface stress induced during adsorption of biomolecules is illustrated in Volume I, Example 4.1.

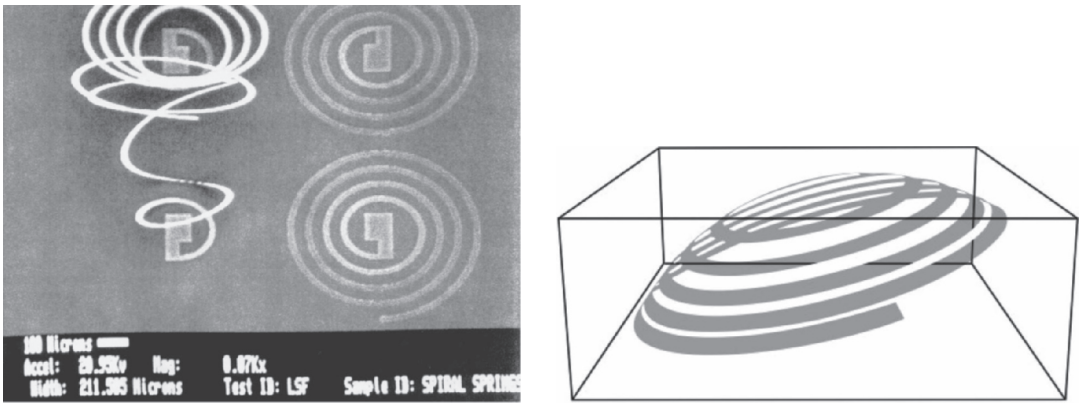
Cantilever Spirals

Residual stress gradients also can be measured by Fan's cantilever spiral, as shown in Figure 6.59a.[45] Spirals anchored at the inside spring upward, rotate, and contract with positive strain gradient (tending to curl a cantilever upward), whereas spirals anchored at the outside deflect in a similar manner in response to a negative gradient. Theoretically, positive and negative gradients produce spirals with mirror symmetry.[45] The strain gradient can be determined from spiral structures by measuring the amount of lateral contraction, the change in height, or the amount of rotation. Krulevitch presented the computer code for the spiral simulation in his doctoral thesis.[23] Figure 6.59b shows a simulated spiral with a bending moment of $\Gamma = \pm 3.0\ \text{mm}^{-1}$ after release.

Krulevitch compared all the above surface micromachined structures for stress and stress gradient measurements on polysilicon films. His comments are summarized in Table 6.10.[23] He found that the fixed-fixed beam structures for determining compressive stress from the buckling criterion produced remarkably self-consistent and repeatable results. Wafer curvature stress profiling proved reliable for determining average stress and the true stress gradient as compared with micromachined spirals. Measurements of curled cantilevers could not be used a lot, as the strain gradients mainly were negative for polysilicon, leading to cantilevers contacting the substrate. The strain gauge dial structures were useful over a rather limited strain-gradient range. With too large a strain gradient, curling of the long beams overshadows expansion effects and makes the vernier indicator unreadable.

Modeling

Modeling is the use of mathematical equations to simulate and predict real events and processes. In this modeling section we introduce finite element analysis and review software packages helpful in the



(a) (b)

FIGURE 6.59 Cantilever spirals for stress gradient measurement. (a) Scanning electron microscope micrographs of spirals from an as-deposited polysilicon. (Courtesy of Dr. L.-S. Fan, IBM, Alamaden Research Center.) (b) Simulation of a thin-film micromachined spiral with $G = 3.0\ mm^{-1}$. (From Krulevitch, P. A. 1994. Micromechanical investigations of silicon and Ni–Ti–Cu thin films. PhD thesis, University of California, Berkeley. With permission.[23])

design phase of miniaturized devices, that is, software for computer-aided design (CAD).

After brainstorming about sensor specifications and sensor transduction principles and making preliminary designs on paper and the white board, it is a good idea to initiate a CAD of the overall microsystem to grasp how all the components, including the package, fit together. In Figure 6.60 we show a three-dimensional visualization (construction) of a NiTi-based valve (the principle of this shape memory type valve is discussed in Chapter 8, this volume, on actuators).

More than visualization is needed though for CAD, and a software package in general is structured as sketched in Figure 6.61 with the design aids used to create the design, simulation to develop the technology, and verification to check the design. The final verification is during fabrication itself; the goal here is to avoid wasteful and slow experiments by carrying out less costly computer work and to get the fabrication "right" the first time.[53]

During design, the process may be separated in a conceptual design and simulation phase and a phase of final design of masks and processes. The ideal suite of CAD tools required for each activity is summarized in Table 6.11;[54] besides visualization, simulation, and verification, a good MEMS CAD package typically includes an extensive materials database.

With CAD for MEMS application in mind, IC development software packages have been expanded with finite element analysis (FEA) programs and

TABLE 6.10 Summary of Various Techniques for Measuring Residual Film Stress

Measurement Technique	Measurable Stress State	Remarks
Wafer curvature	Stress gradient, average stress	Average stress over entire wafer, provides true stress gradient, approximately 5 MPa resolution
Vernier strain gauges	Average stress	Local stress, small dynamic range, resolution = 2 MPa
Spiral cantilevers	Stress gradient	Local stress, provides equivalent linear gradient
Curling beam cantilevers	Large positive stress gradient	Local stress, provides equivalent linear gradient
Fixed-fixed beams	Average compressive stress	Local stress measurement

Source: Krulevitch, P. A. 1994. Micromechanical investigations of silicon and Ni–Ti–Cu thin films. PhD thesis, University of California, Berkeley. With permission.[23]

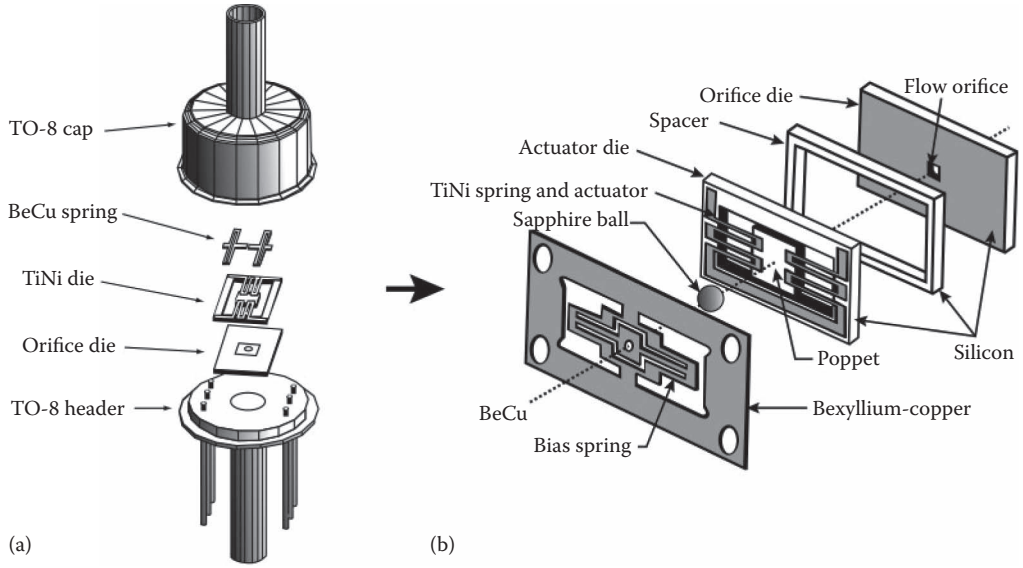


FIGURE 6.60 Microflow Analytical's shape memory alloy valve. Normally closed micro valve. A current through the NiTi die lifts it from the orifice die. (Courtesy of Dr. D. Johnson, TiNi Corporation.)

materials databases. This type of software is paying off, even for IC development where there is a growing need for the ability to perform mechanical analysis of microelectronic devices, both in ensuring structural reliability against failure of thin film layers and in evaluating the effects of various external loads, including the effects of temperature and humidity. An understanding of the required processes and materials properties that ensure reproducibility for commercial MEMS products has emerged slowly, and several CAD systems that have facilitated the wider acceptance of MEMS are now on the market. Most include a materials database that can be updated by the user. Details on CAD packages for MEMS is preceded by a short introduction to finite element analysis. We finish this chapter by referring to software packages of interest in NEMS and bio-NEMS modeling.

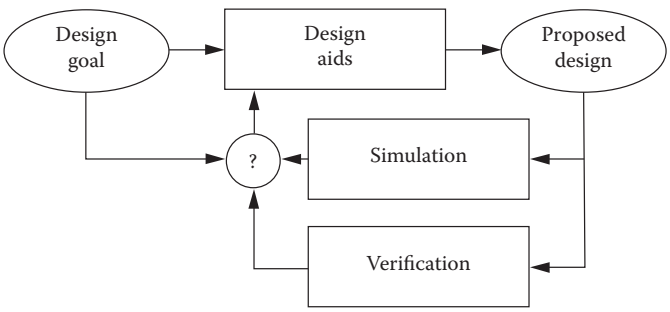


FIGURE 6.61 Design, simulation, and verification.[54]

Finite Element Analysis

Introduction

Finite element analysis is used in almost every engineering discipline. Early FEA packages were limited to simulators using linear static stress analysis. Today, they have been extended to include nonlinear static stress, dynamic stress (vibration), fluid flow, heat transfer, electrostatic forces, etc. These capabilities are frequently combined to perform coupled-force analyses that consider multiple physical phenomena and are tightly integrated within a CAD interface. The best-developed application areas, which also happen to be very important for microsystem design, are electrostatics, stress/strain analysis of solid structures, heat conduction analysis, and fluid dynamics. FEA typically includes the steps outlined below, where we will assume that we are dealing with a stress/strain analysis of a solid structure.

Geometry Development

The essence of FEA is "divide and conquer," and the first step consists of dividing the whole MEMS structure under analysis into a finite number of subdivisions of special shapes or elements that are interconnected at specific points, called *nodes* (Figure 6.62). The nodes are discrete points of the MEMS structure where the analysis will reflect the

Table 6.11 Ideal Tool Suites for MEMS

Conceptual Design and Simulation	Final Design of Masks and Process
Rapid construction and visualization of three-dimensional solid models	Process simulation or process database, including: • Lithographic and etch process biases (the difference between as-drawn and as-fabricated dimensions) • Process tolerances on thicknesses, lateral dimensions, doping and resistivity levels
A database of materials properties	Design optimization and sensitivity analysis: • Variation of device sizing to optimize performance • Analysis of effects of process tolerances
Simulation tools for basic physical phenomena. For example: • Thermal analysis: heat flow • Mechanical and structural analysis: deformation • Electrostatic analysis: capacitance and charge density • Magnetostatic analysis: Inductance and flux density • Fluid analysis: pressures and flow	Mask layout
Coupled-force simulators. For example: • Thermally induced deformation • Electrostatic and magnetostatic actuators • Interaction of fluids with deformable structures	Design verification, including: • Construction of a three-dimensional solid model of the design, using the actual masks and process sequence • Checking the design for violation of any design rules imposed by the process • Simulation of the expected performance of the design, including the construction of macromodels of performance useable in circuit simulators to assess overall system performance
Formulation and use of macromodels. For example: • Lumped mechanical equivalents for complex structures • Equivalent electric circuit of a resonant sensor • Feedback representation for coupled-force problems	

Source: Senturia, S. D., and R. T. Howe. 1990. Lecture notes. Boston: Massachusetts Institute of Technology.[26]

response of the component to an applied loading. This response is defined in terms of nodal degrees of freedom. In the case of stress analysis, up to six degrees of freedom are possible at each node (three components of translation and three components of rotation), depending on the element type selected (e.g., beam, plate, 2D, and 3D elements). Element selection is a function of product geometry and loading conditions. The element selected affects the results as each element has characteristic properties. A model can use more than one type of element.

Material Property Assignment

In the second stress/strain finite element analysis step, the modulus of elasticity or rigidity and Poisson ratio are used to define material properties for each element. For a nonlinear material condition (elastic-plastic), a stress-strain curve represents the material properties.

Mesh Generation

The grid of connecting elements at common nodes is called the *mesh* (Figure 6.62). Based on the element

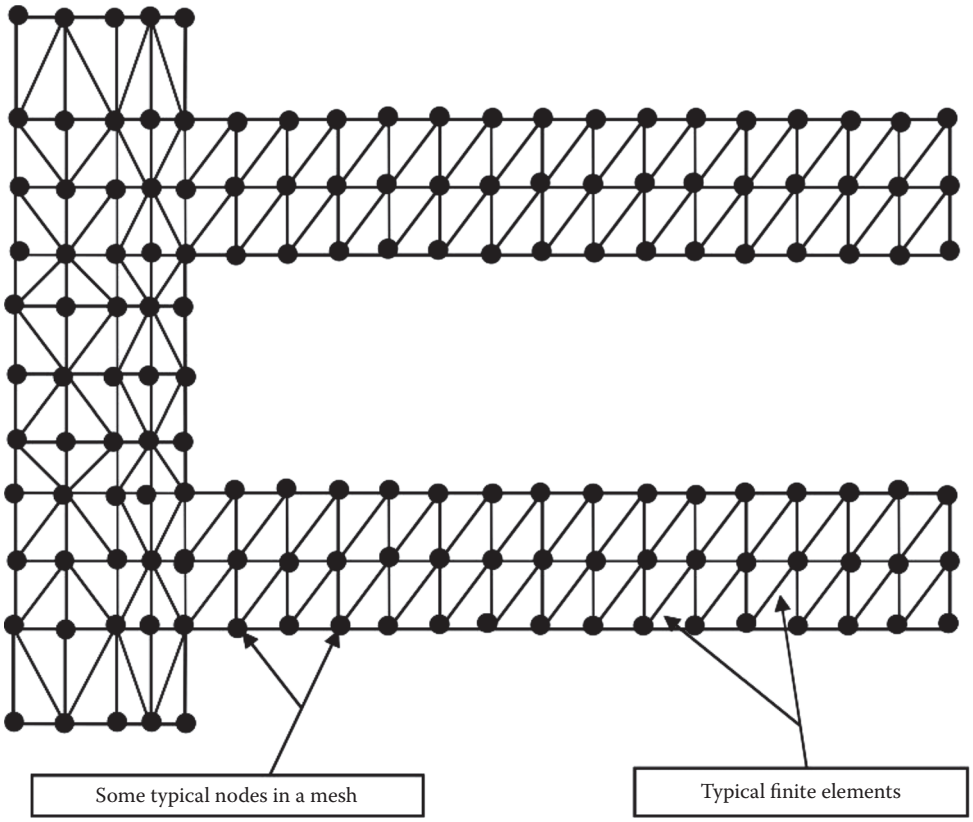


FIGURE 6.62 Finite element analysis mesh.

types selected, automatic mesh generation subdivides the geometry into finite elements in a third step of the FEA process. The element density within each segment of the structure is either chosen or automatically determined. When adjacent elements share nodes, the displacement field is continuous across the shared element boundaries and loads can be transferred between the elements. Analysis is performed on the elements instead of the whole structure, and solutions at the element level are "assembled" to get the corresponding solutions for the whole structure. Mathematically, this is achieved by calculating nodal stiffness properties for each element and arranging these into matrices. The appropriate matrix transformation generates a global stiffness matrix from the existing element matrix.

Boundary and Loading Conditions

In a fourth step, the appropriate boundary and loading conditions apply constraints to the model. In stress analysis, the boundary conditions specify displacement constraints (i.e., in the *x*-, *y*-, or *z*-direction), and the loading conditions apply loads to the model such as concentrated forces at specified nodes or pressure at specified edge surfaces of the element.

Run Analysis

During analysis, in a fifth step, the program processes the equation matrices with the applied loads and boundary conditions to calculate displacements, strain, natural frequencies, or other data specified by the user. This provides a stress distribution across the entire MEMS structure. The high-stress regions should have the highest element density, as a finer mesh increases the accuracy of the model. Various adaptive methods find the critical regions in the model and make the necessary mesh refinement to reduce the error for the next iteration before reaching convergence.

Results

Results obtained in a sixth step include nodal and element information, displacement at nodes, stresses and strains in each element, and various forms of graphical display of the solution. In this regard, a von Mises stress plot often is used. The von

Mises stress (also called *effective* or *equivalent stress*) is defined as follows:

$$\sigma_M = \sqrt{\frac{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_1 - \sigma_3)^2}{2}} \quad (6.32)$$

where σ_1, σ_2, and σ_3 are the three principal stresses. The von Mises stress is compared with the yield strength for plastic yielding and for the prediction of rupture of the structure.[55]

Data Correlation and Design Optimization

Finally, in a seventh step, experimental data are collected to correlate the FEA model results and to formulate a baseline. After comparing the baseline results, design modification and remodeling are available. This iterative process constitutes design optimization, namely, combining the engineering requirements, geometric parameters, CAD model, and performance goals into a computer simulation to achieve the optimum design.

As mentioned in the introduction, in addition to considering a part's ability to withstand mechanical stresses, today's FEA software also enables engineers to predict other real-world stresses. These might include the effects of extreme temperatures or temperature change (heat transfer analysis), the flow of fluids through and around objects (fluid flow analysis), or voltage distributions over the surface or throughout the volume of an object (electrostatic analysis). As these effects often are coupled, it is important that the FEA program consider their effects on one another. For instance, a MEMS chip may be heating up over time, cooling down by airflow from a fan, vibrating against other parts, and be electrically charged. A typical approach is to isolate and calculate each variable then feed the results into the FEA program one at a time. But since each variable could also affect the others, either a coupled analysis or tools for relating results is required.

CAD for MEMS

Introduction

CAD tools were essential to the evolution of MEMS from laboratory status to a bona fide and accessible manufacturing process. Without them, fabrication would have remained in the domain of experts, and evolution of the design process would still rely on empirical approaches. Early developments in CAD for MEMS evolved from 2D IC circuit work and often lacked the tools for a priori design of micromechanical devices in 2.5 or three dimensions essential for high-aspect ratio microsystems. Given the freedom to fabricate 3D systems, designers think in terms of the whole system, not a series of 2D reticulations of the system (i.e., masks). The mask-to-model approach is still somewhat suitable for surface micromachined devices but becomes less useful as the number of processes, materials, and mechanical degrees of freedom increases, as in the case of high-aspect ratio micromachining (especially for a process such as LIGA).[56] A number of current commercial CAD packages for MEMS are competing for market share, and the codes have been improving steadily. For the future, CAD programs addressing generic miniaturization problems rather than just micromachining for Si and poly-Si would be highly desirable. The architecture should have both lithographic and nonlithographic machining options, adhere to accepted format standards for data communication, and integrate available design, modeling, and simulation software from both traditional and nontraditional (i.e., IC based manufacturing) industries wherever possible.

Derived CAD Packages for MEMS

For some of the simplest micromachined structures, modest modification of existing IC software packages enabled designing of the micromachine as in regular IC processing. Marshall et al.,[57] for example, modified the standard Magic VLSI mask layout package (http://opencircuitdesign.com/magic/archive/papers/Magic_A_VLSI_Layout_System.pdf) to permit the design of simple micromachined structures using CMOS processes. This modification makes standard foundry work through the MOSIS service (http://www.mosis.org) possible.[58] Other traditional IC mask layout software packages are used by MEMS engineers from companies such as TMA and Cadence. In the public domain, one can turn to KIC from the University of California, Berkeley.[59] For a comparison of various layout design and verification software packages, visit

http://www.ee.rochester.edu/users/sde/cad/ LDV.html. Mechanical drafting tools with interfaces to mask making, such as AutoCAD, have full geometric flexibility. Other popular mechanical CAD tools for 3D solid model construction, meshing, application of loads and boundary conditions, and visualization of results include Pro/Engineer, PATRAN, and I-DEAS. Suppliers of CAD systems to the semiconductor industry also offer one- and two-dimensional process simulators. These simulators cover standard semiconductor processes such as oxidation, diffusion, thin film deposition, and plasma etching. An example is SAMPLE[60–62] for simulation of projection lithography, deposition, and etching. Simulation tools for basic physical properties include ABAQUS for structural and thermal finite-element simulation and FASTCAP (with its follow-on MEMCAP) (available for download from http://www.rle.mit.edu/cpg/research_codes.htm) for electrostatic boundary-element simulation.

COSMOS (http://www.cosmosm.com) has a MEMS analysis capability for mechanical, thermal, electrostatic, electromagnetic, and fluid flow modeling. COSMOS sells a version that is tightly integrated with the 3D modeling software SolidWorks™. SolidWorks is a player in the mechanical CAD tools arena for 3D solid model construction. Its tight integration with the COSMOS FEA package also enabled it to enter the MEMS simulation industry. AnSoft (http://www.ansoft.com) has software products for 2D and 3D electromagnetic field and thermal field finite-element simulation capabilities applicable to RF-MEMS analysis and design. AnSoft's products integrate electromagnetic, circuit, and system engineering modules. Field equation solvers such as QuickField (http://www.quickfield.com) are increasingly being used for MEMS modeling. Such software provides a quick analysis tool for 2D field problems such as electromagnetics, heat transfer, and fluid flow.

CAD Specifically Developed for MEMS

Some early examples of CAD programs developed specifically for MEMS include Oyster from IBM,[63] MEMCAD from Massachusetts Institute of Technology (MIT),[64] CAEMEMS from the University of Michigan,[65] and the SESES system from ETH (Zurich).[66] These programs mainly address bulk Si micromachining and poly-Si surface micromachining applications.

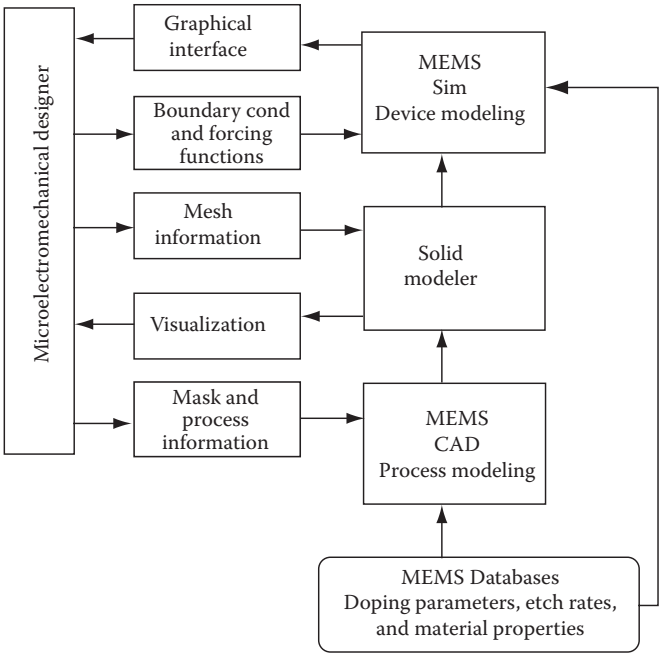


FIGURE 6.63 Block diagram of the CAEMEMS system. [Crary, S., and Y. Zhang. 1990. In *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '90)*, 113–14. Napa Valley, CA. With permission.[65]]

The CAEMEMS program provides a database of material properties and process model parameters, a process modeler that takes process information as input and produces a solid model of the device to be fabricated, a solid modeler that allows for visualization and design verification, and a device modeler that performs finite element simulations of components and systems. The database can be updated by the user as new information becomes available. A block diagram of the CAEMEMS software package is shown in Figure 6.63.[65] ANSYS Multiphysics is the version of ANSYS FEA software that has been used for MEMS modeling. ANSYS Multiphysics has a broad physics capability, such as structural modal, static, transient, electrostatic-structural, high-frequency and low-frequency electromagnetics, electrothermal-structural coupled physics, free surfaces and capillary action, capacitance extraction, Newtonian/non-Newtonian continuum flow, and fluidic structural capability to evaluate damping effects, among others.

The architecture of MEMCAD, illustrated in Figure 6.64a, integrates various simulators, databases, and a solid state modeler with a user interface.[54] As outlined by the dashed blocks in the figure, the CAD system consists of three sections: the microelectronic CAD section, the mechanical CAD section, and the material property simulator. The interactions among these three sections and the flow of information is denoted by the direction of the arrows. In MEMCAD V1.0, the primary interface for mechanical modeling was through PATRAN. MEMCAD V2.0 provides this function from I-DEAS. In the microelectronic section, the mask is created using KIC. SUPREM 3 and SAMPLE are integrated to provide depth and cross-sectional modeling capabilities. Microcosm licensed the MEMCAD technology from MIT and released the first redesigned commercial version in June 1996. Besides MEMCAD V4.0, Microcosm also developed FlumeCAD, geared toward modeling of microfluidic devices. In January 2001 Microcosm Technologies changed its name to Coventor (http://www.coventor.com).

IntelliSense Corp. sells IntelliSuite™ (http:// www.intellisense.com),[67] incorporating the MEMaterial® database and the anistropic etching module AnisE®. The material database contains electrical, mechanical, optical, and physical properties of semiconductor thin films collected from the literature. As in the case of CAEMEMS, the database can be updated by the user. The architecture of IntelliCAD, with a central graphical user interface, is shown in Figure 6.64b. Figure 6.65 shows a typical output of MEMaterial. The selected 3D plot shows the density variation of a plasma-enhanced chemical vapor-deposited Si nitride film as a function of deposition temperature and pressure. AnisE simulates single-crystal silicon anistropic etching. It predicts the effects of etchant's temperature, concentration, and etch time on the final 3D geometry. Users can study etch stops, corner compensation, higher-order etch planes, and process tolerances for single- or double-sided masks. The program added RF-MEMS design tool (2002) and a bio-MEMS design tool (2003).

Other simulators for anistropic etching of Si are available. For example, the Anisotropic Crystalline Etch Simulation (ACES™) is a first personal computer-based 3D etch simulator from the University of Illinois at Urbana-Champaign. The program can simulate silicon etching with different front-surface orientations in different etchants. 2D and 3D simulation programs for anisotropic etching profiles have also been developed by Sequin at the University of California, Berkeley,[68] Koide et al. at Hitachi,[69] and Li et al. at Caltech.[70] These programs enable the prediction of a change in cross-sectional shape of a feature in a Si wafer with arbitrary crystallographic orientation and with a mask including concave and convex edges.

Coventor now markets CoventorWare,™ which consists of a system-level modeling environment, a 2D layout editor, materials property editor and database, a 3D model generator, and a multiphysics numerical analysis framework. It has also added a new module that interfaces with IC design environment by exporting models to IC simulators from Synopsys and Cadence.

CFD Research Corporation (CFDRC) is another provider of advanced software tools for modeling, simulation, and design of MEMS. Their software has been used for the analysis of a wide variety of physical phenomena such as fluid flow, heat transfer, combustion, and fluid-structure interaction. Industrial applications include MEMS, electronics packaging, semiconductor turbomachinery, and automotive engineering (http://www.cfdrc.com/).

MEMS Pro, originally from MEMSCAP, provides a unified MEMS and IC design environment. It includes certain software elements licensed from Tanner Research. MEMS Pro was sold in 2004 to SoftMEMS (http://www.softmems.com/) and is still marketed under the same name. It provides an integrated MEMS and IC- design environment with system-level tools, layout editing, 3D Solid Modeler (3D view generator from layout and fabrication process description), FEA modeling tools, and MEMS verification tool for design rule checks.

SUGAR V3.1 from University of California, Berkeley, is a free simulation program for MEMS devices composed of beams, electrostatic gaps, anchors, circuits elements etc. The approach is based on a modified nodal analysis to solve coupled nonlinear differential equations. The simulation algorithms are implemented in MATLAB®

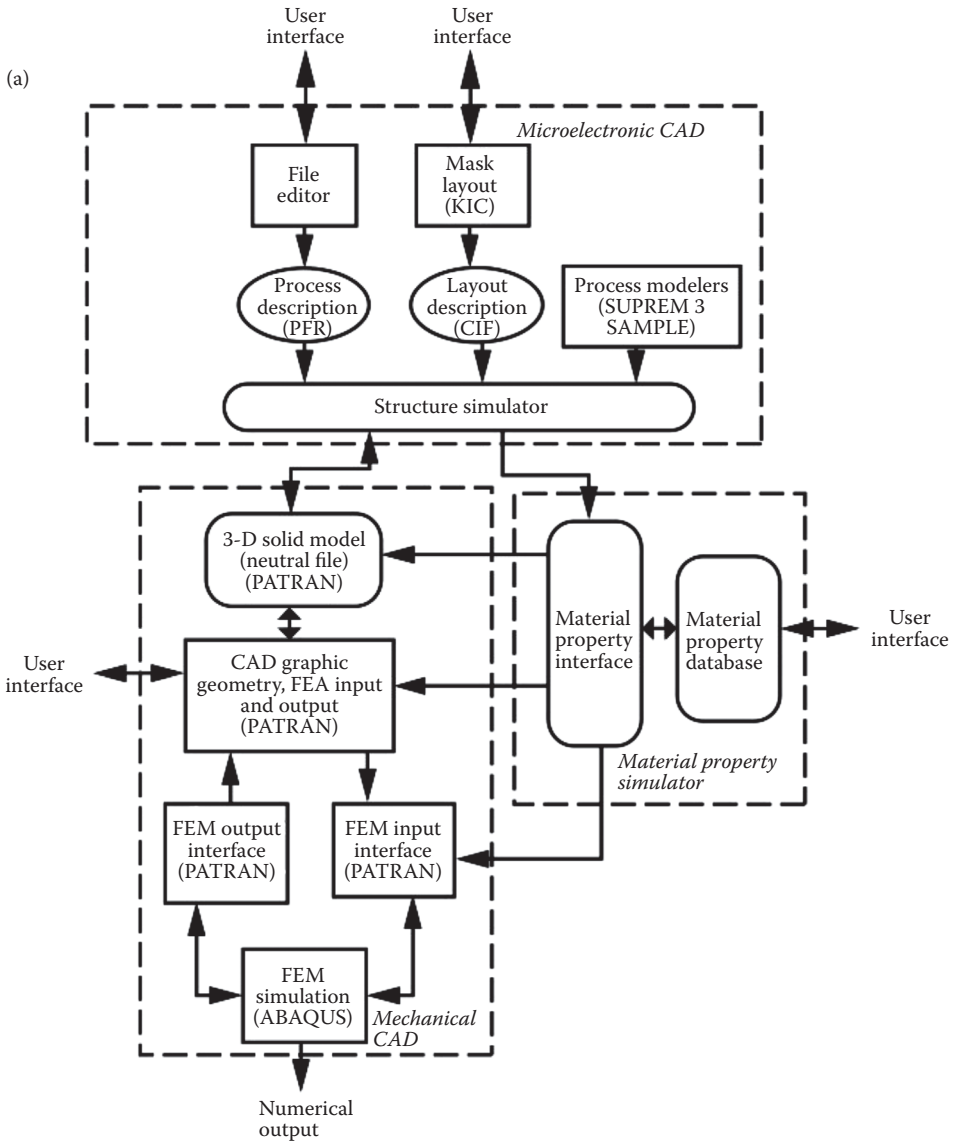


FIGURE 6.64 (a) Outline of the Massachusetts Institute of Technology MEMCAD. (b) Schematic of IntelliCAD architecture. (Courtesy of Dr. F. Maseeh, IntelliSense.)

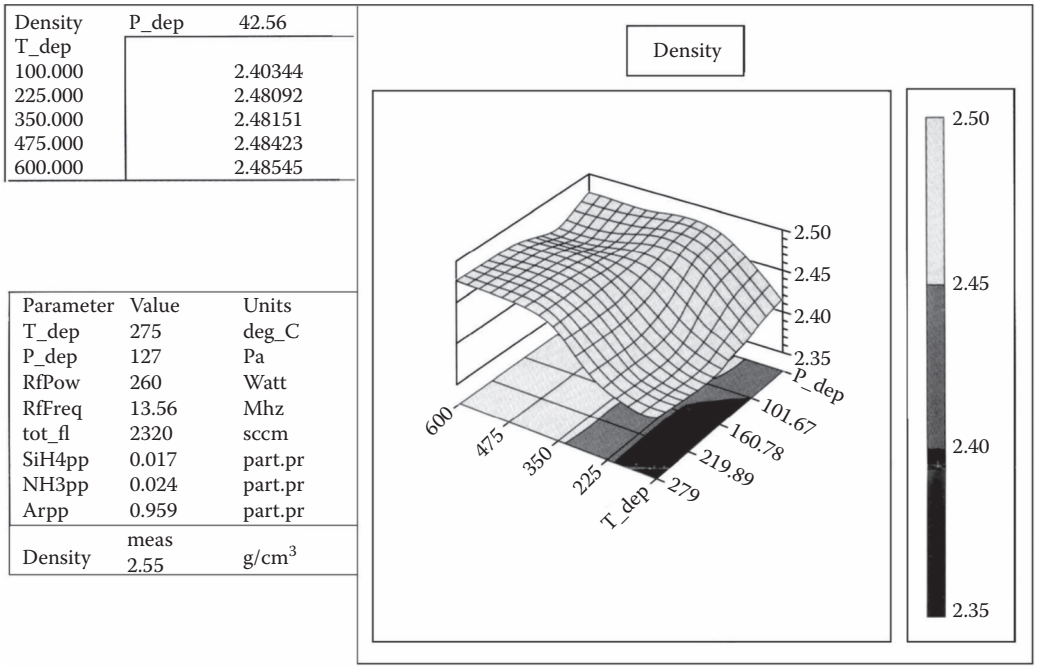


FIGURE 6.65 The MEMaterial design window from IntelliSense Corp. The three-dimensional graph shows the variation of Si nitride density deposited by plasma-enhanced chemical vapor-deposition. The table at the bottom left shows the process parameters, their numerical values, and units. The resulting density is shown at the bottom of the table. The user can alter any of the parameters. (Courtesy of Dr. F. Maseeh, IntelliSense.)

and are portable across all Unix and personal computer platforms.

FEMLAB, marketed by COMSOL (http://www.comsol.com), has entered the MEMS market with its MEMS Module. FEMLAB is a partial differential equation-based FEA modeling tool. It has modules for mechanical, thermal, DC and AC electrical, and fluidic modeling of devices among others. The FEMLAB MEMS Module includes ready-made applications that cover microfluidics plus electromagnetic-structural, thermal-structural, and fluid-structural interactions. Its MEMS-specific application modes are electrostatics, stress and strain, piezoelectric, and electrokinetics.

ALGOR (http://www.algor.com) has a MEMS module for mechanical, thermal, and fluid flow modeling. It also has electrostatic analysis and piezoelectric material model capabilities.

The major improvements in the leading MEMS software have been in a tighter integration with solid modeling as well as IC design environments, system-level modeling for quick analysis, capabilities of modeling microfluidics for the bio-MEMS industry, and capabilities for modeling RF-MEMS. Some FEA software packages have also added MEMS modules.

Future of CAD for MEMS and NEMS

Future MEMS CAD packages should seamlessly integrate 2D IC processes with 3D MEMS and traditional machining in a wide variety of materials. A conspicuous example for the need of such a 3D microdesign tool is in micromolding of LIGA shapes, in which nontraditional (lithography) and traditional (plastic molding) technologies are combined (see Volume II, Chapter 10). As an example of work in this direction, Hill et al.[71] explored the use of the computer software I-DEAS Master Series™ Thermoplastic Molding to model the micromolding aspect of such LIGA parts and have found that it accurately describes the filling characteristics of the microparts. What is needed now is to extend this package to include lithography steps.

A good introduction to the available CAD programs for IC processes comes from Fichtner.[72] A most detailed account of simulation and design issues of microsystems and microstructures is by Adey et al.[73] For access to continuously updated information on MEMS modeling, visit the MEMS Interchange on the Web (http://www-mtl.mit.edu/semisubway/semi subway.html); equally useful is http://www.memsnet.org.

With the advent of NEMS and bio-MEMS, molecular modeling has come to the forefront. Molecular modeling is a collective term that refers to theoretical methods and computational techniques to model or mimic the behavior of molecules. For molecular modeling (say on protein or DNA sensors), software can now successfully model the dynamics of most molecular interactions under numerous static and dynamic conditions. Molecular dynamics (MD) is a specialized discipline of molecular modeling and computes the behavior of a system as a function of time. The main justification of the MD method is that statistical ensemble averages are equal to time averages of the system. Because molecular systems generally consist of a vast number of particles, it is impossible to find the properties of such complex systems analytically; MD simulation circumvents this problem by using numerical methods. Molecular dynamics simulations are playing an increasingly important role in many areas of science and engineering, from biology and pharmacy to nanoelectronics and structural materials. We review only some available software packages for molecular modeling.

A lot of the modeling work is geared toward DNA and proteins; software is available to graphically manipulate 3D images of proteins and nucleic acids. The modules perform complex molecular and quantum mechanical calculation and protein structure homology analysis. Files in standard formats such as that of the Protein Data Bank can be imported.

Some other programs to check out are the following:

- CPMD (Car-Parrinello Molecular Dynamics)—This code can be used to perform *ab initio* molecular dynamics calculations. It allows for time-dependent density functional calculations, wave function optimization, and path-integral molecular dynamics.
- Siesta (Spanish Initiative for Electronic Simulations with Thousands of Atoms)—This code uses numerically truncated orbitals (single- and double-zeta approach) to build an order-N density functional code. This code is ideal for modeling large scale nanostructures (i.e., nanotubes, nanowires, and molecules).
- NWChem—This code is a highly parallel quantum chemistry package capable of self-consistent field, Hartree-Fock (RHF, UHF), and Gaussian density functional calculations, as well as several other techniques.
- Atomistix (ATK and VNL)—This commercial package provides the ability to calculate I-V curves in nanostructures and molecular junctions. The code is based on a nonequilibrium Green's function approach, and it also takes advantage of numerically truncated orbitals to allow for calculations of large systems.
- LAMMPS—This is a general purpose molecular dynamics simulator that has the option to use Leonard Jones potentials, embedded atom potentials, and potentials for biomolecules and proteins. This parallel code can easily handle systems with thousands of atoms. The ability to incorporate the effect of temperature provides an important complement to density functional techniques.
- MPB (MIT Photonic Bands package)—This code can calculate the band structure and electromagnetic modes of periodic dielectric structures.
- UTQUANT—This is a quasi-static CV simulator for one-dimensional silicon MOS structures.
- SEMC-2D (Schrödinger Equation Monte Carlo 2D)—This is a simulator for quantum transport and inelastic scattering effects in nanoscale semiconductor devices, such as nanoscale double-gate MOSFETs and tunnel injection lasers.

Questions

Questions by Chuan Zhang, UC Irvine

6.1: Compare the similarities and differences of scanning near-field optical microscopy (NSOM) and traditional optical microscopy. Explain how NSOM exceeds the resolution limit of traditional optical microscopes.

6.2: Classical far-field microscopy was introduced by Anton van Leeuwenhoek in the 16th century. It was a major breakthrough that played an important role in the advancement of science, engineering, and medicine. Briefly explain what is the underlying physical

limitation regarding the magnification/resolution of classical far-field optical microscopy.

6.3: What does the NSOM acronym stand for? What are the principles of NSOM? What is the difference between apertureless NSOM and aperture-based NSOM and what dictates their resolution?

6.4: Scanning tunneling microscopes (STMs) and atomic force microscopes (AFMs) are two so-called *proximal probe techniques* with the power to reveal the atomic world to us.
(a) What is the difference between these two?
(b) A crucial component in an STM is the integrated sharp tip, which has a curvature radius of about 50 Å, an angle of aperture of 5°, and a height of 10 µm. Sketch a process to fabricate this integrated tip on a Si wafer.

6.5: What is the effect of the following changes on quantum yield of fluorescence? Explain the reason for your answer.
(a) Raising the temperature.
(b) Adding a heavy atom to the solvent.
(c) Lowering solvent viscosity.
(d) What is the physical process that makes phosphorescence rare?

6.6: What does AFM stand for and what is the principle of operation of AFM? What kind of information can you obtain by using AFM? Briefly describe an application where AFM is employed.

6.7: Briefly state the difference between fluorescence and phosphorescence.

6.8: Please define the following analytical terms. Use equations when appropriate.
(a) Beer's law
(b) Transmittance
(c) Quantum yield
(d) Sensitivity
(e) Precision
(f) Accuracy

6.9: Suggest the type of system that you would employ for the separation of the following compounds. Justify your choice.
(a) DNA with a size of 500 bases in length and DNA with a size of 2,000 bases in length
(b) A protein of a MW = 15,000 Daltons and another protein of a MW = 100,000 Daltons
(c) A protein with a MW = 50,000 Daltons and a fluorophore with a MW = 300
(d) A mixture of long-chain unsaturated aliphatic compounds

6.10: Compare the merits of an AFM with transmission electron microscopy (TEM) and a scanning electron microscopy (SEM).

6.11: Why do the samples for TEM need to be extremely thin, while those for SEM don't need to be?

6.12: Both energy dispersive spectroscopy (EDS) and Auger electron spectroscopy (AES) can be utilized to conduct the composition analysis of a sample. Which one of them can be used to analyze light elements such as Li, Be, and C?

6.13: How does two-photon confocal microscopy enhance the resolution of fluorescence microscopy?

6.14: List the signals generated when electrons interact with a solid. What information about a material can be obtained from each of those signals?

6.15: Explain the Auger process.

6.16: Review the various steps in a typical finite element analysis (FEA), e.g., a stress-strain problem.

6.17: The identity of a semiconductor can be determined by measuring its intrinsic carrier concentrations at various temperatures. An unknown semiconductor is submerged in ice water (273K) and boiling water (373K). Its intrinsic carrier concentrations, n_i, were found to be 4.6×10^{18} m^{-3} and 3.1×10^{20} m^{-3} at 273K and 373K, respectively. What is the bandgap and identify of the semiconductor?

Further Reading

ASTM Standard Guide for Computed Tomography (CT) Imaging. Designation E 1441–93, Philadelphia: ASTM.

Binnig, G., C. Quate, and C. Gerber. 1986. Atomic force microscopy. *Phys Rev Lett* 56:930–33.

Binnig, G., and H. Rohrer. 2000. Scanning tunneling microscopy. *IBM Journal of Research and Development* 44(1–2):279–93.

Cao, S., U. Brand, T. Kleine-Besten, W. Hoffmann, H. Schwenke, S. Bütefisch, and S. Büttgenbach. 2002. Recent developments in dimensional metrology for microsystem components. *Microsystems Technologies* 8:3–6.

Ceremuga, J. 2003. Obtaining inspection of high aspect ratio microstructures using a programmable optical microscope. Masters thesis, Georgia Institute of Technology, Atlanta, GA.

Chinn, D., P. Ostendorp, L. Garrett, P. Guthrei, M. Haugh, R. Kershmann, and T. Kurfess. 2002. Three dimensional imaging of LIGA-made micromachines using digital volumetric imaging. *Proceedings of the Japan-USA Symposium on Flexible Automation*, Hiroshima, Japan, Vol. 1:565–68.

Chinn, D., P. Ostendorp, M. Haugh, R. Kershmann, T. Kurfess, A. Claudet, and T. Tucker. Three dimensional imaging of LIGA-made microcomponents. *ASME Journal of Manufacturing Science and Engineering* 126:813–21.

Corle, T., and G. Kino. 1996. *Confocal Scanning Optical Microscopy and Related Imaging Systems.* San Diego: Academic Press.

El-Hakim, S. 1990. Some solutions to vision dimensional metrology problems. *Close-Range Photogrammetry Meets Machine Vision*. SPIE 1395:480–87.

Fan, K., C. Chu, and J. Mou. 2001. Development of a low-cost autofocusing probe for profile measurement. *Measurement Sci Tech* 12:2137–46.

Fukiwara, M., A. Yamaguchi, K. Takamasu, and S. Ozono. 2001. Evaluation of stages of nano-CMM. *Initiatives of Precision Engineering at the Beginning of a Millennium*, 634–638, New York: Springer.

Griffith, J., H. Marchman, G. Miller, and L. Hopkins. 1995. Dimensional metrology with scanning probe microscopes. *J Vacuum Sci Tech* 13(3):1100–05.

Groot, P., and L. Deck. 1994. Surface profiling by frequency-domain analysis of white light interferograms. *Proceedings of SPIE—The International Society for Optical Engineering* 2248:101–04.

Hall, A., and L. Degertekin. 2002. Integrated optical interferometric detection method for micromachined capacitive acoustic transducers. *Appl Phys Lett* 80:3859–61.

Heller, A. 2003. Nondestructive characterization at the meso-scale. *Science and Technology Review*, Livermore, CA: Lawrence Livermore National Laboratory.

Hibbard, R., and M. Bono. 2003. Meso-scale metrology tools: A survey of relevant tools and a discussion of their strengths and weaknesses. *Proceedings of Machines and Processes for Micro-scale and Meso-scale Fabrication, Metrology, and Assembly*, 70–72.

Kim, B., A. Razavi, L. Degertekin, and T. Kurfess. 2003. Micromachined interferometer for MEMS metrology. *Proceedings of Machines and Processes for Micro-scale and Meso-scale Fabrication, Metrology, and Assembly*, 73–78.

Kim, B., A. Razavi, F. Degertekin, and T. Kurfess. 2002. Micromachined interferometer for measuring dynamic response of microstructures. *Proceedings of 2002 ASME International Mechanical Engineering Congress, and Exposition, MEMS Symposium*, New Orleans, LA.

Kim, B., Schmittdiel, M. C., Degertekin, F. L., and T. R. Kurfess. 2004. Scanning grating microinterferometer for MEMS metrology. *ASME Journal of Manufacturing Science and Engineering* 126:807–12.

Kirkland, E. 2003. A nano coordinate machine for optical dimensional metrology. Master's thesis, Georgia Institute of Technology, Atlanta GA.

Kramar, J., J. Jun, W. Penzes, F. Scire, E. Teague, and J. Villarrubia. 1999. Grating pitch measurements with the molecular measuring machine. *SPIE 3806*, *Conference on Recent Advances in Metrology, Characterization, and Standards for Optical Digital Data Disks*, 46–53.

Kramar, J., E. Amatucci, D. Gilsinn, J. Jun, W. Penzes, F. Scire, E. Teague, and J. Villarrubia. 1999. Toward nanometer accuracy measurements. *SPIE* 3677, *Metrology, Inspection and Process Control for Microlithography XIII*, 1017–28.

Kurfess, T., D. Chinn, P. Ostendorp, A. Claudet, and T. Tucker. 2002. Metrology for micro-components. *Proceedings of the JSME/ASME International Conference on Materials and Processing 2002*, Honolulu, Hawaii, 2:77–82.

Lagerquist, M., W. Bither, and R. Brouillette. 1996. Improving SEM linewidth metrology by two-dimensional scanning force microscopy. *SPIE 2725*, *Metrology, Inspection and Process Control for Microlithography X*, 494–503.

Larrabee, R., and M. Postek. 1994. Parameters characterizing the measurement of a critical dimension. *Proceedings of SPIE—The International Society for Optical Engineering* CR52:2–24.

Li, X., M. Rahman, K. Liu, K. Neo, and C. Chan. 2003. Nano-precision measurement of diamond tool edge radius for wafer fabrication. *J Materials Process Tech* 140:358–62.

Mack, C., S. Jug, R. Hones, P. Apte, S. Williams, and M. Pochkowski. 2001. Metrology and analysis of two dimensional SEM patterns. *SPIE 4344*, *Metrology, Inspection and Process Control for Microlithography XV, 169–76*.

Marchman, H. 1996. Nanometer-scale dimensional metrology with noncontact atomic force microscopy. *SPIE* 2725, *Metrology, Inspection and Process Control for Microlithography X*.

Marchman, H., and N. Dunham. 1998. AFM: A valid reference tool? *SPIE 3332*, *Metrology, Inspection and Process Control for Microlithography XII*, 2–9.

Marchman, H. 1998. Critical dimension metrology. *Microlithography: Science and Technology*, New York: Marcel Dekker, Inc.

Marschner, T., G. Eytan, and O. Dror. 2001. Determination of best focus and exposure dose using CD-SEM side-wall imaging. *SPIE 4344*, *Metrology, Inspection and Process Control for Microlithography XV*, 355–65.

Martz, H., and G. Albrecht. 2003. Nondestructive characterization technologies for metrology of micro/mesoscale assemblies. *Proceedings of Machines and Processes for Micro-scale and Meso-scale Fabrication, Metrology, and Assembly*, 131–141.

Meyyappan, A., M. Klos, and S. Muckenhirn. 2001. Foot (bottom corner) measurement of a structure with SPM. *SPIE 4344*, *Metrology, Inspection and Process Control for Microlithography XV*, 733–38.

Nichols, J. F., and Kurfess, T. R. 2004. Metrology of high aspect ratio MEMS. *J Microsyst Tech* 10:556–59.

O'Mahony, C., M. Hill, M. Brunet, R. Duane, and A. Mathewson. 2003. Characterization of micromechanical structures using white-light interferometry. *Measurement Sci Tech* 14:1807–14.

Ogura, I., and Y. Okazaki. 2003. A study of development of small-cmm probe detecting contact angle. *ASPE Summer Topical Meeting—Coordinate Measuring Machines, 349*.

Opsal, J., H. Chu, Y. Wen, Y. Chang, and G. Li. 2002. Fundamental solutions for real-time optical CD metrology. *SPIE* 4344, *Metrology, Inspection and Process Control for Microlithography XVI*, Santa Clara, CA, March 4–7, 2002, p. 163.

Peggs, G., A. Lewis, and R. Leach. 2003. Measuring in three dimensions at the mesoscopic scale. *Proceedings of Machines and Processes for Micro-scale and Meso-scale Fabrication, Metrology, and Assembly*, 53–57.

Peggs, G., A. Lewis, and S. Oldfield. 1999. Design for a compact high-accuracy CMM. *CIRP Annals—Manufacturing Technology* 48(1):417–20.

Postek, M. 1994. Scanning electron microscope metrology. *Proceedings of SPIE—The International Society for Optical Engineering* CR52:46–90.

Rizivi, S., and A. Meyyappan. 1999. Atomic force microscopy: A diagnostic tool (in) for mask making in the coming years. *SPIE* 3677, *Metrology, Inspection and Process Control for Microlithography XIII*.

Roth, J., E. Felkel, and P. Groot. 2003. Optical metrology. *Proceedings of Machines and Processes for Micro-scale and Meso-scale Fabrication, Metrology, and Assembly*, 87–92.

Schellekens, P., H. Haitjema, and W. Pril. 2001. A silicon-etched probe for 3-D coordinate measurements with an uncertainty below 0.1 µm. *IEEE Transactions on Instrumentation and Measurement* 50(6):1519–23.

Shilling, M. 2003. Two dimensional analysis of mesoscale parts using image processing techniques. Masters thesis, Georgia Institute of Technology, Atlanta, GA.

Shiozawa, H., Y. Fukutomi, T. Ushioda, and S. Yoshimura. 1998. Development of ultra-precision 3D-CMM based on 3-D metrology frame. *Proc ASPE* 18:15–18.

Simon, M., and C. Sauerwein. 2001. Quality control of light metal castings by 3D computed tomography. http://www.ndt.net/article/wcndt00/papers/idn730/idn730.htm.

St. Clair, L., and A. Mirza. 2000. Metrology for MEMS manufacturing. *Sensors Magazine* 17(7):24–33.

Storment, C., D. Borkholder, V. Westerlind, J. Suh, N. Maluf, and G. Kovacs. 1994. Flexible, dry-released process for aluminum electrostatic actuators. *J Microelectromech Syst* 3(3):90–96.

Svetkoff, D., and D. Kilgus. 1991. Influence of object structure on the accuracy of 3D systems for metrology. *SPIE 1614 Optics, Illumination, and Image Sensing for Maching Vision VI*, 218–30.

Takamasu, K., M. Fujiwara, A. Yamaguchi, M. Hiraki, and S. Ozono. 2001. Evaluation of thermal drift of nano-CMM. *Proc. of 2nd EUSPEN International Conference*, Turin, Italy, May 27–31.

Takamasu, K., M. Hiraki, K. Enami, and S. Ozono. 1999. Development of nano-CMM and parallel-CMM. *Proc. of Intl. Dimensional Metrology Workshop*, May 10–13.

Uhlmann, E., D. Oberschmidt, and G. Kunath-Fandri. 2003. 3D-analysis of microstructures with confocal laser scanning microscopy. *Proceedings of Machines and Processes for Micro-scale and Meso-scale Fabrication, Metrology, and Assembly*, 93–97.

Veeco. 2003. *Scanning probe/atomic force microscopy: Technology overview and update.* Santa Barbara, CA: Veeco Instruments, Inc.

VIEW Engineering. Voyager 6x12 specifications. http://www.vieweng.com, 2003.

Walch, K., A. Meyyappan, S. Muckenhirn, and J. Margail. 2001. Measurement of sidewall, line and line-edge roughness with scanning probe microscopy. *SPIE 4344, Metrology, Inspection and Process Control for Microlithography XV*, 726–32.

Wyant, J. 2002. White light interferometry. *Proceedings of SPIE—The International Society for Optical Engineering* 4737:98–107.

Zygo Corporation. 2003. *NewView 5000 specifications.* Middlefield, CT: Zygo Corporation.

References

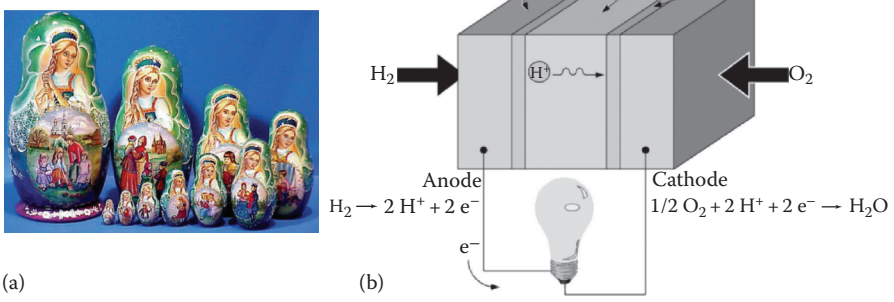
1. Quirk, M., and J. Serda. 2001. *Semiconductor manufacturing technology*. Upper Saddle River, NJ: Prentice Hall.
2. Nichols, J. F. 2004. *Metrology of high aspect ratio MEMS, PhD diss.*, Georgia Institute of Technology.
3. Ehmann, K., D. Bourell, M. L. Culpepper, T. J. Hodgson, T. R. Kurfess, M. Madou, K. Rajurkar, and R. DeVor. 2007. *Micromanufacturing: International research and development.* New York: Springer.
4. Denk, W., J. Strickler, and W. Webb. 1990. Two-photon laser scanning fluorescence microscopy. *Science* 248:73–76.
5. Tompkins, H. G., and E. A. Irene, eds. 2005. *Handbook of ellipsometry.* Berlin: Springer.
6. Goldstein, J., D. Newbury, D. Joy, C. Lyman, P. Echlin, E. Lifshin, L. Sawyer, and J. Michael. 2003. *Scanning electron microscopy and x-ray microanalysis.* New York: Kluwer Academic/Plenum Publishers.
7. Synge, E. H. 1928. A suggested method for extending microscopic resolution into the ultramicroscopic region. *Phil Mag* 6:356–62.
8. O'Keefe, J. A. 1956. Resolving power of visible light. *J Opt Soc Am* 46:359.
9. Ash, E. A., and G. Nicholls. 1972. Super-resolution aperture scanning microscope. *Nature* 237:510.
10. Pohl, D. W., W. Denk, and M. Lanz. 1984. Optical stethoscopy: Image recording with resolution lambda/20. *Appl Phys Lett* 44:651–53.
11. Harootunian, A., E. Betzig, M. Isaacson, and A. Lewis. 1986. Super-resolution fluorescence near-field scanning microscopy. *Appl Phys Lett* 49:674–76.
12. Vinci, R. P., and J. C. Braveman. 1991. Mechanical testing of thin films. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, June 24–27, 1991, pp. 943–48.
13. Bushan, B., ed. 1995. *Handbook of micro/nanotribology.* Boca Raton, FL: CRC Press.
14. Wong, S. M. 1978. Residual stress measurements on chromium films by x-ray diffraction using the sin2 Y method. *Thin Solid Films* 53:65–71.
15. Nishioka, T., Y. Shinoda, and Y. Ohmachi. 1985. Raman microprobe analysis of stress in Ge and GaAs/Ge on silicon dioxide-coated silicon substrates. *J Appl Phys* 57:276–81.
16. Marco, S., J. Samitier, O. Ruiz, J. R. Morante, J. Esteve-Tinto, and J. Bausells. 1991. Stress measurement of SiO_2-polycrystalline silicon structures for micromechanical devices by means of infrared spectroscopy technique. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, June 24–27, 1991, pp. 209–12.
17. Mehregany, M., R. T. Howe, and S. D. Senturia. 1987. Novel microstructures for the in situ measurement of mechanical properties of thin films. *J Appl Phys* 62:3579–84.
18. Jaccodine, R. J., and W. A. Schlegel. 1966. Measurements of strains at Si-SiO_2 interface. *J Appl Phys* 37:2429–34.
19. Bromley, E. I., J. N. Randall, D. C. Flanders, and R. W. Mountain. 1983. A technique for the determination of stress in thin films. *J Vac Sci Technol* B1:1364–66.
20. Allen, M. G., M. Mehregany, R. T. Howe, and S. D. Senturia. 1987. Microfabricated structures for the in situ measurement of residual stress, Young's modulus and ultimate strain of thin films. *Appl Phys Lett* 51:241–43.
21. Zhang, L. M., D. Uttamchandani, and B. Culshaw. 1991. Measurement of the mechanical properties of silicon microresonators. *Sensors Actuators* A A29:79–84.
22. Pratt, R. I., G. C. Johnson, R. T. Howe, and J. C. Chang. 1991. Micromechanical structures for thin film characterization. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, June 24–27, 1991, pp. 205–08.

23. Krulevitch, P. A. 1994. Micromechanical investigations of silicon and Ni–Ti–Cu thin films. PhD thesis, University of California, Berkeley.
24. Hoffman, R. W. 1976. In *Physics of nonmetallic thin films (NATO advanced study institutes series: Series B, physics)*, ed. C. H. S. Dupuy and A. A. Cachard, 273–353. New York: Plenum Press.
25. Singer, P. 1992. Film stress and how to measure it. *Semicond Int* 15:54–58.
26. Senturia, S. D., and R. T. Howe. 1990. Lecture notes. Boston: Massachusetts Institute of Technology.
27. Maseeh-Tehrani, F. 1990. Characterization of mechanical properties of microelectronic thin films. PhD thesis, Massachusetts Institute of Technology.
28. Maseeh, F., and S. D. Senturia. 1989. In *Polyimides: Materials, chemistry, and characterization*, ed. C. Feger, M. M. Khojasteh, and J. E. McGrath, 575–84. Amsterdam, the Netherlands: Elsevier Science Publishers B.V.
29. Senturia, S. 1987. Can we design microbotic devices without knowing the mechanical properties of materials? *Proceedings: IEEE Micro Robots and Teleoperators Workshop*. Hyannis, MA, November 1987, 3/1–5.
30. Biebl, M., and H. von Philipsborn. 1993. Fracture Strength of Doped and Undoped Panical Filters. PhD thesis, University of California, Berkeley.
31. Greek, S., F. Ericson, S. Johansson, and J.-Å. Schweitz. 1995. In situ tensile strength measurement of thick-film and thin film micromachined structures. *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, pp. 56–59.
32. Ericson, F., and J.-Å. Schweitz. 1990. Micromechanical fracture strength of silicon. *J Appl Phys* 68:5840–44.
33. Vinci, R. P., and J. C. Braveman. 1991. Mechanical testing of thin films. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, June 24–27, 1991, pp. 943–48.
34. Sekimoto, M., H. Yoshihara, and T. Ohkubo. 1982. Silicon nitride single-layer x-ray mask. *J Vac Sci Technol* 21:1017–21.
35. Guckel, H., T. Randazzo, and D. W. Burns. 1985. A simple technique for the determination of mechanical strain in thin films with applications to polysilicon. *J Appl Phys* 57:1671–75.
36. Lin, L. 1993. Selective encapsulations of MEMS: Micro channels, needles, resonators, and electromechanical filters. PhD thesis, University of California, Berkeley.
37. Guckel, H., D. W. Burns, C. C. G. Visser, H. A. C. Tilmans, and D. Deroo. 1988. Fine-grained polysilicon films with built-in tensile strain. *IEEE Trans Electron Devices* 35:800–01.
38. Kim, C. J. 1991. Silicon electromechanical microgrippers: Design, fabrication, and testing. PhD thesis, University of California, Berkeley.
39. Goosen, J. F. L., B. P. van Drieenhuizen, P. J. French, and R. F. Wolfenbuttel. 1993. In *7th International Conference on Solid-State Sensors and Actuators (Transducers '93)*. Yokohama, Japan, June 7–10, 1993, pp. 783–86.
40. Ericson, F., S. Greek, J. Soderkvist, and J.-Å. Schweitz. 1995. High sensitive internal film stress measurement by an improved micromachined indicator structure. *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, pp. 84–87.
41. Benitez, M. A., J. Esteve, M. S. Benrakkad, J. R. Morante, J. Samitier, and J. Å. Schweitz. 1995. Stress profile characterization and test structures analysis of single and double ion implanted LPCVD polycrystalline silicon. *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, pp. 88–91.
42. Biebl, M., G. Brandl, and R. T. Howe. 1995. Young's modulus of in situ phosphorus-doped polysilicon. *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, pp. 80–83.
43. Core, T. A., W. K. Tsang, and S. J. Sherman. 1993. Fabrication technology for an integrated surface-micromachined sensor. *Solid State Technol* 36:39–47.
44. Chu, W. H., M. Mehregany, X. Ning, and P. Pirouz. 1992. Measurement of residual stress-induced bending moment of p+ silicon films. *Mat Res Soc Symp* 239:169.
45. Fan, L.-S., R. S. Muller, W. Yun, J. Huang, and R. T. Howe. 1990. Spiral microstructures for the measurement of average strain gradients in thin films. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '90)*. Napa Valley, CA, February 11–14, 1990, pp. 177–81.
46. Campbell, D. S. 1970. In *Handbook of thin film technology*, eds. L. I. Maissel and R. Glang. New York: McGraw-Hill.
47. Bushan, B. 1996. In *Proceedings: IEEE Ninth Annual International Workshop on Micro Electro Mechanical Systems*, 91–98. San Diego, CA.
48. Bushan, B., Ed. 1995. *Handbook of micro/nanotribology.* Boca Raton, FL: CRC Press.
49. Howe, R. T., and R. S. Muller. 1983. Polycrystalline silicon micromechanical beams. *J Electrochem Soc* 130:1420–23.
50. Howe, R. T. 1985. In *Micromachining and micropackaging of transducers*, eds. C. D. Fung, P. W. Cheung, W. H. Ko, and D. G. Fleming, 169–87. New York: Elsevier.
51. Yun, W. 1992. A surface micromachined accelerometer with integrated CMOS detection circuitry. PhD thesis, University of California, Berkeley.
52. Adams, A. C. 1988. In *VLSI technology*, ed. S. M. Sze. New York: McGraw-Hill.
53. Maseeh, F., and S. D. Senturia. 1990. In *Technical Digest: 1990 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC, June 1990, pp. 55–60.
54. Senturia, S. D., and R. T. Howe. 1990. Lecture notes. Boston: Massachusetts Institute of Technology.
55. Cook, R. D. 1995. *Finite element modeling for stress analysis*. New York: Wiley.
56. Murphy, M. 1995. Personal communication.
57. Marshall, J. C., M. Parameswaran, M. E. Zaghloul, and M. Gaitan. 1992. High-level CAD melds micromachine devices with foundries. *IEEE Circuits Devices Mag* 8:10–17.
58. Tomovich, C., ed. 1988. *MOSIS user manual.* Los Angeles, CA: University of Southern California.
59. Billingsley, G. C. 1983. *Report no. UCB/ERL M83/62*. Berkeley, CA: University of California, Berkeley.
60. ERL. 1985. *SAMPLE version 1.6a user's guide.* Berkeley, CA: Electronics Research Laboratory, University of California, Berkeley.
61. Oldham, W. G., S. N. Nandgaonkar, A. R. Neureuther, and M. O'Toole. 1979. A general simulator for VLSI lithography and etching processes—Part I: Application to projection lithography. *IEEE Trans Electron Devices* ED-26:717–22.

62. Oldham, W. G., A. R. Neureuther, J. L. Reynolds, S. N. Nandgaonkar, and C. Sung. 1980. A general simulator for VLSI lithography and etching processes—Part 2: Application to deposition and etching. *IEEE Trans Electron Devices* ED-27:1455–59.
63. Koppelman, G. M. 1989. In *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '89)*, 88–93. Salt Lake City, UT.
64. Gilbert, J. R., G. K. Ananthasuresh, and S. D. Senturia. 1996. 3-D modeling and simulation of contact problems and hysteresis in coupled electromechanics. *Ninth Annual International Workshop on Micro Electro Mechanical Systems (MEMS '96)*. San Diego, CA, February 11–15, 1996, pp. 127–32.
65. Crary, S., and Y. Zhang. 1990. CAEMEMS: an integrated computer aided engineering workbench for micro electromechanical systems. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '90)*. Napa Valley, CA, pp. 113–14.
66. Korvink, J. G., J. Funk, M. Roos, G. Wachutka, and H. Baltes. 1994. SESES: a comprehensive MEMS modeling system. *IEEE International Workshop on Micro Electro Mechanical Systems (MEMS '94)*. Oiso, Japan, January 25–28, 1994, pp. 22–27.
67. Maseeh, F. 1994. A novel multidimensional semiconductor material analysis tool. *Solid State Technol* 37:83–84.
68. Sequin, C. H. 1992. Computer simulation of anisotropic crystal etching. *Sensors Actuators* A A34:225–41.
69. Koide, A., K. Sato, and S. Tanaka. 1991. Simulation of two-dimensional etch profile of silicon during orientation-dependent anisotropic etching. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '91)*. Nara, Japan, pp. 216–20.
70. Li, G., T. Hubbard, and E. K. Antonsson. 1998. SEGS: on-line etch simulator. *1998 International Conference on Modeling and Simulation of Microsystems.* Santa Clara, CA.
71. Hill, S. D. J., K. P. Kamper, U. Dasbach, J. Dopper, W. Ehrfeld, and M. Kaupert. 1995. In *Simulation and design of microsystems and microstructures,* eds. R. A. Adey, A. Lahrmann, and C. Lessmollmann, 276–83. Billerica, MA: Computational Mechanics Publications.
72. Fichtner, W. 1988. In *VLSI technology,* ed. S. M. Sze, 422–65. New York: McGraw-Hill.
73. Adey, R. A., A. Lahrmann, and C. Lessmollmann, eds. 1995. *Simulation and design of microsystems and microstructures*. Billerica, MA: Computational Mechanics Publications.

Part II

Scaling Laws, Actuators, and Power and Brains in Miniature Devices



(a) Nested Russian dolls as an example of isometric scaling. (b) Proton exchange membrane fuel cell.

> The civilization of abundance is also that of famines in African and other places. The collapse of totalitarian regimes has left intact all the evils of the democratic liberal societies, ruled by the demon of money. Marx's famous phrase about religion as the opiate of the masses can now be applied, and more accurately; to television, which will end up anaesthetizing the human race, sunk in an idiotic beatitude.
>
> **Octavio Paz in a passage on the science and technology of the West in *In Light of India***

> Napoleon asked: "Monsieur Laplace, why wasn't the Creator mentioned in your book on Celestial Mechanics?"
> Laplace replied: "Sir, I have no need for that hypothesis."

Introduction to Part II

Chapter 7 Scaling Laws
Chapter 8 Actuators
Chapter 9 Power and Brains in Miniature Devices

Introduction to Part II

Of the three chapters in Part II, the first deals with scaling laws, the second deals with actuators, and the third discusses surrounding power generation and the implementation of brains in miniaturized devices.

In Chapter 7, on scaling laws, we look at scaling from both intuitive and mathematical points of view. We are especially interested in deviations from linear scaling, where downscaling reveals new unexpected physical and chemical phenomena. The scaling chapter also constitutes an introduction to the subsequent treatises on actuators in Chapter 8 and on power and brains in Chapter 9. An actuator—like a sensor—is a device that converts energy from one form to another. In the case of an actuator, one is interested in the ensuing action; in the case of a sensor, one is interested in the information garnered. Scaling enables one to compare various actuator mechanisms, such as those proposed to propel a rotor blade, bend a thin Si membrane, or move fluids in fluidic channels. Power generation poses quite a challenge for miniaturization science; the smaller the power source, the less its total capacity. We will consider the powering of miniaturized equipment, as well as the miniaturization of power sources themselves, in Chapter 9. In the same chapter, we also discuss the different strategies for making micromachines smarter.

Part II should help the reader to choose a preferred actuator mechanism and/or power source through a better understanding of scaling and give some idea on how to incorporate intelligence in MEMS.

7

Scaling Laws

Our solar system as an example of isometric scaling.

Outline

Introduction

Microintuition

Mathematical Approach to Scaling

Scaling Examples

Breakdown of Scaling Laws

Appendix 7A: Common Dimensional Numbers

Questions

References

Introduction

The electronics revolution, which began in the 1960s, drove the miniaturization of radio, television, hard disk drives, camcorders, personal digital assistants, etc., to the point where miniaturization now appears commonplace. More recent is the miniaturization of components that are not based on integrated circuits in hardware such as physical, chemical, and biological sensors and analytical instruments. Progress in this newer area is derived from mature, traditional manufacturing technologies such as mechanical precision machining, as well as from lithography-based nontraditional microelectromechanical system (MEMS) and nanoelectromechanical system (NEMS) techniques. The encroaching limits of our planet's resources and the continuing deterioration of the environment are adding urgency to this trend of scaling things down in size.

In Table 7.1 we list sound reasons for miniaturizing systems, actuators, power sources, sensors, and components. In some cases, the reason is favorable scaling—for example, smaller devices tend to be faster and consume less power. Scaling may also disfavor miniaturization, as often happens with actuators and power sources; smaller actuators exert less force, and smaller power sources harness less power.

In this chapter, we analyze scaling laws and the size regimes where macrotheories start requiring corrections. We aspire to elucidate the unexpected behavior of some of these micromachines and to better understand why, in some cases, it makes sense to miniaturize a device for reasons beyond economics, volume, and weight. Following some comments about human microintuition, we introduce the concepts of

Table 7.1 Motivations to Miniaturize Systems, Actuators, Power Sources, Sensors, and Components

Miniaturization Attributes	Motivations
Low energy* and little material consumed	There are limited resources on planet Earth
Arrays of sensors	Redundancy, wider dynamic range, and increased selectivity through pattern recognition
Small	Smaller is lower in cost, minimal invasive
Favorable scaling laws (in some cases)	Forces that scale with a low power become more prominent in the microdomain; if these are positive attributes then miniaturization is favorable, e.g., surface tension becomes more important than gravity in narrower capillaries
Batch and beyond batch techniques (roll-to-roll manufacturing)	Lowers cost
Disposable	Helps avoid contamination
Breakdown of macrolaws in physics and chemistry	New physics and chemistry might be developed
Increased sensitivity (in some cases)	Nonlinear effects can increase a sensor's sensitivity, e.g., amperometric sensors
Smaller building blocks	The smaller the building blocks, the more sophisticated a system that can be built

* There is a significant debate about this point. If the total environmental "cost" for a product is considered, miniaturized products might be worse for the planet. It is similar to the debate about hybrid cars: they save on gas, but the environmental impacts of producing the batteries are often ignored.

isometry, allometry, and dimensional analysis—mathematical techniques that help us predict the influence of downscaling more reliably. The lessons learned are then applied to some scaling applications in both natural and human-made "machines." We end the chapter by introducing some examples where scaling laws break down and where noncontinuum models must be invoked. In a continuum, all quantities are defined everywhere and vary continuously from point to point; in a noncontinuum this continuity breaks down.

Microintuition

As children, we learned to deal with everyday objects having size and scale dimensions that are naturally "sized up" by our senses, and we were mystified and enthralled by anything much larger (Gulliver) or smaller (Tom Thumb) (see Figure 7.1). It is indeed much more difficult to get a feel for or understand the consequences of dimensions of physical systems many orders of magnitude larger or smaller than what we commonly experience. Of all our senses, vision informs our mind the most. The unaided eyes generally give us the ability to discern things as small in size as an ant's eye or the point of a pin (approximate size, 50 μm). However, telescopes and microscopes continue to improve and greatly extend our sensory capabilities in our quest to understand and utilize our material world better, including systems spanning extreme length scales from the astronomical to the atomic. These new sets of eyes have greatly assisted us in assessing the size and scale of diverse natural and artificial material systems. Today telescopes allow us to see galaxies 13 billion light-years away (10^{25} m), and microscopes allow us to see even single atoms (about 10^{-10} m)!

Q: Gulliver was twelve times as tall as the Lilliputians. How much should they feed him? Could it be twelve times their own food ration? Hint: A person's food needs are related to his or her mass (volume), which depends on the cube of the linear dimension.

A: Let l_G and V_G denote Gulliver's linear and volume dimensions. Let l_L and V_L denote a Lilliputian's linear and volume dimensions. Gulliver is twelve times as tall as the Lilliputians, or

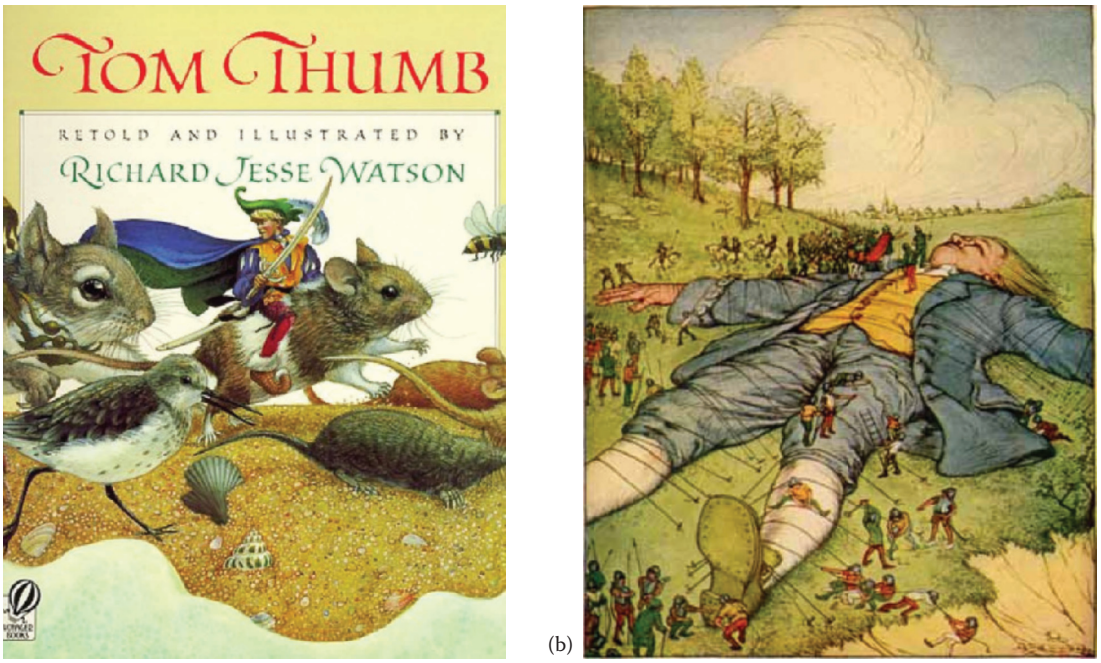


FIGURE 7.1 *Tom Thumb* originally by the Brothers Grimm (a) and *Gulliver's Travels* by Jonathan Swift (b).

$l_G = 12\ l_L$. Then, $V_G \propto (l_G)^3$ and $V_L \propto (l_L)^3$, so $V_G/V_L = (l_G)^3/(l_L)^3 = (12\ l_L)^3/(l_L)^3 = 12^3 = 1728$. Gulliver needs to be fed 1728 times as much food each day as a Lilliputian. This may be the real reason Gulliver was driven out of Lilliput. (Courtesy of W. Wilson, University of Central Oklahoma).

Humans are accustomed to thinking in distances and sizes—even microsizes have become quite intuitive, more so than microtimes. To accommodate the human inability to comprehend minute segments of time, Isaac Asimov introduced a measure based on the speed of light. Asimov's *light-meter* is the time required for light (in a vacuum) to cover the distance of a meter. An illustrative case is the half-life of radioactive particles. Saying that the half-life of a first particle is of the order of a hundredth of a millionth of a second and that of a second particle is of a thousandth of a trillionth of a second makes little impression. On the other hand, saying that the half-life of one type of particle is of the order of a light-meter and that of the other is of a tenth of a light-micrometer is more effective. We find linear extrapolation of length easy but are quickly at a loss when considering the implications that shrinking of time segments has or shrinking of surface-to-volume ratios on the relative strength of external forces working on miniaturized structures. In dealing with very small devices, our *microintuition*—as Bill Trimmer calls it—is often misleading (see also http://home.earthlink.net/~trimmerw/mems/tutorials.html).[1]

The effect of downscaling, especially into the nanodomain, is indeed not well understood and leads to some very surprising and sometimes counterintuitive results. For example, scaling shows that NEMS devices feature fundamental frequencies in the microwave range, active masses in the femtogram range, mechanical quality factors (Qs) in the tens of thousands, force sensitivities in the attonewton level, mass sensitivities at the level of individual atoms, and heat capacities far below yoctocalories. NEMS will make for electromechanical components with response times and operating frequencies that are as fast as most of today's electron devices. We now introduce a rigid mathematical formulation for scaling of a system, illustrating the newly gained insights with observations from nature. Our aim is to develop a systematic approach about the likely behavior of downsized systems so we do not need to rely on microintuition alone.

In Figure 7.2 we illustrate some known and not-so-well-known objects and their respective sizes,

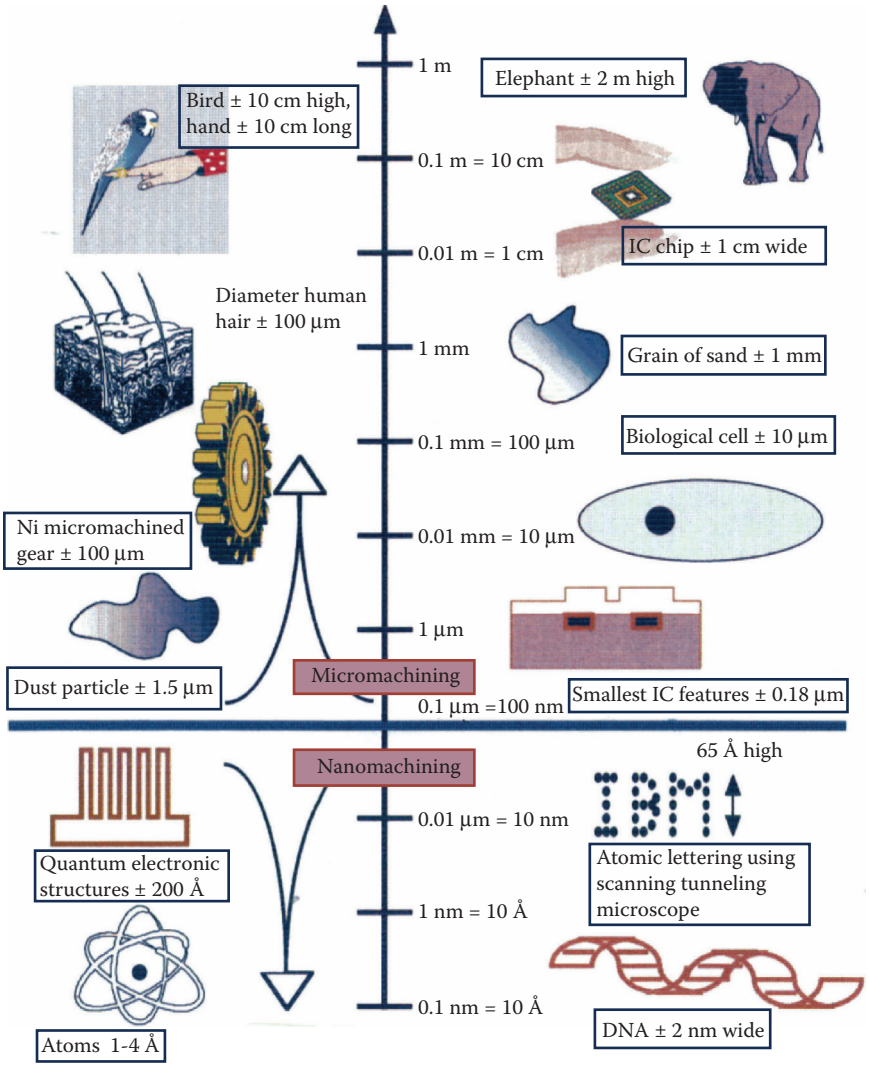


FIGURE 7.2 Various objects and their linear size.

including MEMS and NEMS, organized in powers of ten.

Mathematical Approach to Scaling

Introduction

In his book *On Growth and Form* (first published in 1917), D'Arcy Wentworth Thompson explains the profound effects of scaling in nature. Thompson (1860–1948) was a mathematically trained zoologist who pioneered the field we call *biomathematics* today (scaling laws in nature are a subset of biomathematics). He recognized many strong mathematical patterns in nature (spiral seashells, zebra stripes, etc.) and tried to find an explanation for the physical principles behind them. His book is the work of a true maverick and still inspires those interested in a mathematical understanding of biology. For a more recent account on growth and form in nature, including a more detailed mathematical treatment of scaling in animals, we refer to Knut Schmidt-Nielsen.[2] A popularizing and exquisitely illustrated work on the same topic is Thomas McMahon and John Bonner's *On Size and Life*.[3]

Similarity or Isometry, Allometry, Fractals, and Dimensional Analysis

A scaling law describes the variation of physical quantities with the size of the system. To understand scaling, it is essential to develop a notion for the concept of similarity or isometry. The idea is familiar from geometry: for example, two triangles are

said to be similar or isometric if all of their angles are equal, even if the sides of the two triangles are of different lengths. The two triangles have the same shape; the larger one is simply a scaled up version of the smaller one. This notion can be generalized to include complex organisms or human-made artifacts such as engines, boats, and aircraft.

In the most general case, to compare the relative sizes of two structures, x and y, one may write:

$$y = bx^a \qquad (7.1)$$

where a and b are constants. If the exponent a equals zero, then y and x are independent. This holds, for example, for the size of a cell (y) as a function of animal size (x); the cell diameter (b) is independent of animal size (about 10 µm, see Chapter 2). In case of a linear relationship between y and x, a equals 1, and x and y are said to be isometric or exhibit geometric similarity—the proportions of the structures that are being compared do not change with a change in size. Isometric change occurs when each dimension of a structure is scaled up or down regardless of the disproportionate change in such parameters as surface area and volume. If x doubles in length, then so does y. The proportions remain constant and do not change with a change in size. Besides the aforementioned triangles, the nine planets and the ten nested Russian dolls shown above are all simple examples of isometry, in which the proportions do not change with size.

From the rules of isometry, we expect that for geometrically similar animals of the same mass density, the body mass, m, will be proportional to the cube of the body length, l^3, and that the area, S, will be proportional to l^2. In a number of species, body surface areas proportional to the 2/3 power of body mass are just what is found. This is illustrated in Figure 7.3 where we have plotted the surface area versus body mass for a species of salamander in a log-log plot. To elucidate the plot in Figure 7.3[4], we first rewrite Equation 7.1 in its logarithmic form and obtain the following.

$$\log y = \log b + a \log x \qquad (7.2)$$

In the example case the slope a equals 2/3, or 0.67, and fits the isometric prediction exactly. The value of b is of significance as well; it fixes the value for y when x is equal to 1. Comparisons made over a limited size range, say within a given species of plants or animals and types of bridges or boats, are often isometric.

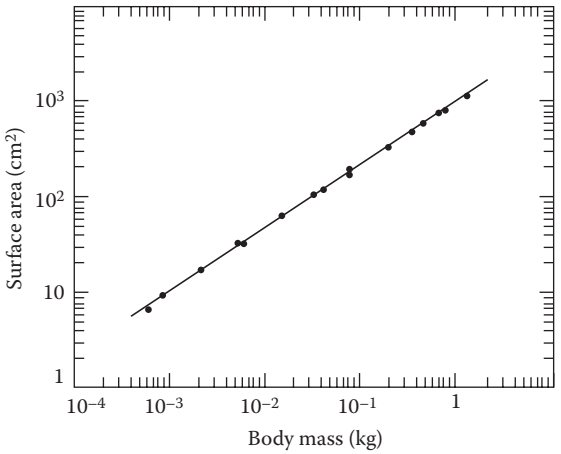


FIGURE 7.3 Surface area versus body mass for a species of salamander, the dwarf siren (*Pseudobranchus striatus*). The surface area, *S*, is proportional to the body mass, *m*, raised to the 0.67 power.

When, on the other hand, comparing animals, plants, engines, or boats of greatly differing size, proportions do often change with size, and one speaks about allometry. Allometric change occurs when the proportions of an organism adjust to the different rates at which surface area, volume, and other physical parameters change with differing sizes. It is sometimes referred to as *differential scaling*. Perhaps the most famous allometric relationship pertains to the metabolic machinery of animals. In 1932, Max Kleiber, an American veterinary scientist, found that the basal metabolic rate (the rate at which the body generates heat when the organism is at rest), as measured by the rate of oxygen consumption or heat production, is proportional to $m^{0.75}$ over a very wide range of animals. In Figure 7.4, we show what has become known as *Kleiber's law*, which says that the metabolic rate for a wide range of different species is proportional to $m^{0.75}$ (the broken line with a slope of 0.67 = 2/3 has been added for comparison). This law has been shown to be approximately true over twenty orders of magnitude (from single cells to elephants). There are a large variety of other phenomena that change with body size according to quarter-power law scaling. For example, a cat,

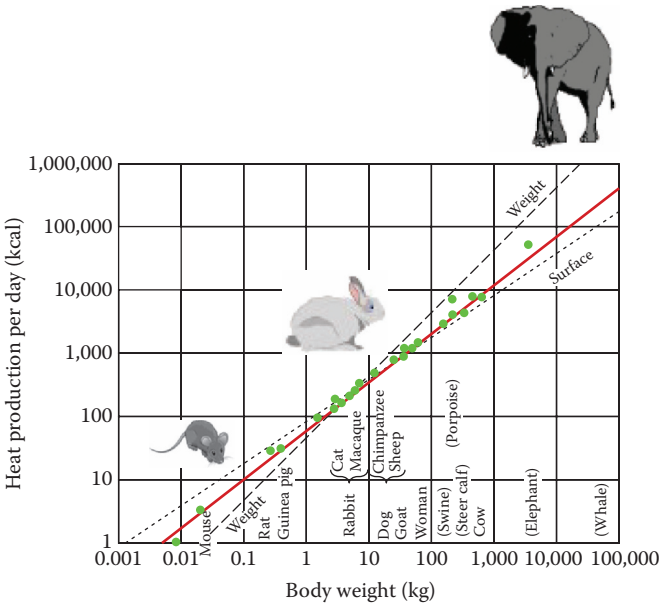


FIGURE 7.4 A log-log plot of metabolic heat production against body weight, taken from Kleiber's 1947 article. It is clear that most points (green) closely fit the red line with a slope of three quarters, rather than the weight line (slope of 1) or the surface line (slope of two thirds). Note that although the differences between the lines seem small, this is a log-log plot, so the deviations are actually vast.

100 times the weight of a mouse lives about $100^{1/4}$, or about three times, as long as a mouse, and heartbeat scales as $m^{-1/4}$, so the cat's heart beats one-third as fast as a mouse's.

The reason for Kleiber's allometric relation has eluded scientists for a long time. For metabolism, molecules inside and outside cells, such as carbohydrates, enzymes, hormones, nucleic acids, must interact and react with each other. If a cell and its organelles were not limited to a size of only a few micrometers in diameter, then it would be difficult for the various molecules to diffuse and come in contact within a reasonable amount of time. The small cell size maximizes surface-to-volume ratio (S/V), allowing for the most efficient metabolism. In larger organisms, individual cells are accessed via distribution networks: in plants, a vascular system of xylem and phloem carries water and nutrients to and from the different parts of the organism, and in animals, the circulatory system of veins plays the same role. If S/V was the main factor involved in controlling metabolic rate, that rate would be expected to scale to the body mass of an organism to the two-thirds power. Plotting the basal rate of oxygen consumption rate within a given species is indeed often found to be proportional to $m^{0.67}$. On this basis, for a cat, 100 times more massive than a mouse, we expect to have a basal metabolism of $100^{0.67}$, or about 21.5 times that of the mouse. But Kleiber's law, which applies when one looks across different species and over a wide range of sizes, means that a cat's metabolic rate is not 21.5 times greater than a mouse's, but 31.6 times. One of the more inspiring explanations for Kleiber's law came from ecologists Enquist and Brown and physicist West in 1997. These authors explain that the generic principle underlying Kleiber's law is that nutrient supply networks in animals and plants form a branching fractal network* to reach all cells in an organism. They showed that they could derive Kleiber's 3/4-power scaling law by invoking these fractal networks. In their analysis, then, the metabolism scales as the dimension of the circulatory system, which is after all what delivers the energy being metabolized.† Allometric changes in proportions are often compensations for competing requirements of surface area, cross-sectional area, and volume. In this case, nature evidently evolved a method to overcome the natural limitation of simple geometric scaling by developing these more efficient fractal-like webs to connect all the similarly sized cells within an organism. Fractals are an optimal geometry for minimizing the work lost due to the transfer network while maximizing the effective surface area (see Figure 7.5, with a tree as an example of a fractal). We believe that this insight might be of great value in human engineering endeavors where reagents and reaction products must be distributed with maximum efficiency throughout the bulk of a device, be it a fuel cell, a battery or a sensor. In electrochemical systems, it is advantageous to have a large S/V when there is the need to transfer the signal or power effectively to an electrical network.

* In a fractal network, each tiny part is a replica of the whole. In the current context, think about blood capillaries in our bodies and their botanical equivalent in trees: a silhouette of the human circulatory system and of roots and branches of a tree actually look remarkably similar.

† The theory is not without is detractors; Dodds at MIT, for example, disputes the very existence of the 3/4 law.[5] A 2005 paper by West and coworkers seems to vindicate this team's work.[6]

FIGURE 7.5 A tree as an example of a fractal.

We will demonstrate in Chapter 9, this volume, in the section on miniaturizing power sources, that fractal geometries for the electrodes might be the best solution for maximizing the current while minimizing the internal resistance.

Having reviewed the basics of isometry and allometry, we now introduce the concept of dimensional analysis, one of the simplest yet most powerful tools in a physicist's bag of tricks. Dimensional analysis is important when modeling physical phenomena, for instance, predicting the performance of a large plane on the basis of observing a small-scale model in a wind tunnel or the reverse—more relevant in miniaturization science—predicting the behavior of a miniaturized machine on the basis of the performance of a larger model. Besides predicting the behavior in scale-up or scale-down experiments, dimensional analysis is of use in the deduction of physics laws (physics demands dimensional consistency in equations), reduction of number of variables by introducing dimensionless groups and order of magnitude analysis (which terms dominate in an equation). Langhaar described dimensional analysis as "a method by which we deduce information about a phenomenon from the single premise that the phenomenon can be described by a dimensionally correct equation among certain variables."[7]

The dimensional specification for a physical variable is its *dimensional formula*. The dimensional formula shows the fundamental or base quantities (denoted by capital letters) raised to the appropriate powers and enclosed in brackets. There are seven fundamental base quantities in these dimensional formulas: time, length, mass, electrical current, temperature, amount of a substance in moles, and luminous intensity. All other quantities, dimensions, and units are known as *derived quantities*. For example, the dimensional formula for acceleration, a, is $[LT^{-2}]$. In the simplest use, dimensional analysis helps one decide whether an equation is correct as the dimensions on both sides of the equality sign must be the same and one can figure out the dimensions of a particular variable if the others are known.

In a more advanced use of dimensional analysis, the object is to substitute a set of dimensionless numbers for the dimensional physical variables that describe the problem—pressure, velocity, density, and so on. As the dimensionless numbers are products or ratios of the physical variables, this process always succeeds in reducing the number of variables in the problem. This reduction in the number of variables is a very worthwhile exercise since it frequently saves a great deal of experimental effort.

We introduce first the two most important theorems of dimensional analysis and follow that by listing a number of detailed steps to carry out a typical dimensional analysis using the exponent method or Buckingham method. Finally, we present a couple of examples and an exercise that will turn out to be of use later on in this chapter.

First Theorem

The dimensional homogeneity theorem states that any physical quantity is dimensionally a power law monomial: $[\text{Any Physical Quantity}] = [M^a L^b T^c]$.

Second Theorem

The Buckingham Π theorem (Pi theorem) states that if a system has k physical quantities of relevance that depend on r independent dimensions, then there are a total of $k - r$ mutually independent dimensionless groups $\pi_1, \pi_2, \ldots, \pi_{k-r}$. The dimensionless groups are called *pi-terms*. The behavior of the system is then describable by a dimensionless equation $F(\pi_1, \pi_2, \ldots, \pi_{k-r}) = 0$. The number of pi-terms is thus fewer than the original number of variables by r.

In practice, one proceeds as follows:

1. Make an informed guess about which parameters (variables), k, are likely to be important

in the phenomenon under scrutiny. These are typically all variables required to describe the problem geometrically (e.g., diameter, D), materials properties [e.g., fluid properties (ρ,μ)], physical constants (e.g., gravity, g—would things work differently on the moon?), and external effects (dp/dx, a pressure differential).
2. Express each variable in terms of its fundamental quantities (dimensions), [M], [L], [T], etc.
3. Determine the required number of dimensionless parameters ($k - r$).
4. Select a number of repeating variables equal to the number of dimensions (= r). All fundamental quantities must be included in this set, and each repeating variable must be dimensionless independent of the others.
5. Form a dimensionless parameter π by multiplying one of the nonrepeating variables by the product of the repeating variables, each raised to an unknown exponent.
6. Solve for the unknown exponents.
7. Repeat this process for each nonrepeating variable.
8. Express results as a relationship among the dimensionless parameters: $F(\pi_1, \pi_2, \pi_3, \ldots) = 0$.

Example 7.1

We apply the Buckingham method now to some example problems, starting with a simple one. Say you want to find out the drag force F_D on a body immersed in a flowing fluid (say a fish or a submarine moving underwater—let's call it a swimmer). The drag force F_D is the force opposing the object's velocity and will depend on the following:

$$F_D = f(l, v, \rho, \mu) \tag{7.3}$$

where k = 5 (number of variables/parameters) and r = 3 number of dimensions (L, T, and M).

Thus, the number of dimensionless numbers is expected to be 5 – 3 = 2. This is summarized in Table 7.2.

Select "repeating" variables: l, v, and ρ. Combine these with the rest of the variables: μ and F_D. So the first pi-term becomes $\pi_1 = \mu(l^a v^b \rho^c)$.

Table 7.2 There Are Five Parameters or Variables (k) and Three Fundamental Quantities

F_D	l	v	ρ	μ
MLT^{-2}	L	LT^{-1}	ML^{-3}	$ML^{-1}T^{-1}$

To calculate the exponents, one writes the following equation:

$$\pi_1 = \mu(l^a v^b \rho^c) \tag{7.4}$$

and since this combination is supposed to be dimensionless:

$$M^0L^0T^0 = (ML^{-1}T^{-1})(L)^a(LT^{-1})^b(ML^{-3})^c$$

$$M: 0 = 1 + c \rightarrow c = -1$$

$$L: 0 = -1 + a + b - 3c \rightarrow a = -1$$

$$T: 0 = -1 - b \rightarrow b = -1$$

The first π term is thus as follows:

$$\pi_1 = \mu/lv\rho,\ \pi_1 = 1/\text{Reynolds number (1/Re)} \tag{7.5}$$

usually written as follows:

$$Re = \rho vl/\mu \tag{6.71}$$

The Reynolds number, familiar from Volume I, Chapter 6, appears in a large number of fluid flow problems. It is the ratio of inertial forces over viscous forces. A more complete discussion of the Reynolds number is given in Volume I, Chapter 6, on fluidics. A list of other common dimensionless groups is represented in Appendix 7A. The second dimensionless group (pi-term) should be chosen in such a way that it is certain to be independent of the first pi-term. One way to do that is to make sure that at least one physical parameter that was present in the first group is missing from the second and vice versa. In the example case, we replace the viscosity μ by the drag force F_D, as follows:

$$\pi_2 = F_D(\rho^d l^e v^f) \tag{7.6}$$

$$M^0L^0T^0 = (MLT^{-2})(ML^{-3})^d(L)^e(LT^{-1})^f$$

$$M: 0 = 1 + d \rightarrow d = -1$$

$$L: 0 = 1 - 3d + e + f \rightarrow e = -2$$

$$T: 0 = -2 - f \rightarrow f = -2$$

This way we obtain a second dimensionless group, given as follows.

$$\pi_2 = F_D/\rho\, l^2 v^2 \tag{7.7}$$

This second group, $F_D/\rho\, l^2 v^2$, may be considered a dimensionless drag force. Using A as the wetted surface of the submerged swimmer, this dimensionless group may be rewritten as the drag coefficient C_D. Since $\pi_2 = f(\pi_1)$ we can also write the following.

$$C_D = F_D/\rho l^2 v^2 = f(\text{Re}) \tag{7.8}$$

For a given Reynolds number, the best design for a race car, submarine, or airplane is the one that minimizes drag coefficient C_D (see below under "Scaling in Flying").

Example 7.2

In a second example, we apply dimensional analysis to the drag force on a rough sphere in a flowing solution. Drag force on a rough sphere (with diameter D and roughness k) is a function of D, ρ, μ, v, and k. The relative roughness of the sphere is given as k/D:

$$k = 6 \text{ (number of variables/parameters)}$$

$$r = 3 \text{ number of dimensions } (L, T, \text{ and } M)$$

Thus, the number of dimensionless numbers is expected to be $6 - 3 = 3$. This is summarized in Table 7.3.

Select "repeating" variables: D, v, and ρ. Combine these with the rest of the variables: F_D, μ, and k. So the first pi-term is given as follows:

$$\pi_1 = \mu(D^a v^b \rho^c) \tag{7.9}$$

and since this combination is dimensionless:

$$M^0L^0T^0 = (ML^{-1}T^{-1})(L)^a(LT^{-1})^b(ML^{-3})^c$$

$$M: 0 = 1 + c \rightarrow c = -1$$

Table 7.3 There Are Six Parameters or Variables (k) and Three Fundamental Quantities

F_D	D	v	k	ρ	μ
MLT^{-2}	L	LT^{-1}	L	ML^{-3}	$ML^{-1}T^{-1}$

$$L: 0 = -1 + a + b - 3c \rightarrow a = -1$$

$$T: 0 = -1 - b \rightarrow b = -1$$

we again obtain $\pi_1 = \mu/lv\rho$, $\pi_1 = 1/\text{Reynolds number}$, abbreviated as $\text{Re} = \rho vl/\eta$ (Volume I, Equation 6.71). For the second dimensionless group we choose the repeating variables D, v, and ρ and combine those with the nonrepeating variables F_D, μ, and k.

$$\pi_2 = k(D^a v^b \rho^c) \tag{7.10}$$

$$M^0L^0T^0 = (L)(L)^a(LT^{-1})^b(ML^{-3})^c$$

$$M: 0 = c \rightarrow c = 0$$

$$L: 0 = 1 + a + b - 3c \rightarrow a = -1$$

$$T: 0 = -b \rightarrow b = 0$$

$$\pi_2 \equiv k/D \tag{7.11}$$

For the third pi-term, we write a dimensionless group as follows:

$$\pi_3 = F_D(D^a v^b \rho^c) \tag{7.12}$$

$$M^0L^0T^0 = (MLT^{-2})(L)^a(LT^{-1})^b(ML^{-3})^c$$

$$M: 0 = c + 1 \rightarrow c = -1$$

$$L: 0 = 1 + a + b - 3c \rightarrow a = -2$$

$$T: 0 = -2 - b \rightarrow b = -2$$

$$\pi_3 \equiv F_D/D^2 v^2 \rho \tag{7.13}$$

And since $\pi_3 \equiv f(\pi_2, \pi_1)$, we can also write the following equation:

$$F_D/D^2 v^2 \rho = f(\rho vl/\mu, k/d) \tag{7.14}$$

We encountered this expression when discussing the friction factor for pipe flow in Volume I, Chapter 6 (see Volume I, Equation 6.73). With the math introduced so far, we can express similitude or similarity a bit better now than the way we represented it in Equation 7.1. As we noted before, similarity is very important to be able to predict the behavior of a model with respect to the real life object. In miniaturization science the model is usually very large compared with the actual prototype device. Suppose the physical problem described in terms of dimensionless parameters is given as $\pi_1 = f(\pi_2, \pi_3, \ldots, \pi_n)$; applied

to the prototype, this transforms into $\pi_{1p} = f(\pi_{2p},\pi_{3p},\ldots,\pi_{np})$, and for the model of the prototype it reads as $\pi_{1m} = f(\pi_{2m},\pi_{3m},\ldots,\pi_{nm})$. The similarity requirements for modeling then become the following:

$$\begin{aligned} \pi_{2m} &\equiv \pi_{2p} \\ \pi_{3m} &\equiv \pi_{3p} \\ &\ldots\ldots \\ \pi_{nm} &\equiv \pi_{np} \\ \Rightarrow \text{then } \pi_{1m} &\equiv \pi_{1p} \end{aligned} \qquad (7.15)$$

In other words, under these conditions the dependent variables for prototype and model will be the same. Similarity comes in different flavors. There is, first of all, geometric similarity (scale factor); i.e., the ratio of lengths in the scale model compared with the full scale case is constant (all angles are identical). Then there is kinematic similarity, in which the ratios of velocity, acceleration, and flow rate between the scale model and the full scale are constant (time-dependent behavior of the model). Finally, there is dynamic similarity, where the ratios of forces on scale model and full-scale case are constant.

To illustrate these distinctive similarities, consider predicting the drag on a thin rectangular plate ($w \times h$) from a model placed normal to a flow. The drag function in this case will be $F_D = f(w, h, \mu, \rho, \text{and } v)$. Dimensional analysis shows that $\pi_1 = f(\pi_2, \pi_3)$, and this applies to both model and prototype. Working out the problem leads to an expression very similar to the one in Equation 7.14 (drag on rough sphere instead of on a plate):

$$F_D/ w^2v^2\rho = f(\rho vl/\mu, w/h) \qquad (7.16)$$

We can design a model to predict the drag on the prototype, since the drag on the model will be $F_{Dm}/w_m^2v_m^2\rho_m = f(\rho_m v_m l_m/\mu_m, w_m/h_m)$ as $\pi_{1m} = f(\pi_{2m}, \pi_{3m})$, and for the prototype we will have $F_{Dp}/w_p^2v_p^2\rho_p = f(\rho_p v_p l_p/\mu_p, w_p/h_p)$ as $\pi_{1p} = f(\pi_{2p}, \pi_{3p})$. The geometric similarity conditions give us the size of the model, with $\pi_{2m} \equiv \pi_{2p}$, $w_m/h_m = w_p/h_p$. For the kinematic similarity, we look at the π_3 terms ($\pi_{3m} = \pi_{3p}$), namely $\rho_m v_m l_m/\mu_m = \rho_p v_p l_p/\mu_p$, or the velocity of the model is as follows:

$$v_m = \frac{\mu_m \rho_p l_p}{\mu_p \rho_m l_m} v_p \qquad (7.17)$$

For the dynamic similarity—the drag force—on the prototype, we calculate from $\pi_{1m} \equiv \pi_{1p}$.

$$F_{Dp} = \left(\frac{w_p}{w_m}\right)^2 \frac{\rho_p}{\rho_m} \left(\frac{v_p}{v_m}\right)^2 F_{Dm} \qquad (7.18)$$

Exercise 7.1

A model of a submarine is 1/20th scale and is moving in fresh water (20°C) at a speed of 2 m/s. Find the speed of the model and the ratio of the drag force on the model to that of the real submarine moving in seawater. The density of seawater ρ is 1015 kg/m^3, and $\nu = \mu/\rho = 1.4 \times 10^{-6}$ m^2/s.

From the Reynolds number ($Re_m = Re_p$), we can calculate the speed of the model in fresh water (kinematic similarity).

$$\frac{v_m l_m}{\nu_m} = \frac{v_p l_p}{\nu_p} \text{ or } v_m = \frac{\nu_m l_p}{\nu_p l_m} v_p = \frac{1}{1.4} \times \frac{20}{1} \times 2\text{m/s}$$

$$v_m = 28.6 \text{ m/s}$$

To calculate the ratio of the drag forces we rely on the dynamic similarity.

$$\frac{F_m}{\rho_m v_m^2 l_m^2} = \frac{F_p}{\rho_p v_p^2 l_p^2} \text{ or}$$

$$\frac{F_m}{F_p} = \frac{\rho_m v_m^2 l_m^2}{\rho_p v_p^2 l_p^2} = \frac{1000}{1015}\left(\frac{28.6}{2}\right)^2\left(\frac{1}{20}\right)^2$$

$$\frac{F_m}{F_p} = 0.504$$

Trimmer's Vertical Bracket Notation

From the preceding section, we appreciate that as the scale of structures decreases, so does the importance of phenomena that vary with the largest power of the linear dimension l. Phenomena that are more weakly dependent on size dominate in small dimensions: electrostatics (l^2), friction (l^2), surface tension (l), diffusion ($l^{1/2}$), and van der Waals forces ($l^{1/4}$). High-power forces include gravity (l^3) (the mass of a system, m, scales as l^3), inertia (l^3), magnetism (l^2, l^3, or l^4, depending on the exact configuration), flow (l^4), and thermal emission

(l^2 to l^4). Scaling laws may be positive or negative order. Positive-order laws imply that the property grows with increasing scale, and negative-order laws imply that the property grows with decreasing scale. A zero-order law implies that a property is invariant with scale. In other words, positive orders affect large objects and negative orders small objects.

Trimmer introduced an elegant method to express different scaling laws by using a vertical bracket notation. For different possible forces, he writes:

$$\begin{bmatrix} l^1 \\ l^2 \\ l^3 \\ . \\ . \\ l^n \end{bmatrix} = F \tag{7.19}$$

The top element in this notation refers to the case where the force scales as l^1. The next one down refers to a case where the force scales as l^2, etc. If the system becomes one-tenth its original size, all the dimensions decrease by a tenth. The mass of a system, m, scales as l^3, and as systems become smaller, the scaling of the force also determines the acceleration a, transit time t, and the amount of power per unit volume (PV^{-1}). For a generalized case with a force F scaling as l^F, we obtain

$$a = \frac{F}{m} = \left[l^F\right]\left[l^{-3}\right] \tag{7.20}$$

for acceleration a,

$$t = \sqrt{\frac{2xm}{F}} = \left(\left[l^1\right]\left[l^3\right]\left[l^{-F}\right]\right)^{\frac{1}{2}} \tag{7.21}$$

for transit time t, with x = distance, and

$$\frac{P}{V} = \frac{Fx}{tV} \tag{7.22}$$

for power per unit volume. Applying Equation 7.20 to Equation 7.22 to calculate a, t, and PV^{-1} for forces scaling with varying power of the linear dimension, we obtain the following:

$$\begin{bmatrix} l^1 \\ l^2 \\ l^3 \\ . \\ . \\ l^n \end{bmatrix} = F \rightarrow \begin{bmatrix} l^{-2} \\ l^{-1} \\ l^0 \\ . \\ . \\ l^{n-3} \end{bmatrix} = a \rightarrow \begin{bmatrix} l^{1.5} \\ l^1 \\ l^{0.5} \\ . \\ . \\ l^{\sqrt{4-n}} \end{bmatrix} = t \rightarrow \begin{bmatrix} l^{-2.5} \\ l^{-1} \\ l^{0.5} \\ . \\ . \\ l^{n-2-\sqrt{4-n}} \end{bmatrix} = \frac{P}{V} \tag{7.23}$$

We can best appreciate the usage of the vertical bracket representation from an example. In the case of electrostatic actuation, $l^F = l^2$, and from Equation 7.23 one deduces $a = l^{-1}$, $t = l^1$, and for the power density $PV^{-1} = l^{-1}$ (an increase of a factor of 10).

For force laws with a power higher than l^2, the power generated per volume degrades as the scale decreases. Even in a case with $F = 1^4$, the time required to perform a task remains constant when the system is scaled down. This is an observation we understand intuitively: small things tend to be quick. We should keep in mind that beneficial effects are easily overshadowed by loss mechanisms, which scale in the same way or become even more important in the microdomain.

In Table 7.4 we list various physical phenomena and their scaling. The highest-order law is that for the moment of inertia (l^5); mass moment of inertia becomes rapidly unimportant for small-scale systems, and consequently small motors, both electric and combustion powered, are able to reach top speed much faster than large motors.

Scaling Examples

The Importance of the Surface-to-Volume Ratio

If a system is reduced isomorphically in size (i.e., scaled down with all dimensions of the system decreased uniformly), the changes in length, area, and volume ratios alter the relative influence of various physical effects that determine the overall operation in unexpected ways. For geometrically similar (isometric) objects of the same mass density, body mass (m), is proportional to the cube of the body length, l^3, and the surface area is proportional to l^2 or the surface-to-volume ratio (S/V) equals $l^{2/3}$, i.e., $l^{0.67}$.

Table 7.4 Scaling of a Selected Set of Physical Phenomena

Physical Quantity	Scaling Exponent of l	Units
Area	2	m^2
Bending stiffness	1	$N\ m^{-1}$
Buoyant force	3	N
Capacitance	1	F
Capacitor electric field	–1	Vm^{-1}
Deformation	1	m
Drag and lift forces	2 + 2n (n is fluid relative velocity)	N
Electrostatic energy	3	J
Electrostatic force	2	N
Frictional force	2	N
Heat capacity	3	$J\ K^{-1}$
Inductance	1	H
Magnetic force	4	N
Mass (m)	3	kg
Mass moment of inertia	5	$Kg{\cdot}m^2$
Ohmic current	2	A
Resistance	–1	Ω
Resistive power loss	1	$A^2\ \Omega$
Shear stiffness	1	$N\ m^{-1}$
Strength	2	$N\ m^{-2}$
Strength-to weight ratio	–1	m
Surface tension force	1	N
Thermal conductance	1	$W\ K^{-1}$
Thermal time constant	2	s
Viscous forces	1 + ν (ν is fluid relative velocity)	N
Voltage	1	V
Volume (V)	3	m^3

As isometric objects or organisms shrink in size, the S/V increases. The surface area of a smaller structure is relatively larger compared to its volume than that of a larger object. There are many examples in biology and daily experience to choose from to illustrate the pervasive importance of increasing S/V with decreasing size. The abundance curve of types of animals as a function of their size as shown in Figure 7.6 illustrates the importance of a change in S/V as a function of size. For warm-blooded mammals there are significant problems associated with being too small. The very small pygmy or dwarf shrew (about 4 cm long), for example, must eat continuously or freeze to death. The heat loss from a living creature is roughly proportional to its surface area (l^2), and the rate of compensating heat generation through eating is roughly proportional to its volume (l^3). As animals get smaller, a greater percentage of their energy intake is required to balance the heat loss. A warm-blooded animal smaller than the dwarf shrew or the Cuban bee hummingbird (about 5 cm) becomes improbable; it could neither obtain nor digest the food required to maintain its constant body temperature. Insects circumvent the problem of heat loss by being cold-blooded.

On the large end of the animal size spectrum, there are no land animals larger than the African elephant at 3.80 m. Land animals fighting the resistance of gravity can grow only so big before becoming too clumsy and inefficient. Large animals also find few niches on earth that can accommodate them, and as a consequence there are fewer large animals than small ones. For creatures of the sea, the same physical barrier of gravity to size does not

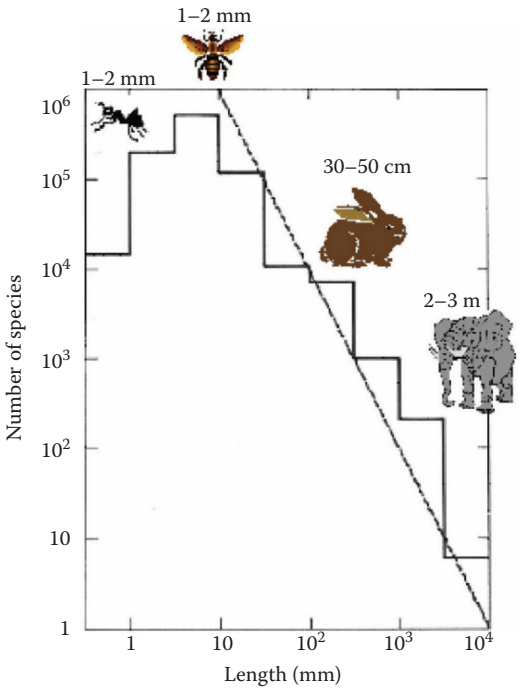


FIGURE 7.6 An abundance of types of animals in various size categories.

exist. The largest sea creature in volume is a blue whale, a little more than 20 m (60 ft.) long, much larger than the elephant. In terms of their length, the range of living things stretches over about 8 orders of magnitude, from the minute prokaryotic bacterium-like *Mycoplasma* (0.3 µm) to the massive blue whale (20 m).

At the small end of the size spectrum, smaller birds and animals are quicker and more agile, and insects are most plentiful. Scaling laws and the many niches in nature where they can hide and survive make insects perhaps the most successful animal species. Morowitz postulates that scaling laws must impose a lower limit for life in a dry environment even for cold-blooded animals, explaining why the abundance curve exhibits a peak.[8] Smaller organisms, because of their very large S/V, cannot retain their vital fluids long enough to survive, and no independent living organism (we exclude viruses) is smaller than the smallest bacteria at 0.2 µm. Water-based life increases its range of sizes both above and below that of terrestrial animals by evading gravity in the large-size range and drying out in the small-size range.

Faster evaporation in dry environments associated with larger surface-to-volume ratios also has important consequences for the retention of liquids in small biosensors, the evaporation of ejected drops in drop delivery equipment, the aqueous cocktails in miniaturized polymerase chain reaction (PCR) chambers, etc. In all these, evaporation quickly becomes a problem with decreasing drop size. Solutions include adding a hygroscopic material; mixing with solvents with lower vapor pressure (e.g., glycerol); topping of the solution with a low-vapor-pressure, nonmixing liquid; and working in a solvent-saturated environment.

Scaling in Surface Swimming

Earlier we mentioned the work on scaling in nature by Knut Schmidt-Nielsen, McMahon and Bonner, and Vogel in the context of Biomimetics (see also Chapter 2). In the same context, Hayashi presented an analysis of the natural laws that govern the relationship between speed and body length in water, on land, and in air.[9] Hayashi also analyzed the effect that miniaturization has on the strength of materials, various surface phenomena, and manufacturing accuracy. The latter is an aspect we covered in Volume II, Chapter 6, on mechanical energy-based removing. Here we will briefly look at scaling in surface swimming followed by an introduction to scaling in flight.

The resistance fish and other swimmers must overcome moving on the surface of a body of water is not gravity (l^3) but "skin friction," which increases only as the square of the linear dimensions (l^2). In this case, a larger size leads to a distinct advantage in that the larger the surface of the swimmer, the greater its swimming speed. This can be understood as follows. The available energy (E) for swimming speed (v) depends on the mass of the creature's muscles or engines (l^3), while its motion through water is opposed by the skin-friction resistance, R (l^2). This leads to $E \sim R \times v^2$ or $v \sim \sqrt{1}$; in other words, the bigger fish or the bigger ship moves faster, but only in the ratio of the square root of the increasing length. The latter is captured in the Froude number. The Froude number Fr is the ratio of inertial to gravitational forces. William Froude (1810–1879), British naval architect and engineer, realized that the factor that most limits the speed of ships or swimmers is the energy involved in the production

of waves by the swimmer and by noting that wave speed increases with the square root of the wavelength. Surface swimmers find it hard to go faster than the waves they create, and by making waves of longer length, the bigger swimmer makes waves that travel faster. Wave production depends on two forces, inertia and gravity. The inertia of the moving water keeps it wavy, and gravity tries to flatten the surface; the interaction of these two phenomena sets the speed of the wave. Surface swimmers thus vary in how they produce waves on the basis of their length, so replace the mass in Equation 7.24 by l^3 times density. Inertial acceleration is hard to measure, so we replace it by velocity divided by time and obtain Equation 7.25. Time in Equation 7.25 can be gotten rid of by noting that length over time is velocity, and the Froude number is obtained in Equation 7.26.

$$\mathrm{Fr} = \frac{ma}{mg} \tag{7.24}$$

$$\mathrm{Fr} = \frac{\rho l^3 v t^{-1}}{\rho l^3 g} \tag{7.25}$$

$$\mathrm{Fr} = \frac{\rho l^3 v t^{-1}}{\rho l^3 g} = \frac{\rho l^2 v^2}{\rho l^3 g} = \frac{v^2}{gl} \tag{7.26}$$

Scaling in Flying

Aircraft and ocean liners operate in fluid environments with a range of Reynolds numbers in the millions to hundreds of millions because the viscous forces are small compared with inertial forces (see Volume I, Equation 6.71). At the other end of the spectrum are small insects, birds, and fish, whose viscous forces are of the same order of magnitude as the inertial forces. Obviously, depending on size, different strategies will be required for both flying and swimming. Although the same fluid mechanical rules govern motion through air and water, flying, as Icarus could attest (see the inset heading Chapter 2, this volume), is considerably harder than swimming. The lower density of air helps a flying machine go forward faster (thrust), but it necessitates a larger force to keep it aloft. For an object to hover immobile in air takes an upward force that counterbalances weight, while forward thrust requires a force equal to the drag force (F_D) at its specific flying speed. Weight depends on volume, while drag depends on area. Thus the smaller creature finds weight less troublesome, but drag becomes a real "drag"—in other words, for a tiny insect, staying aloft against gravity's pull is easy but making headway against the drag of the air is tougher than for a bird or a plane.

The drag coefficient C_D, a dimensionless number, is given by $C_D = F_D/\rho\, l^2 v^2 = f\,(\mathrm{Re})$ (Equation 7.8), with surface area A replacing l^2 and a factor 1/2 added:

$$C_D = \frac{F_D}{\frac{1}{2}\rho v^2 A} \tag{7.27}$$

where ρ = density of the fluid
v = free stream velocity
A = characteristic area of the body*
F_D = drag force

The drag coefficient, C_D, is the parameter that best describes the resistance to motion. As we gleaned from Equation 7.8, the drag coefficient varies with the Reynolds number, but the coefficient also depends on the angle of attack (AOA). This is the angle that the flying object penetrates the air with. It is measured in degrees. In the case of an airfoil (Figure 7.7), at zero degrees AOA, the front of the airfoil is pointed directly into the stream. At a 90° AOA, it is pointed straight up.

We learned in Volume I, Chapter 6, on fluidics, that C_D also depends on surface roughness; the drag coefficient decreases with increasing roughness as it delays the boundary layer separation. In Figure 7.8A we show the historical reduction in drag on cars. The C_D of a Porsche 924 Turbo, for example, is 0.34, while that for a Toyota Tercel (liftback) is 0.54. In Figure 7.8B we also show the drag coefficient of some different familiar shapes.

Lift force, F_L, is the force normal to the direction of the object's velocity, v. Lift force, like drag force, depends on surface area and goes up with the object's

* A = frontal area (thick body, cylinder, cars, projectiles) = plan area (wide/flat body, wings, hydrofoils) = wetted area (surface ships, barges). This is a simplification, as both frictional and pressure drag forces should be considered, and factor A has a different meaning in both. For a more rigorous treatment, see Granger.[6]

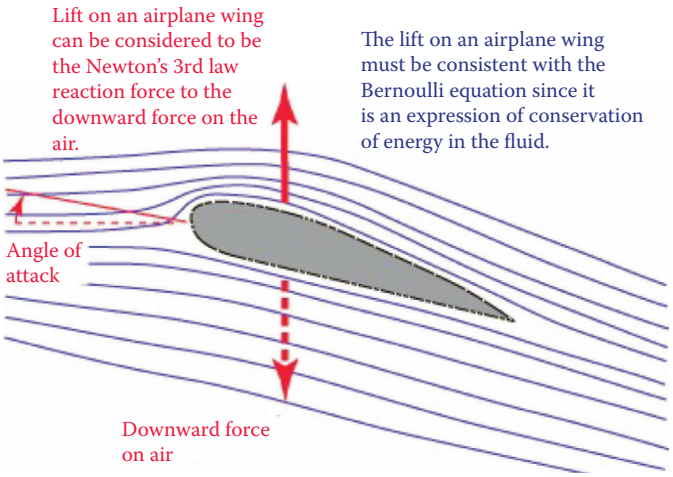


FIGURE 7.7 Angle of attack of an airfoil. For Bernoulli equation, see Volume I, Chapter 6, on fluidics (Equation 6.44). Notice that the Bernoulli interpretation of lift is not entirely correct (see below).

speed through the air. Many of us were taught that airplanes fly as a result of the Bernoulli's principle, which says that if air speeds up, pressure is lowered (Volume I, Equation 6.44). From the shape of the wing shown in Figure 7.7 the path over the top is longer than over the bottom, and in the Bernoulli model a wing generates lift because the air goes faster over the top, creating a region of low pressure, and thus lift. This begs the question of how inverted flight is possible, and naturally one may also ask why the air goes faster over the top of the wing than over the bottom. The usual explanation is that when the air separates at the leading edge of the wing, the part that goes over the top must converge at the trailing edge with the part that goes under the bottom. This is the so-called *principle of equal transit times* (http://www.allstar.fiu.edu/AERO/airflylvl3.htm, http://www.grc.nasa.gov/www/K-12/airplane/

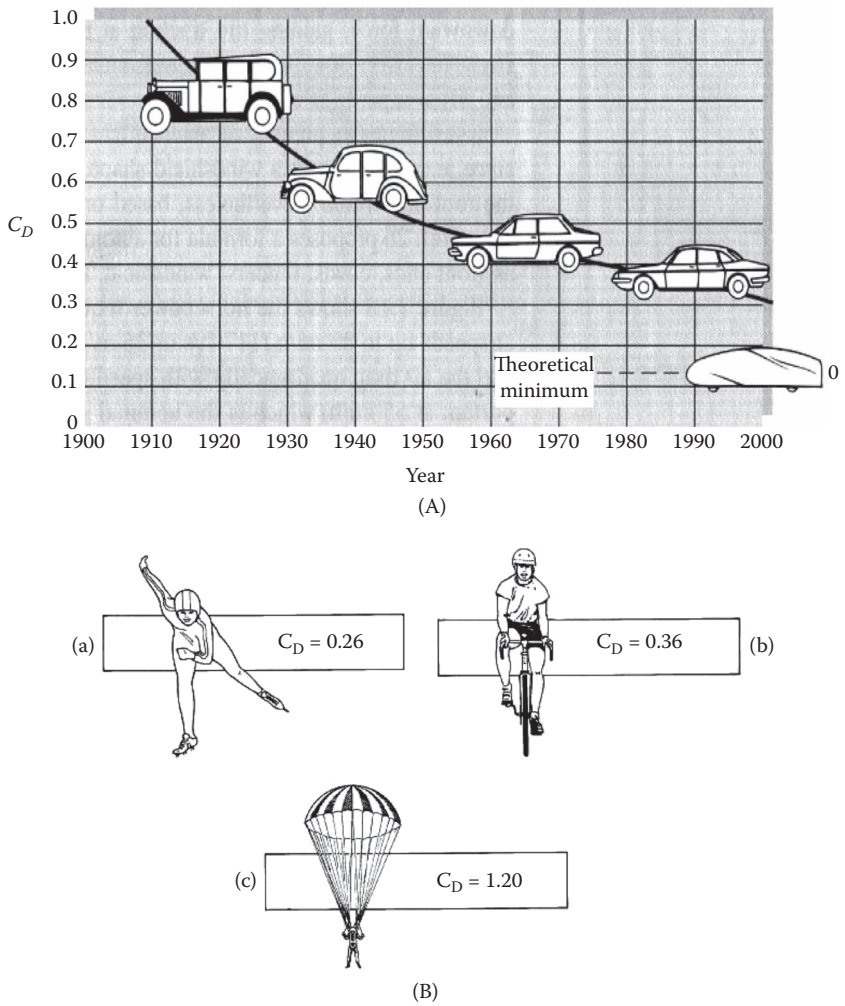


FIGURE 7.8 (A) Historical reduction of drag on cars. (B) Drag on a variety of shapes: (a) skater; (b) cyclist; (c) parachutist.

lift1.html). But in a wind tunnel simulation, where colored smoke is introduced, one sees that the air that goes over the top of the wing gets to the trailing edge before the air that goes under the wing. In fact, closer inspection reveals that the air going under the wing is slowed down from the "free-stream" velocity of the air. A calculation of the average speeds of the air over and under the wing can be made, and from the Bernoulli equation the resulting pressure forces and thus lift can be determined. The problem is that this calculation shows that in order to generate the required lift for a typical small airplane, the distance over the top of the wing must be about 50% longer than under the bottom; this would make for a very odd wing shape. An airplane wing more typically has a top surface that is only 1.5–2.5% longer than the bottom surface. So a Cessna 172 would have to fly at over 400 mph to generate enough lift. Clearly, something in this description of lift is wrong. As Newton's laws suggests, the wing must change something of the air to get lift, and that something, as revealed from wind tunnel tests, is that air passes over the wing and is bent down. The bending of the air is the action, and the reaction is the lift on the wing. The natural question now becomes, how does the wing divert the air down? When a moving fluid, such as air or water, comes into contact with a curved surface, it tries to follow that surface; this is known as the *Coanda effect*. The reason behind the Coanda effect is viscosity; viscosity in air is small, but it is enough for the air molecules to want to stick to the wing surface. Because the fluid near the surface has a change in velocity, the fluid flow is bent toward the surface. When the air is bent around the top of the wing, it pulls on the air above it, accelerating that air down; otherwise there would be voids left in the air above the wing. This pulling causes the pressure to become lower above the wing. It is the acceleration of the air above the wing in the downward direction that gives lift. The lift of a wing is equal to the change in momentum of the air it is diverting down.

The dimensionless lift coefficient C_L is given by the following equation:

$$C_L = \frac{F_L}{\frac{1}{2}\rho v^2 A} \tag{7.28}$$

Both lift and drag coefficients are a function of the Reynolds number, AOA, geometric shape (i.e., the thickness distribution), and speed. The lift and drag forces can be changed by changing (most easily) the AOA or velocity (v). The surface area (A) can also be changed by changing the shape or configuration of, say, the wing flaps during flight. As in swimming, larger objects mean faster flight: a fruit fly might hit 3 mph, a bumblebee can do about 12 mph, and a Boeing 747 can do 920 kph.

Doubling the length of a wing gives a flying machine four times the lift (and drag), but it also increases the weight by a factor of 8. The best-designed wings produce a lot of lift while suffering little drag. The largest flying animals alive today are the soaring birds such as the 10-kg wandering albatross, with a wingspan of 3.5 meters. Lift relative to drag gets worse as wings get smaller, and gliding becomes impossible with wing lengths of less than 1 mm.[10] As a point of reference, ordinary airplanes have a lift/drag ratio of 10–100, whereas that of a micromachined airplane fabricated from Si by Kubo et al. (1.56 mm long and with a mass of 10.8 mg) was measured at 0.4.[11] The low lift/drag ratio of the latter is due to the influence of viscosity, which gets more pernicious the smaller the flying object. When a glider descends, the angle depends entirely on the ratio of lift to drag. The larger this ratio, the more horizontal the glide. Kubo et al.'s micromachined airplane had an AOA less than 45°, so it could not glide. Small flying animals use fast-beating wings to produce both lift and thrust. Their wings not only beat up and down but to some extent also fore and aft. The drag force produced by down beating must exceed that produced by up beating.

Scaling in Surface Tension and Adhesion

Inertia and gravity forces rule the world of large organisms and machines, while the forces of attraction between molecules rule the world of minute organisms. A large and lumbering elephant cares little about viscosity and even less about surface tension and diffusion, but insects are similar in size to many human-made miniature machines, and viscosity, surface tension, and diffusion play an important role for both.

Capillary forces, as we saw above, are caused by surface tension and provide a striking example of scaling laws. The mass of a liquid in a capillary tube, and hence the weight, scales as l^3 and decreases more rapidly than the surface tension, which scales as l, as the system becomes smaller. That is why it is more difficult to empty liquids from a capillary than to spill coffee from a cup.

One easily recognizes how insects take advantage of surface tension and several other scaling phenomena: some insects can walk on water (e.g., a water strider), and most jump farther in proportion to their size than humans can. Walking on water is based on the surface tension at the water surface (l^1), which easily supports an insect's weight (l^3). A 10-mg, mosquito-sized insect needs a mere millimeter of total foot edge to be supported by surface tension; a 60-kg human would need 8000 m of foot edge. The effort for a jump is proportional to the mass (m) and the height (h) to which that weight is raised. In other words, $E \sim m \times h$. The biological force in a muscle available for this work is proportional to the mass of the muscle or to the mass of the animal. It follows that h is, or tends to be, a constant. As a consequence, animals tend to jump to the same actual height independent of their size. Insects have other advantages: to circumvent problems with excessive heat loss, insects are cold-blooded. Having adapted so well to so many niches in nature, it is no wonder insects are so abundant (Figure 7.6).

As we saw in Volume I, Chapter 7, molecular adhesive forces increase in proportion to the surface l^2. The strength of the attraction force between two solid surfaces drops off very quickly, though, as the distance between the two solids increases (see Volume I, Equation 7.26 and Table 7.8). The larger the bodies, the smaller the chance they come in close contact, and intimate contact is greatly aided by small size. Bonner suggested that adhesive forces, just like surface tension, scale as l^1. In Figure 7.9, following Bonner,[4] we plot the forces due to mass (l^3) and molecular cohesion (l^1) (both surface tension and adhesion) as a function of length in a log-log plot. The vertical axis measures force in arbitrary units. The lines for l^3 and l^1 are made to cross arbitrarily at 1 mm, based on the empirical observation that structures of that length have to deal in

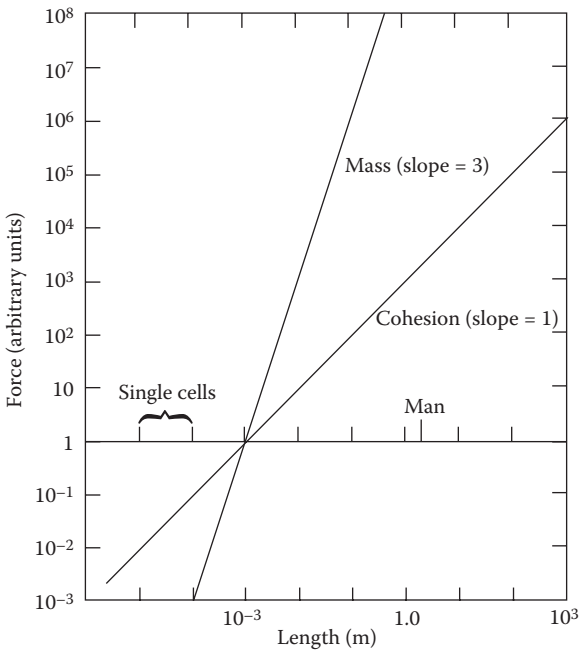


FIGURE 7.9 Log-log plot of forces due to mass (l^3) and molecular cohesion (l^2) (both surface tension and adhesion) as a function of length. The vertical axis measures force in arbitrary units. The lines for l^3 and l^1 are made to cross arbitrarily at 1 mm.[4]

equal amounts with the impact of gravity and cohesion forces. This graph helps explain how creatures below a certain size, can walk on walls and ceilings. Protuberances on the feet of these creatures function to enhance the molecular adhesion. From Figure 7.9, we see that small insects and other wall-walking animals must operate below a critical size at which the line proportional to body weight intersects the line proportional to cohesive force.

Kellar Autumn et al. confirmed, using a MEMS force sensor, that in the case of a gecko, the gecko setae (microhairs) are responsible for the remarkable van der Waals-based dry adhesion to walls.[12] This effect is almost exclusively based on the size and shape of the gecko's foot hair or setae and is quite independent of surface chemistry. Gecko foot hair is made of five microdiameter seta fibers, and each of these fibers is equipped with between 100 and 1000 nanofibers (spatulae), which are 200 nm in diameter (see Figure 7.10 for a picture of a gecko, a gecko foot, setae, and spatulae). In order for this type of dry and clean adhesion (the hairs are hydrophobic in nature) to work, only a small preload force normal to the surface is required to force the compliant

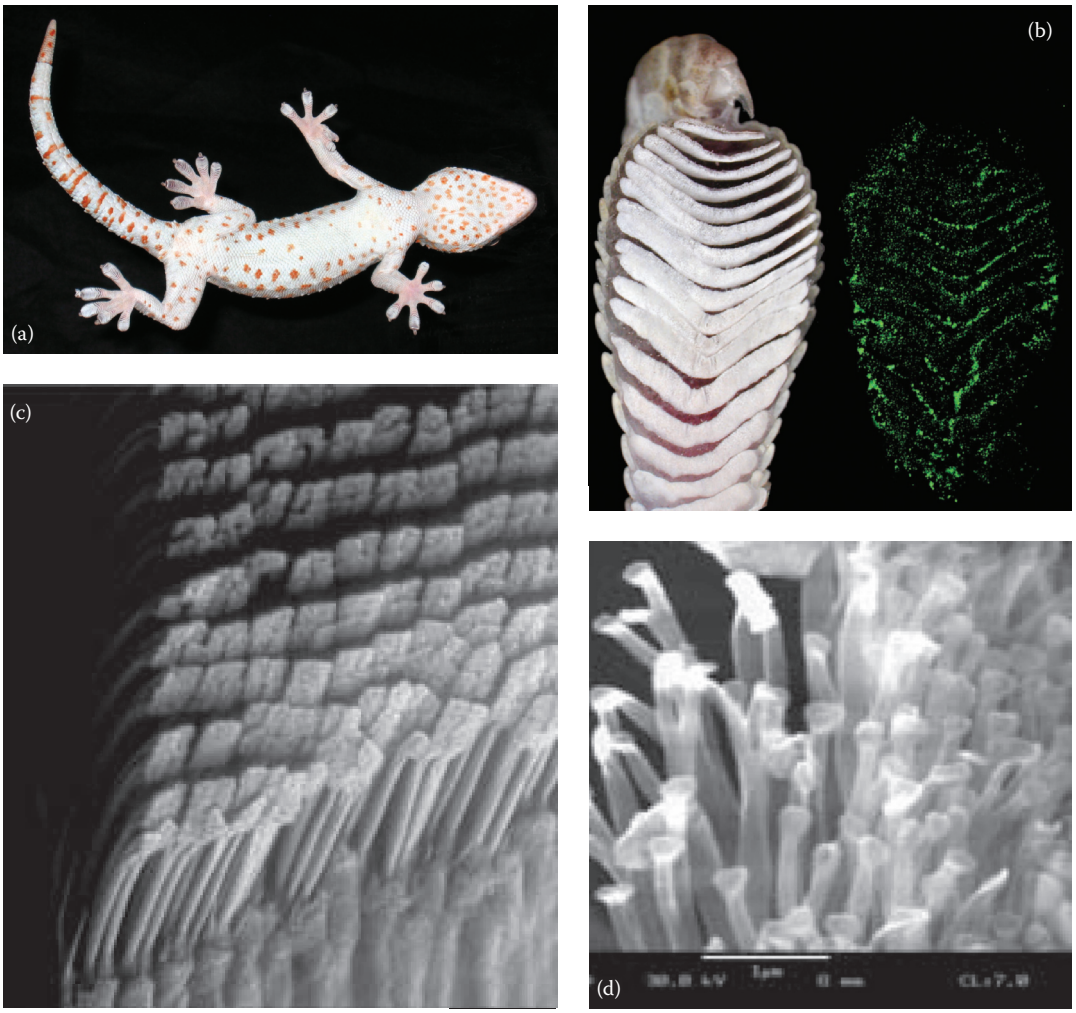


FIGURE 7.10 A gecko (a), gecko foot (b), setae (c), and spatulae (d). (Courtesy of Dr. K. Autumn, http://www.lclark.edu/~autumn/PNAS.)

hairs to configure themselves properly to a rough or smooth surface. The adhesion force can be as high as 10 N per 1 cm^2 area, and the foot will remain on until peeled off. Menon et al. have been working on synthetic gecko hair to enable gecko inspired climbing robots.[13] For an update on gecko adhesion, visit http://www.lclark.edu/~autumn/PNAS.

Scaling and Diffusion

Diffusion plays a particularly significant role in the lives of very small organisms. Diffusion is the consequence of the endless and random wandering of every molecule of every gas or liquid at temperatures above 0 K. Diffusional effects come into play at yet a smaller length scale than surface tension. For a spherical molecule, the Stokes-Einstein diffusion coefficient, D (m^2/s), we saw in Volume I, Chapter 6, is given by (Volume I, Equation 6.29) the following:

$$D = \frac{k_B T}{6\pi\eta r} \qquad \text{(Volume I, Equation 6.29)}$$

where k_B = Boltzmann constant ($1.38 \times 10\, m^{-23} J\ K^{-1}$)
T = absolute temperature (K)
η = absolute viscosity (kg/m·s)
r = hydrodynamic radius

A small molecule, with a molecular weight of between 500 and 1000, will have a diffusion coefficient of about 5×10^{-5} m^2/s. According to the random walk equation, the diffusion length, x, of a molecule in solution is given by the Einstein-Smoluchowski expression (Volume I, Equation 6.100):

$$x = \sqrt{2D\tau} \qquad \text{(Volume I, Equation 6.100)}$$

Table 7.5 A Drop of Saline Solution in a Small Bucket

Volume	1 μL	1 nL (10^{-9} L)	1 pL (10^{-12} L)	1 fL (10^{-15} L)	1 aL (10^{-18} L)
Length of each cube side	1 mm	100 μm	10 μm	1 μm	100 nm
Time to diffuse over a cube side	500 s	5 s	0.050 s	0.5 ms	0.05 ms
Number of molecules in a 1 μM solution	6×10^{11}	6×10^{8}	6×10^{5}	600	6

Sources: Manz, A., J. Harrison, E. Verpoorte, J. Fettinger, A. Paulus, H. Lüdi, and H. Widmer. 1992. Planar chips technology for miniaturization and integration of separation techniques into monitoring systems: Capillary electrophoresis on a chip. *J Chromatogr* 593:253–58;[14] and Brown, P. R., and E. Grushka, eds. 1993. *Advances in chromatography.* New York: Marcel Dekker.[15]

where τ is the time required for a molecule to diffuse over a distance *x*. From Volume I, Equation 6.100, diffusion of a molecule in the bulk of a liquid over a length of 10 μm is a million times faster than diffusion over 1 cm. Table 7.5[14,15] illustrates how the volume of a drop of liquid (say, of a saline solution, in liters), relates to the linear dimension of a cube containing it. Assuming a diffusion constant *D* of 10^{-5} cm^2s^{-1}, the table also lists the time it takes for a molecule to diffuse across one side of the containment volume; it will take a molecule about 500 s to diffuse over a distance of 1 mm but only 0.5 ms to cross a distance of 1 μm. Finally, the table lists the total number of molecules the various volumes filled with a 1 μM solution contain.

As we saw in Volume I, Chapter 6, on fluidics, mixing, although only mediated by diffusion, is very fast at the microlevel and should allow for reaction times to be determined by inherent kinetics rather than the time it takes for reactant species to meet in solution. In nature, only the smallest animals rely on diffusion for transport; animals made up of more than a few cells cannot rely on diffusion anymore to move materials within themselves. They augment transport with hearts, blood vessels, pumped lungs, digestive tubes, etc. These distribution networks typically constitute fractals. Fractals are an optimal geometry for minimizing the work lost due to the transfer network while maximizing the effective surface area. Along this line, one can envision making arrays of parallel microchemical reactors. Mixing small amounts of fluid in a large set of these parallel microreactors leads to a much higher mixing and reaction efficiency than when mixing the same total amount of reagents in one big reactor vessel all at once. Based on the latter, it is possible to design chemical reactors where scaling up is perfectly linear, a feat not possible when scaling up chemical reactors in the traditional mode from milliliters to hundreds of liters in big tanks.

Mixing of liquids in small microfluidic manifolds is addressed in Volume I, Chapter 6, on fluidics.

Scaling of Strength-to-Weight Ratio and Inertia

From the previous section, we recognize the important role of gravity for larger systems. Gravity, for example, does not allow large creatures to support themselves with surface tension on the surface of water. It makes large objects fall faster than small ones and necessitates that larger aircraft fly faster to stay aloft. To counteract gravity, large land mammals have stiffer bones than small animals and require thicker legs. The legs of the elephant appear as rather straight, stubby columns, while those of an insect are long and spindly. The diameter of a tall homogeneous body such as a tree must grow as the power 3/2 of its height, which accounts for Goethe's *Es ist dafür gesorgt, dass die Bäume nicht in den Himmel wachsen** in *Dichtung und Wahrheit*.[16] Human design reflects the same influence. If we build two geometrically similar bridges, the larger will be the weaker of the two, and it will be so in the ratio of their linear dimension, *l*. The strength of the iron girders in the two bridges varies with the square of the linear dimension, that is, l^2, but the weight of the whole structure varies with the cube of its linear dimension,

* Translation: It has been arranged for trees not to grow all the way into heaven.

that is, l^3. Scaling up the entire bridge by a factor of two gives us columns that are four times as strong but must also bear eight times the load—the safety factor is halved. The girders, being thicker, also suffer greater self-loading, which requires them to be thicker yet. To engineer around gravity, stiffer materials and alternative designs are required. The larger the structure is, the more severe the strain. By reducing the size of a device, the structural stiffness generally increases relative to inertia-imposed loads.

The strength-to-weight ratio scales as l^{-1} (area over weight, i.e., l^2/l^3), and as a consequence small things are relatively stronger. Consider the strength comparison between a human and an ant. The human body is 300 times that of an ant, and a human can carry approximately one body weight. The strength-to-weight ratio law predicts that a 1/300th scale human Lilliputian could carry 300 times its body weight. Although not exactly right, an ant can carry objects 10–50 times (but not 300 times) its own weight. The discrepancy is caused by the difference between a scaled human form to that of an ant. Because of the increased strength-to-weight ratio, an ant is proportionately more slender than a human; in other words, an ant scaled to human size would have legs far too thin to support its own weight.

Inertia is the tendency of both fluids and solids to remain at rest or to keep moving. The force involved in starting a movement or stopping it equals the mass of the object times the acceleration or deceleration. The highest-order scaling laws are those involving moments of inertia, that is:

$$I = \int r^2 dm \tag{7.29}$$

with l^5. Small things are easier to start and stop. An example of this is small motors, electric or combustion, which are able to reach top speed in a fraction of a second; large motors may require several seconds to reach full speed.

An important bottleneck in the development of microactuators in general is that resistive forces, such as viscous drag (l^2) and surface tension (l), exceed motive forces [mass (l^3) × surface (l^2) = (l^5)] in the microdomain. Miniaturization engineers look for driving forces that scale more advantageously, such as electrostastics (l^2) and capillary force (l).

The above scaling examples constitute but a few illustrations; many more scaling examples can be found in Chapter 8, this volume, on actuators; Volume I, Chapter 6, on fluidics; Chapter 9, this volume, on power and brains in miniature devices; and throughout the book.

Breakdown of Scaling Laws

Introduction

Scaling laws break down when the continuum theory for the phenomenon under study no longer applies. The latter occurs any time boundary conditions for the macroscale case are violated. Many examples of breakdown of scaling laws can be found throughout the book. A detailed example was worked out in Volume I, Chapter 7, where we saw that electrochemical detectors with electrodes of the size of the diffusion boundary layer thickness display noncontinuum behavior. Here we give as an additional example, namely that of electrostatic devices with electrode spacing of the size of the mean free path of the surrounding gas molecules.

Scaling in Electrostatics

Boundaries and Continuum Theory in Electrostatics

Electrostatic charges, we will see in Chapter 8, on actuators, arise from buildup or deficit of free electrons in a material. An electrically charged material can exert an attractive force on oppositely charged objects or a repulsive force on similarly charged ones. To appreciate scaling issues in electrostatics, we follow Trimmer's analysis of isometric scaling of the maximum stored energy in a simple parallel plate capacitor. The ability of a capacitor, C, to store charge is its capacitance, C, and the capacitance is related to the applied potential V as $C = Q/V$. From Gauss's law (Volume I, Equation 5.37) applied to two plates of cross-sectional area A, containing a dielectric (and hence charge density Q/A), one obtains the following:

$$E = \frac{Q/A}{\varepsilon_r \varepsilon_0} \tag{7.30}$$

and since C = Q/V and E is given by V/d, where d is the distance between the two plates of the capacitor, Equation 7.30 can be rewritten as follows.

$$C = \frac{\varepsilon_r \varepsilon_0 A}{d} \tag{7.31}$$

From Equation 7.31, it is clear that the greater the relative permittivity ε_r, of the material between the capacitor plates, the greater the capacitance. This can be understood as follows, since a dielectric is an insulator all a gap material can do in response to a charge on the capacitor plates is to accumulate a countercharge. The greater the relative permittivity, the greater the effect of this charges redistribution; thus the greater the effective "stored charge." Based on Equation 7.30, the insertion of a dielectric reduces the electric field inside the dielectric by a factor ε_r.

The dimensions of the capacitor in Figure 7.11 are w, v, and d, and they all scale with l. We first calculate the incremental work dW to add charge dQ to this capacitor at voltage V:

$$dW = V(Q) \cdot dQ = \left(\frac{Q}{C}\right) \cdot dQ \tag{7.32}$$

The total work W to charge to Q and thus the energy stored in the capacitor is then given as follows:

$$U = W = \frac{1}{C}\int_0^Q Q\, dQ = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}CV^2 \tag{7.33}$$

Above a critical voltage, V_b, the electrostatic forces polarizing the molecules in the gap are so strong that electrons are torn free and charge flows and the maximum electrostatic potential energy, U_m, stored in the capacitor in Figure 7.11 is then given by the following:

$$U_m = \frac{1}{2}CV_b^2 = \frac{\varepsilon_r \nu_0 w v V_b^2}{2d} \tag{7.34}$$

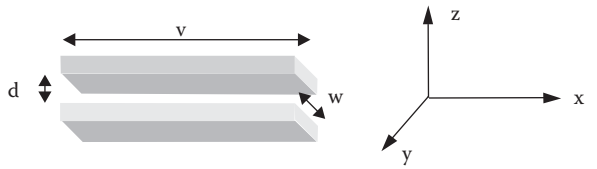


FIGURE 7.11 A parallel plate capacitor of plate size $w \times v$ and separation d stores a maximum potential energy of U_m.

where V_b = electrical breakdown voltage and $\varepsilon_r\varepsilon_0 \times w \times v/d$ = capacitance C. The permittivity of vacuum, ε_0, and the relative permittivity, ε_r, remain unchanged with scaling, so we can assign them an l^0 dimension. We also assume, for now, that V_b scales linearly with d. Intuitively, we expect a smaller gap, d, to result, indeed, in a lower breakdown voltage, V_b. However, as we shall see next (contrary to our macrointuition), at very small capacitor plate separation, the linear relationship between V_b and d breaks down. The breakdown voltage of the capacitor actually starts increasing with decreasing plate separation (see discussion below on continuum breakdown in electrostatics and the discussion on the Paschen curve below, in Figure 7.12).

Writing out all the dimensions in Equation 7.34 in terms of l, we obtain the following:

$$U_m \propto \frac{l^0\, l^0 l^1\, l^1\, (l^1)^2}{l^1} = l^3 \tag{7.35}$$

and the maximum energy stored in the capacitor scales as l^3. It follows that if l decreases by a factor of 10 (i.e., w, v, and d all decrease simultaneously), the stored maximum potential energy in the capacitor decreases by a factor of 1000.

The maximum electrostatic energy density, U'_m (in J/m^3), between the capacitor plates is obtained from Equation 7.34 by dividing the maximum potential energy, U_m, by the volume, V_0, of the capacitor gap, $V_0 = w \times v \times d$:

$$U'_m = \frac{\varepsilon_r \nu_0 E_b^2}{2} \tag{7.36}$$

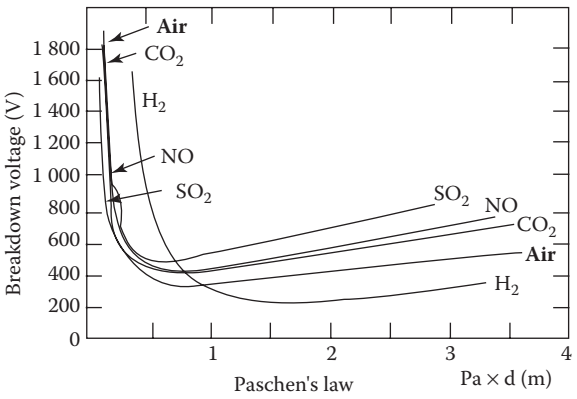


FIGURE 7.12 Breakdown voltage between two conductors, plotted as a function of the product of the distance between them and the surrounding gas pressure (P).

This expression for the energy density is general and is not restricted to the special case of a constant field in a parallel plate capacitor. In the case of an air capacitor, E_b is limited to approximately 3×10^6 Vm^{-1} by the electrical breakdown of air. Thus, the stored energy for an air capacitor is about 40 Jm^{-3}.[17] In power capacitors (see Chapter 9, on power and brains in miniature devices), with special dielectrics and a field of 10×10^6 Vm^{-1}, energy densities of 1 kJm^{-3} have been achieved.

We are interested in using electrostatic force in all types of actuators (see Chapter 8, on actuators), and that force, F, is the negative spatial derivative of the energy, U. In a case where we move the plate of a capacitor with respect to the other plate, the electric force will be the spatial derivative, in that direction, of the stored energy U. A translational movement of one plate with respect to the other in a microactuator in the direction x leads to a force, F_x, as follows:

$$F_x = -\frac{\partial U}{\partial x} = \frac{V^2}{2}\frac{\partial C}{\partial x} = \frac{l^3}{l^1} = l^2 \qquad (7.37)$$

where V is the applied voltage. The electrostatic force for a constant field is thus found to scale as l^2. This is often an advantage, because the mass and, hence, inertial forces scale as l^3. The electrostatic force gains over inertial forces as the size of the system is decreased. A decrease in size by a factor of 10 leads to a decrease of the inertial forces by a factor of 1000, whereas the electrostatic force decreases by a factor of only 100. It is thus not hard to exert rapid accelerations in microsystems with the aid of electrical fields. These are the type of features we will explore further in different types of electrostatic actuators in Chapter 8.

Breakdown of Continuum Theory in Electrostatics

We will now consider where macroscopic electrostatic laws err in the microscopic regime. The fundamental boundary condition in a gas pertains to the main free path, λ, the average distance traveled by a molecule between two successive collisions. The mean free path, λ, can be derived from the kinetic theory of gases as follows:

$$\lambda = \frac{1}{\sqrt{2}\pi n a^2} = \frac{kT}{\sqrt{2}\pi P a^2} \qquad (7.38)$$

where n = density (the number of molecules per unit volume)
T = temperature
k = Boltzmann constant
P = pressure
a = size of the molecule

Using Avogadro's number, we know that at room temperature and 1 atmosphere, 1 cm^3 contains 2.69×10^{19} molecules. Assuming that oxygen and nitrogen have a molecule size of about 3 Å, one calculates then a mean free path in air of about 60 nm.

Electrostatic charges, we saw earlier, arise from buildup or deficit of free electrons in a material. An electrically charged material can exert an attractive force on oppositely charged objects or a repulsive force on similarly charged ones. At small distances between two conductors in air, the electrical field is not isotropic as assumed above. This can be gleaned from the so-called *Paschen curve* in Figure 7.12. This figure represents the breakdown voltage between two conductors, plotted as a function of the product of the distance between them and the surrounding gas pressure (P). In the case of an air capacitor, the field strength required to cause sparks between the two electrodes (E_b), we saw in the section above, is approximately 3×10^6 Vm^{-1}. The general shape of the Paschen curve can be understood more easily by assuming that the pressure, P, remains constant at 1 atm, thus making the x-axis a simple distance axis. We notice that, on the right side of the Paschen curve, at large electrode distances, the field is constant and the scaling laws introduced above all pertain. But for smaller electrode gaps (below about 5 μm in air), the curve sharply reverses, bending upward and leading to higher electrical breakdown fields for thinner air gaps. Electrical breakdown in these small gaps does occur at much larger voltages than those predicted from linear scaling. From Bart et al.[17] and Busch-Vishniac[18] and references therein, we quote observed electrical fields of $10^8 \times$ Vm^{-1} (1.5-μm air gap), 1.7×10^8 Vm^{-1} (2-μm air gap), and 3.2×10^7 Vm^{-1} (12.5-μm air gap). The upper limit is the electrical field measured for small gaps in vacuum, that is, 3.0×10^8 Vm^{-1}.[19] Surface roughness results in a

lower average breakdown voltage, but even then, these fields are significantly higher than the 3×10^6 Vm^{-1} quoted above for breakdown of capacitors with large air gaps. Assuming a field strength of 3×10^8 Vm^{-1}, Equation 7.36 predicts an energy density of about 4×10^5 Jm^{-3}, compared with 40 Jm^{-3} without the Paschen effect. This unexpected result derives from the fact that there are not enough ionizing collisions to induce an avalanche between electrodes with a separation comparable to the mean free path of the surrounding gas. When the gap between two conductors approaches the mean free path of the molecules (λ), statistically fewer molecules are present to be ionized between those closely spaced conductors. Higher fields and higher-energy densities can be achieved before the critical breakdown voltage, V_b, is reached. The Paschen curve also illustrates, how, in terms of influencing the breakdown voltage, reducing the gap between two electrodes is equivalent to reducing the surrounding gas pressure: the x-axis is the product of pressure (P) and distance (d), and they exert the same effect on the breakdown voltage. One can thus actually say that reducing the distance between two conductors in a gas is equivalent to reducing the pressure in a macroscopic set-up.

In addition to the advantage of higher-energy densities in micromachined electrostatic systems, the Paschen effect suggests yet another advantage. The scaling of the electrical field (E_b) in the nonisotropic region is more like $l^{-1/2}$. According to Equation 7.37, in the macroregime, the force on a capacitor scales as l^2, or as a surface force. In the Paschen regime, the force scales with l so that with a factor of 10 size reduction, the inertial force still decreases by a factor of 1000, but the electrostatic force decreases only by a factor of 10. On the basis of Equation 7.23, we further see that acceleration and transit times are also higher than for the isotropic system.

Working on the left side of the Paschen curve renders high-field operation possible for a wide variety of electrostatic microdevices without incurring catastrophic sparking. Examples include the electrostatic motors illustrated in Figure 8.5, the electrostatic switch in Figure 8.19, and the author's volcano ionization sources presented in the example below.

Example: Ion Mobility Spectrometer

In an ion mobility spectrometer (IMS), ions produced at atmospheric pressure in an ionization cell by a ^{63}Ni beta source are accelerated in a drift chamber (uniform field of 150–250 V/cm), where they are separated according to their mobilities, detected as a current on a Faraday plate, and plotted on a time axis in accordance with their time of arrival (see Figure 7.13). About twenty years ago, this author set out to micromachine some critical components of an IMS instrument. We recognized that field uniformity, temperature, and pressure could be controlled with micromachined Si sensors. Our aim, though, was to substitute the ^{63}Ni ionization source

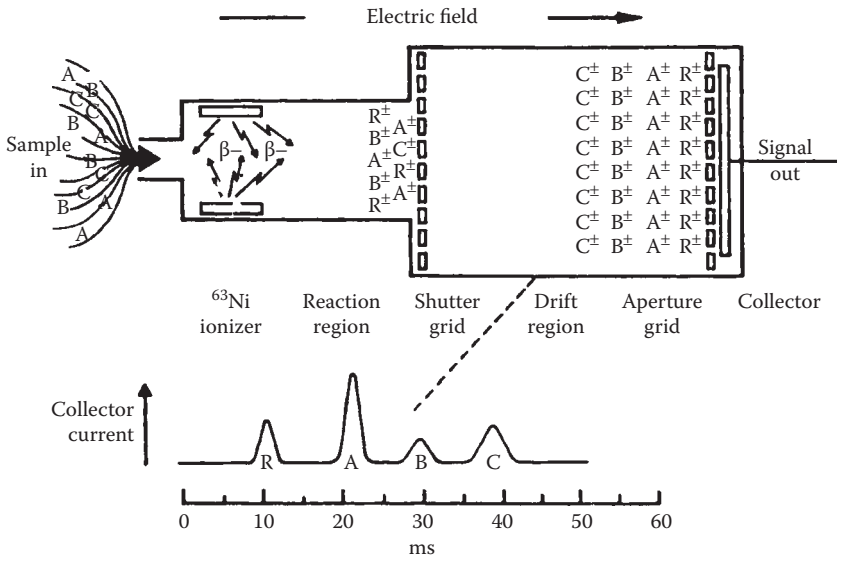


FIGURE 7.13 Ion mobility spectrometry with a Ni ionization source.

with a less fragmenting ionization source employing a "softer" field ionization (FI) technique. In FI, one creates mainly positive parent ions (M^+), and thus simpler spectra for complex environmental gas mixtures result. FI is well known from FI mass spectrometry (FIMS). But mass spectrometry is carried out in vacuum, and considering the extreme electric fields required for FI, one might expect that, under atmospheric pressure conditions, catastrophic electrical breakdown between the ionizing electrodes would occur. The idea about how to avoid breakdown or sparking at atmospheric pressures is based on the peculiar behavior of microelectrodes operating at the left side of the Paschen curve (see Figure 7.12), where much higher voltages are required for breakdown to occur.

For an ionization source operating on the left side of the Paschen curve, we microfabricated an array of microvolcanoes with a typical throat opening of 1 μm and a volcano rim to gate distance of less than 1 μm. Both the volcano rim and gate electrode may be made from a variety of inert metals. A scanning electron microscope (SEM) photomicrograph of a single microvolcano and an array of microvolcanoes is shown in Figure 7.14a (the metal used in this case is Pt). Given the small dimensions of the microvolcanoes, the intense electric fields required for field ionization can be produced with significantly

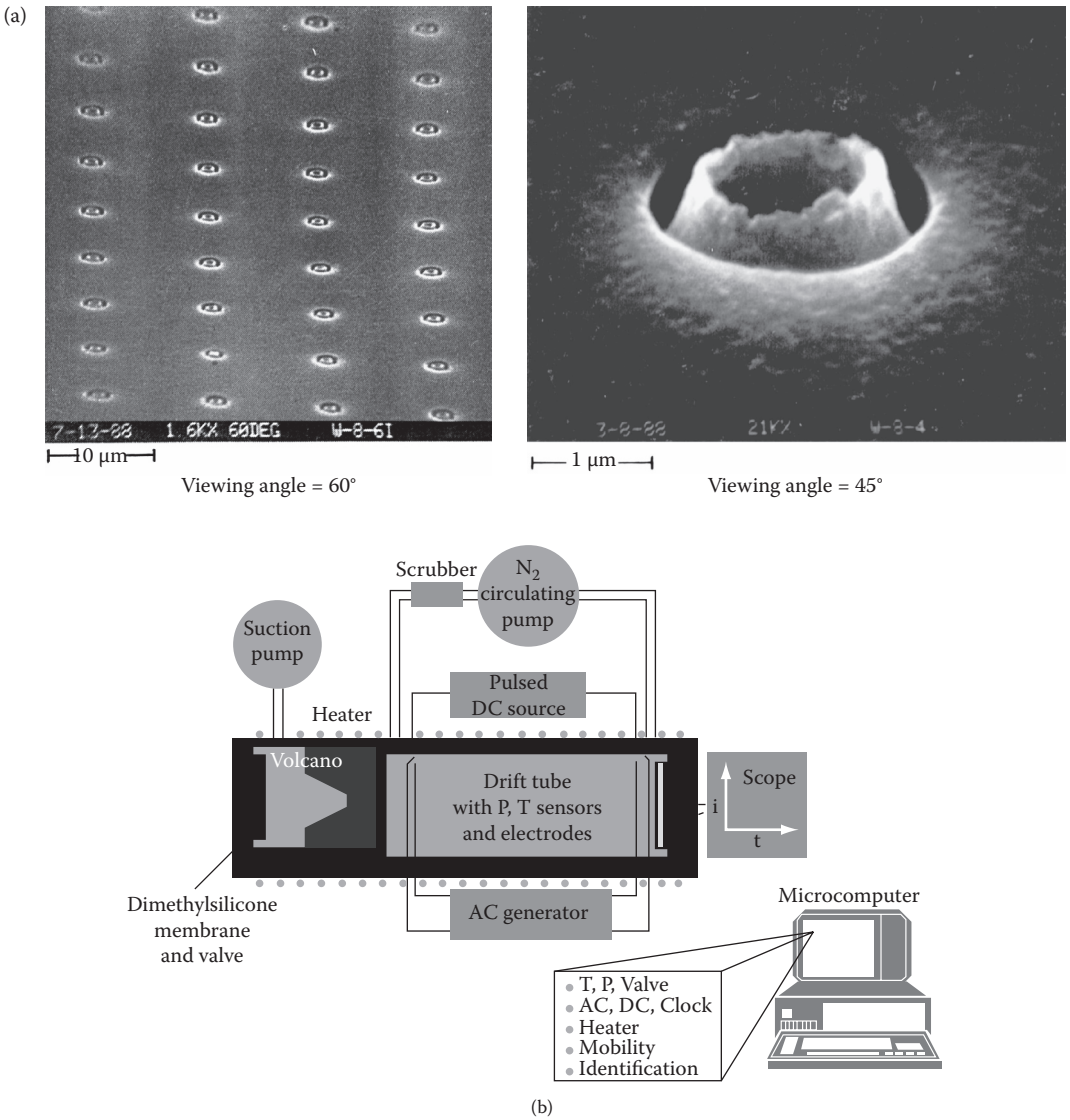


FIGURE 7.14 Ion mobility spectrometry with micromachined volcano sources. (a) Scanning electron microscope image of micromachined volcanoes. (b) Setup for ion mobility spectrometry with microvolcano ionization sources.

lower voltages than other FI sources. Referring back to the Paschen curve, it is clear that the microvolcano sources should operate at atmospheric pressure with voltages up to several hundred volts and perhaps as high as a kilovolt before they induce sparking. In collaboration with Dr. M. Coggiola, at SRI International, we did show that these sources can indeed resist electrical breakdown at atmospheric pressure and produce usable current levels from positive ions formed from gases such as pyridine, butane, and toluene in a setup like that shown in Figure 7.14b.[20]

The fabrication process of the microvolcanoes, which are embedded in a brittle 1-μm-thick Si dioxide layer, was very problematic, and failure rate was very high. Sturdier, more reliable volcano sources are needed to continue this type of development. Today, much better, high-aspect-ratio micromachining techniques are available, and it would be very worthwhile to reinitiate such a study.

Questions

Questions by Omid Rohani, UC Irvine

7.1: What do you expect will be the influence of miniaturization on a potentiometric sensor and on an amperometric sensor?

7.2: Explain in detail the scaling laws for flying and swimming.

7.3: The surface waves in water are characterized by an angular velocity ω and a wave number $k = 2\pi\lambda$ (λ, wavelength). In this problem we intend to find a dispersion relation expressing ω as a function of *k*.

(a) Using Buckingham's π theorem, find a dispersion relation containing ω and *k*. Accept that the depth *d*, the wave height *h*, the acceleration of gravity *g*, the waters density ρ, and the water surface tension γ play roles in this problem and neglect the water viscosity.

Appendix 7A: Common Dimensional Numbers

Parameter	Definition	Qualitative Ratio	Importance of Effects
Reynolds number	$R_E = \rho UL/\mu$	Inertia/Viscosity	Always
Mach number	$MA = U/A$	Flow speed/Sound speed	Compressible flow
Froude number	$Fr = U^2/gL$	Inertia/gravity	Free surface flow
Weber number	$W_e = \rho U^2 L/\gamma$	Inertia/Surface tension	Free surface flow
Cavitation number (Euler number)	$Ca = (p - p_v)/\rho U^2$	Pressure/Inertia	Cavitation
Prandtl number	$Pr = C_p\mu/k$	Dissipation/conduction	Heat convection
Eckert number	$E_c = U^2/c_pT_o$	Kinetic energy/enthalpy	Dissipation
Specific heat ratio	$\gamma = c_p/c_v$	Enthalpy/internal energy	Compressible flow
Strouhal number	$St = \omega L/U$	Oscillation/mean speed	Oscillating flow
Roughness ratio	ε/L	Wall roughness/Body length	Turbulent, rough walls
Grashof number	$Gr = \beta\Delta TgL^3\rho^2/\mu^2$	Buoyancy/viscosity	Natural convection
Temperature ratio	T_w/T_o	Wall temperature/stream temperature	Heat transfer
Pressure coefficient	$C_p = (p-p_o)/0.5\rho U^2$	Static pressure/dynamic pressure	Aerodynamics, hydrodynamics
Lift coefficient	$C_L = L/0.5\rho U^2A$	Lift force/dynamics force	Aerodynamics, hydrodynamics
Drag coefficient	$C_D = D/0.5\rho U^2A$	Lift force/dynamics force	Aerodynamics, hydrodynamics
Euler number	$E_u = g_c\Delta P/\rho V^2$	Friction head × velocity head	Momentum transfer, fluid frictions
Weber number	$W_e = DV^2\,\rho/g_c\sigma$	Inertial force/surface tension force	Momentum transfer

(b) Now suppose that the water is deep compared to the wave length and the wave height is small compared to the wave length. Also suppose that the waves are long. Try to write the relation you got in part (a) in a simpler form expressing ω just as a function of k.

7.4: Explain the DC breakdown voltage versus electrode distance curve and how it is relevant to dry etching. How is miniaturization of an electrode set equivalent to creating a local vacuum?

7.5: Why is the range of sizes in water-based life much greater than terrestrial life range? Discuss both above and below limits of these ranges.

7.6: Fill in the blank with the correct number for these scaling laws.
(a) Mass scales with a factor of _____.
(b) Stress scales with a factor of _____.
(c) Natural frequency scales with a factor of _____.
(d) Viscous damping scales with a factor of _____.
(e) Coulomb damping scales with a factor _____.
(f) Elastic coulomb damping scales with a factor of _____.
(g) Surface adhesion scales with a factor of _____.
(h) Power scales with a factor of _____.

7.7: How is the number of steps one animal must take per unit distance related to the animal's weight?

7.8: How does velocity scale with size (ignore drag/damping)? (A short two- or three- equation answer—the solution may be surprising.)

7.9: What methods do you recommend to overcome the problem of retention of liquids caused by large surface-to-volume ratios in small biosensors?

7.10: A simple and accurate viscometer can be made from a length of capillary tubing. If the flow rate and pressure drop are measured, and the tube geometry is known, the viscosity can be computed. A test of a certain liquid in a capillary viscometer gave the following data:
- flow rate: 880 mm^3/s
- tube diameter: 500 μm
- tube length: 1 m
- pressure drop: 1.0 MPa

Determine the viscosity of the liquid

7.11: The accepted transition Reynolds number for flow in a circular pipe is Re ~2300. For flow through a 6 cm dia. pipe, at what velocity will this occur at 20°C for (a) airflow and (b) water flow?

7.12: What are the forces that play an important role for (a) very large and (b) very small devices?

7.13: Explain Kleiber's law. What is the generic principle underlying Kleiber's law?

7.14: How do the following forces scale?
(a) Electrostatic force in a constant electric field
(b) Pneumatic and hydraulic forces caused by pressure

7.15: How does the acceleration, transit time, and produced power for each case scale in problem 7.14?

7.16: The required power in a stirring device in a tank is a function of diameter of the device d, number of rotations of the motor per unit time N, density of liquid ρ and viscosity of the liquid η.
(a) Using Buckingham's π theorem, obtain a relation between power P and d, N, ρ, η.
(b) Supposing that the power P is proportional to the square of N, determine by what factor would P be expected to increase if d is doubled.

7.17: A centrifugal pump for pumping the lubricating oil at 15°C rotates at 1000 rpm. For testing this pump, we make a model pump which works with 20°C air. Assuming the diameter of model is 4 times the diameter of the prototype, find the speed at which the model pump should run.

7.18: Complete similarity between a ship model and a full-sized ship requires the same Reynolds and Froude numbers. Suppose that

$$Re = \rho VL/\mu = 11 \times 10^8 \text{ and } Fr^{-1} = gL/V^2 = 33.25$$

Assuming a model 1/100 the size of the ship, can you design an experiment having full similarity?

7.19: In the miniaturization of analytical instruments, scaling laws and breakdown of scaling

laws often determine whether miniaturization will favor sensitivity or not. What will happen to the sensitivity of the following techniques upon miniaturization: (a) the optical path in UV spectrometer; (b) a potentiometric sensor (e.g., a pH sensor); (c) an amperometric sensor (e.g., an oxygen gas sensor); (d) the column in a GC?

7.20: If you miniaturize an absorption-based optical analytical instrument and one based on luminescence, which one scales down more favorably? Explain why.

7. 21: What is the primary assumption of continuum mechanics? Give examples where continuum theory breaks down and how MEMS, in some cases, may take advantage of this breakdown.

7.22: Using dimensional analysis, derive the flow Q of an ideal liquid through an orifice in terms of the density of the liquid ρ, the pressure difference ΔP, and the diameter of the orifice d.

References

1. Trimmer, W. S., ed. 1997. *Micromechanics and MEMS classic and seminal papers to 1990*. New York: IEEE.
2. Schmidt-Nielsen, K. 1984. *Scaling: Why is animal size so important?* Cambridge, United Kingdom: Cambridge University Press.
3. McMahon, T., and J. T. Bonner. 1985. *On size and life.* New York: Scientific American Library.
4. Bonner, J. T. 2006. *Why size matters: from bacteria to blue whales.* Princeton, NJ: Princeton University Press.
5. Dodds, S., D. H. Rothman, and J. S. Weitz. 2000. Re-examination of the "3/4-law" of metabolism. *J Theor Biol* 209:9–27.
6. Gillooly, J. F., A. P. Allen, G. B. West, and J. H. Brown. 2005. The rate of DNA evolution: Effects of body size and temperature on the molecular clock. *Proc Natl Acad Sci U S A* 102:140–45.
7. Langhaar, H. L. 1951. *Dimensional analysis and the theory of models.* New York: Wiley.
8. Morowitz, H. J. 1985. *Mayonnaise and the origin of life.* New York: Berkley Books.
9. Hayashi, T. 1994. Micromechanism and their characteristics. *IEEE International Workshop on Micro Electro Mechanical Systems (MEMS '94)*. Oiso, Japan, January 25–28, 1994, pp. 39–44.
10. Fujimasa, I. 1996. *Micromachines: A new era in mechanical engineering.* Oxford, United Kingdom: Oxford University Press.
11. Kubo, Y., I. Shimoyama, and H. Miura. 1993. Study of insect-based flying microrobots. *IEEE International Conference on Robotics and Automation*. Atlanta, GA, May 1993, pp. 386–91.
12. Autumn, K., Y. A. Liang, S. T. Hsieh, W. Zesch, W. P. Chan, T. W. Kenny, R. Fearing, and R. J. Full. 2000. Adhesive force of a single gecko foot-hair. *Nature* 405:681–85.
13. Menon, C., and S. Metin. 2006. A biomimetic climbing robot based on the gecko. *J Bionic Eng* 3:115–25.
14. Manz, A., J. Harrison, E. Verpoorte, J. Fettinger, A. Paulus, H. Lüdi, and H. Widmer. 1992. Planar chips technology for miniaturization and integration of separation techniques into monitoring systems: Capillary electrophoresis on a chip. *J Chromatogr* 593:253–58.
15. Brown, P. R., and E. Grushka, eds. 1993. *Advances in chromatography.* New York: Marcel Dekker.
16. Thompson, D. W. 1992 (reprint of 1917 edition). *On growth and form, Canto ed.* Cambridge, United Kingdom: Cambridge University Press.
17. Bart, S. F., T. A. Lober, R. T. Howe, J. H. Lang, and M. F. Schlecht. 1988. Design considerations for micromachined electric actuators. *Sensors Actuators* 14:269–92.
18. Busch-Vishniac, I. J. 1992. The case for magnetically driven microactuators. *Sensors Actuators A* A33:207–20.
19. Bollee, B. 1969. Electrostatic motors. *Philips Tech Rev* 30:178–94.
20. Madou, M. J., and S. R. Morrison. 1991. High-field operation of submicrometer devices at atmospheric pressure. *1991 International Conference on Solid-State Sensors and Actuators.* San Francisco, June 24–28, 1991, p. 145.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Outline

Introduction

Electric Actuators

Electrostatics

Correspondence between Electrical and Mechanical Circuits

Dynamics of Electrostatic Actuators

Fundamental Frequency and Quality Factor of Electrostatic Resonators

Piezoelectricity

Overview of Types of Acoustic Waves

Piezoelectric Applications

Acoustic Sensing of Mass

Acoustic Imaging

Piezoelectricity Summary

Dielectric Induction

Electrostriction

Pyroelectricity

Thermoelectricity

Electrets

Electrorheological Fluids

Magnetic Actuators

Examples of Magnetic Actuators

Comparing Magnetic with Electrostatic Actuation

Magnetostriction

Magnetohydrodynamic Pumping

Magnetorheological Fluids, Ferrofluids, and Ferrowaxes

Superconductivity

Thermal Actuators

Types of Thermomechanical Actuators

Comparison of Actuator Principles in Commercial Valves

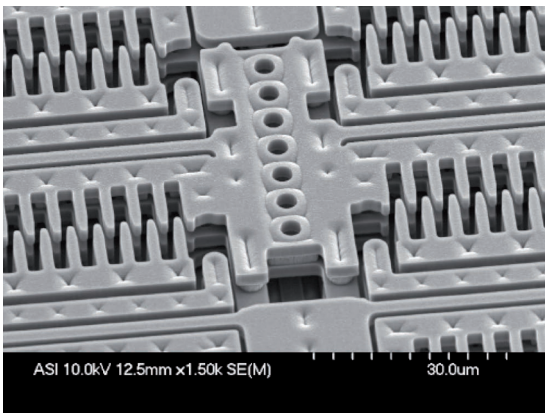
Other Actuators and Some Final Actuator Comparisons: Final Comparison

Questions

References

8

Actuators



MEMS electrostatic comb actuator. (Courtesy of MEMX.)

> Their knowledge did not lead one to the contemplation of divinity or liberation; their science was action: nature obeyed them, and in their cities the power of the rich and strong was less oppressive. The old magic was now in reach of everyone who knew the formula for the spell.
>
> **Octavio Paz in a passage on the science and technology of the West in *In Light of India***

Introduction

An actuator, like a sensor, is a device that converts energy from one form to another (Figure 8.1). In the case of an actuator, one is interested in the ensuing action, while in the case of a sensor what is of interest is the information gained. On the subject of sensor/actuators, we review electrical, magnetic, thermal, chemical, and electrochemical phenomena. From the previous chapter, we know that the dominant physical quantities may change as a function of length scale. In this context, we saw how gravitational (inertial) forces become less effective in the microdomain, whereas van der Waals forces, electrostatic forces, and surface tension forces become dominant. In this chapter we compare different sensor/actuators on the basis of their scaling behavior and other properties,

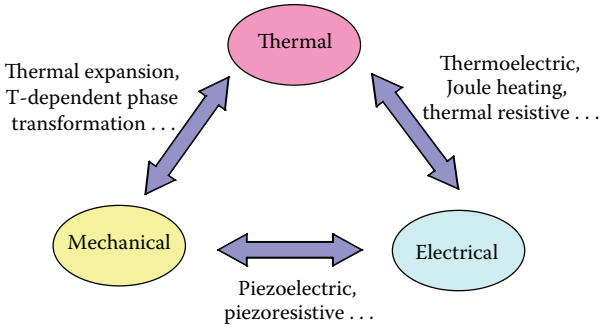


FIGURE 8.1 Energy conversion in sensor and actuator systems.

which, besides size and cost, include dynamic range, displacement (linear or angular), force or torque, response time, and power consumption (see Table 8.1). The results of these comparisons must be seen as guideposts toward a decision only on what actuator to use; yet other considerations (e.g., size of the absolute forces involved, the potential for integration with electronics, materials choice, and manufacturability) may influence the design rules for a particular microsensor or microactuator.

TABLE 8.1 Sensor/Actuator Design Principles and Considerations. Numbers refer to the numbered explanations below the table

• Operational principle (e.g., electrostatic capacitor, shape memory effect, solenoid inductor, thermoelectric effect, piezoelectric, thermal expansion, and themoresistive)
• Scaling behavior
• Size and cost
• Materials choices
• Process feasibility
• Size of absolute forces obtained
• Sensitivity (signal-to-noise ratio) (1)
• Accuracy (3)
• Output force/energy/power
• Dynamic range of frequency response (8)
• Hysteresis (4)
• Resolution (7)
• Nonlinearity (also linearity) (5)
• Noise (6)
• Span or bandwidth range (2)
• Limitations: quality factor, failure mode
• Temperature (or other factors) dependence
• Reliability
• Manufacturability
• Integration and packaging

1. Sensitivity. The sensitivity of a sensor/actuator is defined in terms of the relationship between input signal and output signal. The sensitivity is, in general, the ratio of a small change in output signal to a small change in input signal. As such, it may be expressed as the derivative of the transfer function* with respect to the input signal. A typical unit is volts/kelvin. A thermometer would have high sensitivity if a small temperature change resulted in a large voltage change.
2. Span or bandwidth. The range of input signals that may be converted to output signals by the sensor/actuator. Signals outside of this range are expected to cause unacceptably large inaccuracy. This span or dynamic range is usually specified by the sensor/actuator supplier as the range over which other performance characteristics described in the data sheets are expected to apply.
3. Accuracy. Generally defined as the largest expected error between actual and ideal output signals. Sometimes this is quoted as a fraction of the full-scale output. For example, a thermometer might be guaranteed to be accurate within 5% of full-scale output (FSO).
4. Hysteresis. Some sensors/actuators do not return to the same output value when the input stimulus is cycled up or down. The width of the expected error in terms of the measured quantity is defined as the hysteresis. A typical unit is percentage of full-scale output.
5. Nonlinearity (often called *linearity*). The maximum deviation from a linear transfer function over the specified dynamic range. There are several measures of this error. The most common compares the actual transfer function with the "best straight line," which lies midway between the two parallel lines that encompass the entire transfer function over the specified dynamic

* The functional relationship between input signal and output signal. Usually, this function is represented as a graph showing the relationship between the input and output signal, and the details of this relationship may constitute a complete description of the sensor/actuator characteristics. For expensive sensors/actuators which are individually calibrated, this might take the form of a certified calibration curve.

range of the device. This choice of comparison method is popular because it makes most sensors/actuators look their best.

6. Noise. All sensors/actuators produce some output noise in addition to the output signal. This noise limits the performance of the system. Noise is generally distributed across the frequency spectrum. Many common noise sources produce a white noise distribution, which is to say that the spectral noise density is the same at all frequencies. Since there is an inverse relationship between the bandwidth and measurement time, it can be said that the noise decreases with the square root of the measurement time.
7. Resolution. The resolution of a sensor/actuator is defined as the minimum detectable signal fluctuation. Since fluctuations are temporal phenomena, there is some relationship between the timescale for the fluctuation and the minimum detectable amplitude. Therefore, the definition of resolution must include some information about the nature of the measurement being carried out.
8. Dynamic range or frequency response. All sensors/actuators have finite response times to an instantaneous change in input signal. In addition, many sensors have decay times, which would represent the time after a step change in physical signal for the sensor output to decay to its original value. The reciprocal of these times correspond to the upper and lower cutoff frequencies, respectively. The bandwidth of a sensor is the frequency range between these two frequencies.

Jargon from different science disciplines and various application fields often makes interdisciplinary microelectromechanical system (MEMS) and nanoelectromechanical system (NEMS) discussions challenging. A useful scheme for standardizing the classification of miniaturized sensors and actuators would make communication between researchers from different disciplines (manufacturers, immunologists, mechanical engineers, market researchers, etc.) easier. Several such attempts have been made in that direction. We state our favored approach here. White[1] describes a classification scheme derived from a Hitachi Research Laboratory communication. He distinguishes ten different domains or fields:

1. Acoustic
2. Biological
3. Chemical
4. Electrical
5. Magnetic
6. Mechanical
7. Optical
8. Radiant
9. Thermal
10. Other

Middlehoek and Audet[2] follow Lion[3] and contract this list to six domains:

1. Radiant
2. Mechanical
3. Thermal
4. Electrical
5. Magnetic
6. Chemical

For simplicity, we adopt the latter list here. Based on Gopel et al.[4] and Habekotte,[5] Table 8.2 exemplifies sensing/actuating principles in these six signal domains. Since the listed energy domains can be energy input as well as energy output, a 6×6 matrix results. The table also presents an example device in each field. We may have a mechanical sensor/actuator converting mechanical energy into thermal energy (i.e., a thermomechanical device such as a friction calorimeter), but also a thermal sensor/actuator converting thermal energy into mechanical energy (e.g., a mechanothermal sensor such as a bimetallic strip).

In the literature, no logic seems to be applied concerning the use of "thermo-optical" or "opto-thermal." Middlehoek and Audet[2] further distinguish between self-generating and modulating sensors/actuators. A device based on a modulating principle requires an auxiliary energy source. An example of a modulating sensor is a fiber-optic magnetic-field sensor in which a magnetostrictive jacket is used to convert a magnetic field into an induced strain in the optical fiber. A sensor/actuator based on a self-generating principle does not require an auxiliary energy source (e.g., a thermocouple in

Table 8.2 Example of Transduction Principles

Input (Primary Signal)	Output (Secondary Signal)					
	Mechanical (Mechano-)	Thermal (Thermo-)	Electrical (Electro-)	Magnetic (Magneto-)	Radiant (Photo- or Radio-)	Chemical (Chemo-)
Mechanical	Acoustics, fluidics	Friction calorimeter, cooling effects	Piezoelectricity, piezoresistivity	Piezomagnetic effect	Photoelasticity, Doppler effect	Motor proteins
Thermal	Thermal expansion, bimetallic strip		Pyroelectricity, Seebeck effect		Radiant emission	Reaction activation
Electrical	Piezoelectricity, electrometer	Joule heating, Peltier effect	Langmuir probe	Biot-Savart's law	Electroluminescence, Kerr effect	Electrolysis
Magnetic	Magnetostriction, magnetometer	Thermomagnetic	Magnetoresistance		Faraday effect, Cotton-Moulton effect	
Radiant	Radiation pressure	Thermopile, bolometer	Photoelectric, dember		Optical bistability	Photoreactions
Chemical	Hygrometer	Calorimeter, thermal conductivity	Amperometry, flame ionization, Volta effect	Nuclear magnetic resonance	Chemiluminescence	

which a change in temperature directly results in an electrical signal that can be measured). The difference between the two is illustrated in Figure 8.2.[6] Since the modulating sensor needs another energy input, one also refers to such sensors as *passive*, whereas the self-generating sensors are *active*.

Irrespective of the fact of dealing with a small sensor, an actuator, or a microsystem, we can always describe the function of the miniaturized device in terms of the matrix in Table 8.2.

There are many reasons for wanting to miniaturize systems, actuators, power sources, sensors, and components. In some cases, the reason is favorable scaling—for example, smaller devices tend to be faster and consume less power. Scaling may also disfavor miniaturization, as often happens with actuators

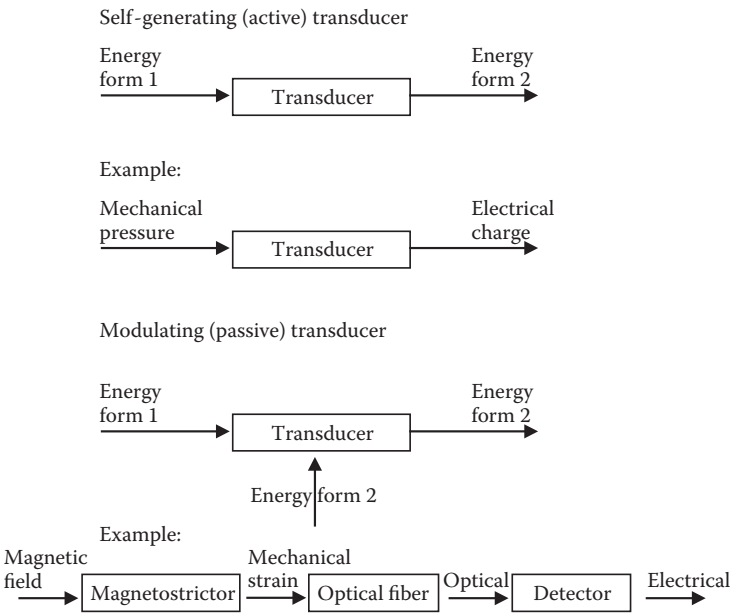


FIGURE 8.2 Comparison of self-generating (active) transducer and modulating (passive) transducer; the piezoelectric sensor represents a self-generating transducer, and the fiber optic magnetic field sensor represents a modulating transducer. (After National Research Council.)

and power sources; smaller actuators exert less force, and smaller power sources harness less power.

Finally, actuator materials are often called *smart materials*, which refers to materials that undergo precise transformations through physical interactions. Examples include piezoelectric materials, shape memory alloys (SMAs), conductive polymers, electrochromic materials, smart gels, rheological materials, biomaterials, electrostrictive materials, magnetostrictive materials, and electrorheological fluids. All these materials have potential uses in actuators that control chatter in precision machine tools, improved robotic parts that move faster and with greater accuracy, smaller microelectronic circuits in machines ranging from computers to photolithography printers, and health-monitoring fibers for bridges, buildings, and utility poles.

Electric Actuators

Under electric actuators, we consider all actuators that involve electrical fields (DC or AC); this includes, foremost, electrostatic and piezoelectric actuators, which we study in the most detail. Besides electrostatic and piezoelectric devices, we consider dielectric induction effects on particles in liquids and on liquids themselves, where one speaks about dielectrophoresis (covered in Volume I, Chapter 6) and electrohydrodynamics (EHD), respectively. Finally, under electric actuators, we also explain how electrostrictive materials, pyroelectrics, thermoelectrics, electrets, and electrorheological fluids work.

Electrostatics

Introduction

In electrostatic sensors and actuators, electrically charged materials are used. These materials can exert an attractive force on oppositely charged objects or a repulsive force on similarly charged ones. We start by introducing the simplest of electrostatic actuators, namely a parallel plate capacitor. This is followed by an introduction to electrical and mechanical equivalent circuits and the correspondence of their respective elements. This gives us the background for considering the dynamic behavior of a parallel plate capacitor. We then further describe the two most important characteristics of electrostatic actuators, i.e., their natural frequency and quality factor. We also introduce some other electrostatic actuators, such as wobble motors and switches, and compare them with other types of actuators. We finish this section with a summary about electrostatic actuators and a more general approach to modeling them.

Parallel Plate Capacitor Based Actuators

A parallel plate capacitor, as shown in Figure 8.3, is an example of a simple electrostatic actuator. We calculate the stored energy density in a capacitor with plate area A and distance between the plates d as follows:

$$U = \frac{1}{2}\varepsilon_0\varepsilon_r\frac{A}{d}V^2 = \frac{1}{2}\varepsilon(Ad)E^2 \qquad (8.1)$$

(see also Equation 7.34, with $\varepsilon = \varepsilon_r\varepsilon_0$ wv = A and $V_b = V$).

The stored energy density is then obtained by dividing U from Equation 8.1 by the volume of the capacitor (Ad), or:

$$U' = \frac{1}{2}\varepsilon E^2 = \frac{1}{2}\mathbf{D}\cdot\mathbf{E} \qquad (8.2)$$

where we bolded D and E to remind the reader that we are dealing with vectors here (see also Equation 7.36, modified to describe all fields, E, not just the breakdown field, E_b, and with $\varepsilon = \varepsilon_r\varepsilon_0$):

The electrostatic surface force is given by the following equation (a generalization of Equation 7.37):

$$F = -\nabla U \qquad (8.3)$$

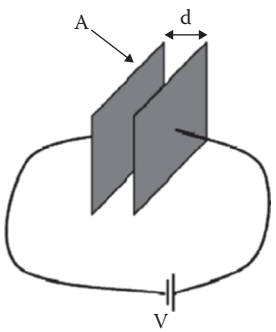


FIGURE 8.3 Parallel plate capacitor.

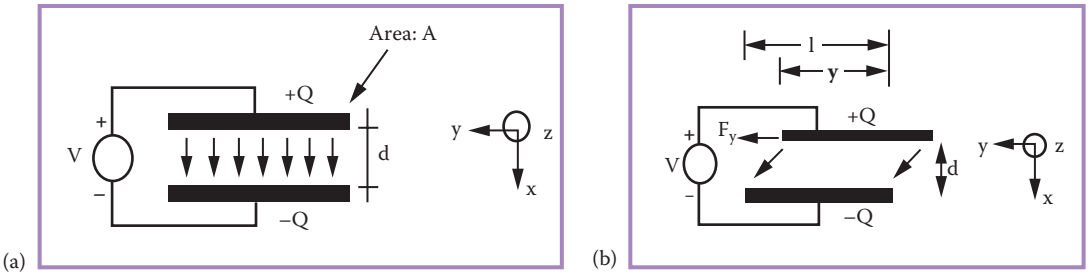


FIGURE 8.4 Vertical-driven (a) and parallel-driven (b) parallel plate capacitor actuators.

Let us now consider parallel- and vertical-driven parallel plate capacitor actuators, as depicted in Figure 8.4.

If the two plates of the parallel capacitor in a microactuator are displaced perpendicular to each other (d changes along the x-direction), the force of attraction in the x-direction, based on Equations 8.1 (with $\varepsilon_r = 1$) and 8.3, is given as follows:

$$F_x = -\frac{\partial U}{\partial x} = \frac{1}{2}\frac{\varepsilon_0 A}{x^2}V^2 \qquad (8.4)$$

Vertically driven resonant microstructures operate in this mode (see Figure 8.5a). Vibration in the x-direction in this case is excited electrostatically, and motion is detected electrostatically as well; that is, by sensing the change in capacitance.[7]

In Chapter 7, we saw that the electrostatic force, F_x, for a constant field, scales as l^2, since:

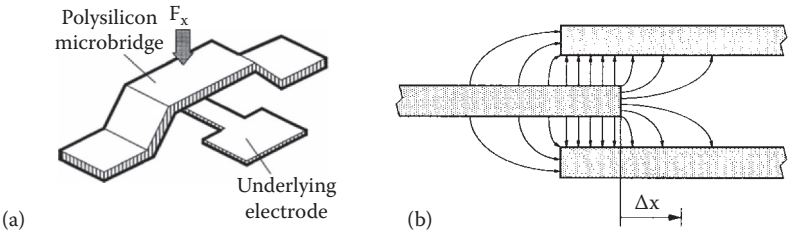
$$F_x = -\frac{\partial U}{\partial x} = \frac{V^2}{2}\frac{\partial C}{\partial x} = \frac{1^3}{1^1} = 1^2 \qquad (7.37)$$

where V is the applied voltage. This l^2 scaling of the electrostatic force often constitutes an advantage compared with other forces when miniaturizing an actuator. For example, mass and hence inertial forces scale as l^3, and from Chapter 7, we know that the lower the exponent of l, the more important the force becomes in the microdomain (see also below and Figure 8.8).

A large displacement in a micromachined electrostatic actuator element can best be achieved if one plate (the actuator) moves parallel to the other capacitor plate rather than perpendicular to it (see Figures 8.4b and 8.5b). For a parallel movement of the plates (in the y-direction or z-direction), Equations 8.1 and 8.3 result in a force of attraction parallel to the plates, which for the y-direction is given by the following equation:

$$F_y = \frac{1}{2}\frac{\varepsilon_0 z}{d}V^2 \qquad (8.5)$$

Notice that the force is constant for a y directional translation, i.e., the force is scale-independent in



Vertical resonator	Horizontal resonator
+ Parallel plates—good drive and sense	− Fringing fields—weak drive and sense
− Amplitude limited to less than 1/3 of gap	+ Larger vibration amplitudes
− Squeeze-film damping in air	+ Couette flow—damping less dissipative
− Difficult to couple mechanically	+ Enables a variety of mechanical coupling

(c)

FIGURE 8.5 Electrostatic actuation. (a) Vertically driven polysilicon microbridge. (b) Laterally driven electrostatic actuator. (After Tang, W. C.-K. 1990. Electrostatic comb drive for resonant sensor and actuator applications. PhD thesis, University of California, Berkeley.[7]) (c) Summary of the pros and cons of vertical and lateral resonators, from a lecture by Dr. Bill Tang (University of California, Irvine).

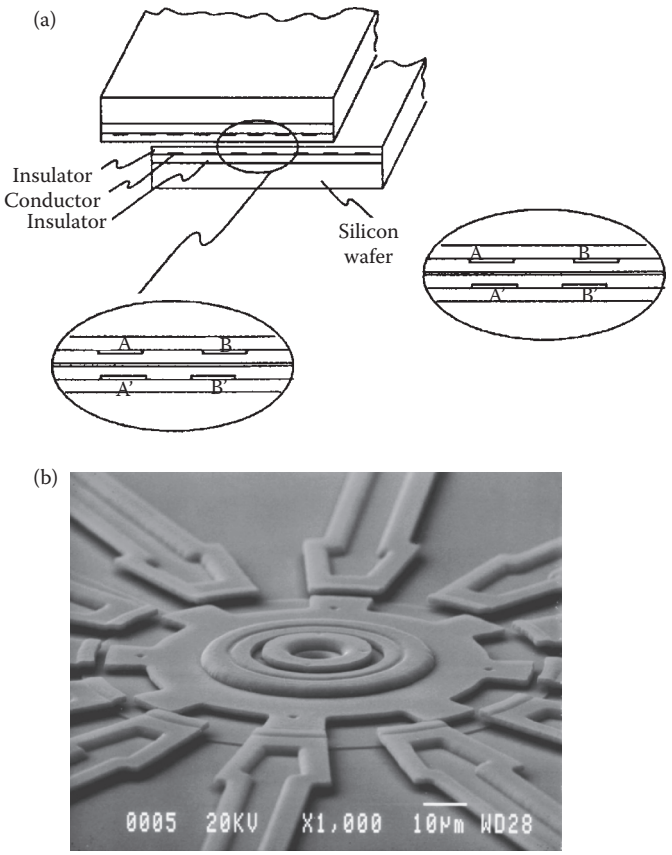


FIGURE 8.6 (a) Linear electrostatic motor: when a voltage is applied to the misaligned plates A-A', a force is exerted, which aligns plates A-A'. Now, plates B-B' are misaligned and in a position to be activated to cause a motion. (After Trimmer, W. S. N., and K. J. Gabriel. 1987. Design considerations for a practical electrostatic micromotor. *Sensors Actuators* 11:189.[11]) (b) Rotational electrostatic motor: scanning electron microscope micrograph of a polysilicon 12:8 salient pole micromotor. The rotor sits atop a 0.5-mm-thick layer of polysilicon that acts as an electrostatic shield. Rotor, hub, and stators are formed from 1.5-mm-thick polysilicon. A 2.0-mm-thick polysilicon disk is attached to the rotor. In turn, the hub overlaps this disk to pin the rotor onto the substrate. (Courtesy of D. Koester, MCNC-MEMS Technology Applications Center.) The first polysilicon motors were made at University of California, Berkeley (Fan, Tai, and Muller), Massachusetts Institute of Technology, and AT&T. Typical starting voltages were >100 V, and operating voltages were >50 V.

that direction ($F \sim l^0V^2$). In this case, the force parallel to the plates wants to realign the plates. This is exploited, for example, in laterally driven linear and rotary electrostatic motors or side-drive motors (see Figure 8.6a and b). The lateral force remains constant during the movement as long as the fringing fields can be neglected (see Figure 8.5b). Besides longer strokes, driving and sensing of planar microstructures parallel to the substrate have two other major advantages over a vertical movement of the plates: the forces change linearly with distance (see Equation 8.5), and dissipative squeeze damping is avoided.[8–10] The latter is related to the magnitude of the viscous losses when moving structures displace fluids (usually air) from small separator gaps. Losses are larger for vertical displacement of fluids, leading to squeeze-film damping and a resultant low-quality factor (Q) (see below) for resonating elements. Couette flow (see Volume I, Chapter 6, on fluidics) in the gap between the structure and the substrate for lateral motion of the actuator is much less dissipative than squeeze-film damping, and higher Q resonators are obtained. Figure 8.5c summarizes the pros and cons of vertical and lateral resonators.

Lateral electrostatic actuation is very important in linear resonators with comb-like structures, such as the device shown in the header of this chapter. These comb drives are used, for example, in accelerometers [see the Analog Devices' ADXL-50 accelerometer, with a surface micromachined capacitive sensor (center), on-chip excitation, self-test, and signal-conditioning circuitry in Volume II, Figure 7.92]. By applying a voltage, a movable shuttle moves toward a fixed part (Volume II, Figures 7.76 and 7.93). For a rectangular design of the resonator beams, the force and the microactuator displacement is proportional to the ratio of structure height, T, versus separation gap, d, that is, the achievable aspect ratio and the number of fingers, N:

$$F = \frac{1}{2}\frac{\varepsilon_0 TN}{d}V^2 \qquad (8.6)$$

An efficient comb drive with a large F can thus be achieved by designing many high comb fingers with narrow gaps. In polysilicon surface micromachining, the thickness, T, is typically limited to about 2–3 µm, whereas with LIGA or pseudo-LIGA methods, a thickness of several hundred micrometers is possible (Volume II, Chapter 10). Clearly, the latter machining methods have the potential to produce the strongest electrostatic actuators. Mohr et al. further showed that by making the comb fingers trapezoidal in shape, the maximum displacement of the interdigitated fingers can be further increased,

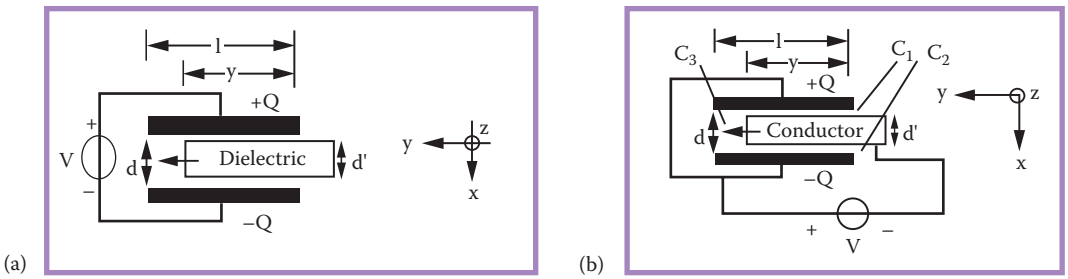


FIGURE 8.7 Dielectric (*a*) or metallic (*b*) slabs of material are moved into the gap of a parallel plate capacitor.

as smaller capacitor gaps are possible this way.[12] For a more rigorous mathematical derivation of Equation 8.6, we refer to the finite-element simulation of comb drives by Tang.[7]

When a dielectric or a metallic slab is moved between the two plates of a parallel plate capacitor as shown in Figure 8.7, Equation 8.5 must be modified as follows:

$$F_y = \frac{1}{2}\frac{(\varepsilon_d - \varepsilon_0)z}{d'}V^2 \tag{8.7}$$

for a dielectric slab of thickness d' and dielectric constant ε_d (Figure 8.7a). For a metal slab of thickness d' inserted in the capacitor gap, one derives the following equation (Figure 8.7b):

$$F_y = \frac{1}{2}\frac{\varepsilon_0 z d'}{d(d-d')}V^2 \tag{8.8}$$

It is important to be able to evaluate the relative importance of different competing forces as one downscales an actuator. In Figure 4.39, we compare gravity, van der Waals, electrostatic, and surface tension forces that are involved when picking and placing small objects during assembly. From this graph it is clear that surface forces become more dominant as structures become smaller. Here we detail the relative importance of gravitational versus electrostatic forces for deciding upon an actuator mechanism in the microworld. In Figure 8.8, we consider competing gravitational and electrostatic forces working on a cube of material with length L, density ρ, and charge q. The gravitational force, F_g, working on the cube is given as follows:

$$F_g = g\rho V = g\rho L^3 \tag{8.9}$$

Based on Equation 8.4, with $\sigma = q/A$ ($A = L^2$) and $AE = q/\varepsilon$, the electrostatic force F_e ($\equiv F_x$) working upon the cube is given by the following equation:

$$F_e = \frac{1}{2}\varepsilon AE^2 = \frac{\sigma^2}{2\varepsilon}L^2 \tag{8.10}$$

Hence, when:

$$L < \frac{\sigma^2}{2g\rho\varepsilon} \tag{8.11}$$

electrostatics dominates (the intersection of the red and blue line in Figure 8.8 occurs at $F_g = F_e$). The electrostatic force, as we learn in Chapter 7, gains over inertial forces as the size of the system is decreased: the electrostatic force, with the lowest exponent in l, becomes the dominant force in the microdomain (surface force is l^2 vs. l^3 for gravitation). A decrease in size by a factor of 10 leads to a decrease of the inertial forces by a factor of 1000, whereas the electrostatic force decreases by a factor of only 100. It is thus not hard to exert rapid accelerations in microsystems with the aid of electrical fields.

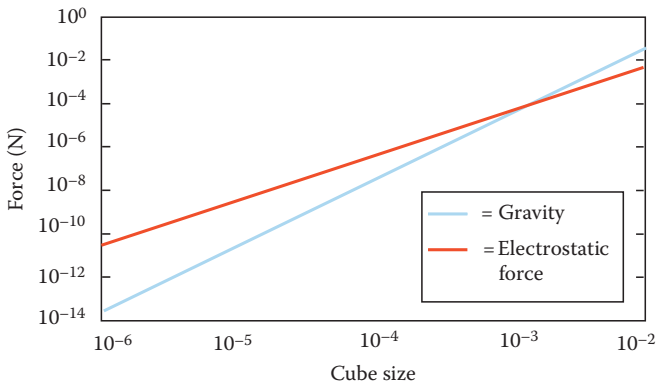
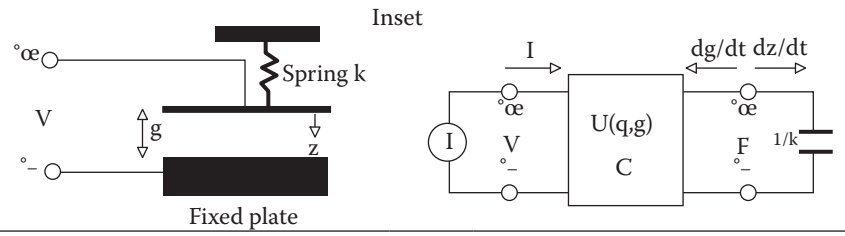


FIGURE 8.8 Competing gravitational and electrostatic forces working on a cube of length L, charge q, and density ρ (see also Figure 4.39).

TABLE 8.3 Equivalent Electrical and Mechanical Terms

Displacement, g	⇔	Charge, q
Force, F	⇔	Voltage, V
Velocity, dz/dt	⇔	Current, I (= dq/dt)
Mass, m	⇔	Inductance, L
Spring, k	⇔	Capacitance, 1/C
Damping, b (dashpot)	⇔	Resistance, R

Mechanical and electrical equivalent circuits.



Correspondence between Electrical and Mechanical Circuits

In this section we briefly introduce the correspondence between electrical and mechanical circuits.

For modeling of electrostatic actuators as shown in Figure 8.7, it is useful to make abstractions of the electrical and mechanical system elements as illustrated in Table 8.3. In this table we list the correspondence between the parameters in electrical and mechanical circuits. In the inset in this table, the gap of the electrostatic actuator is marked with the letter *g*, not to be confused with the gravitational constant. Movement of the electrostatically actuated structure with spring constant *k* is in the *z*-direction.

We will come back to modeling electrostatic actuators at the end of "Electrostatics."

Dynamics of Electrostatic Actuators

Dynamics is the branch of classical mechanics that is concerned with the motion of bodies. Here we consider the dynamics of an electrostatic actuator. In the electrostatic actuator shown in Figure 8.9, the electrostatic force, F_e (V) (see Equation 8.4, with x replaced by g, the instantaneous gap spacing, and ε_0 replaced by ε) works against a mechanical restraint $F_s = kz$. The latter is Hooke's law, with k the spring constant and z the displacement of the movable plate, $z = g - g_0$, with g_0 the initial gap spacing, i.e., the gap without applied voltage.

A stable equilibrium results when the electrostatic and mechanical force are equal in absolute magnitude but opposite in sign, i.e., $F_e = -F_s$, or as follows:

$$F_e = \frac{1}{2}\frac{\varepsilon A V^2}{g^2} = -kz = k(g_0 - g) \qquad (8.12)$$

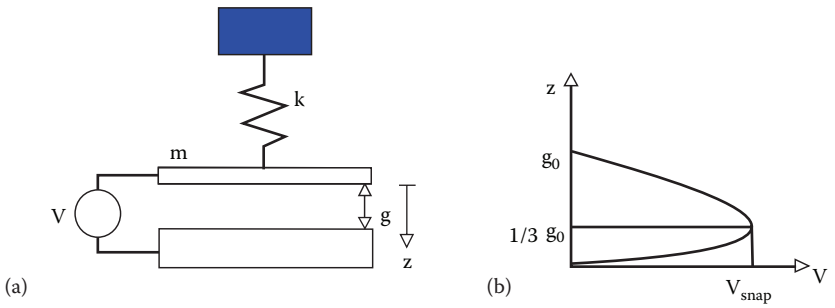


FIGURE 8.9 (a) Schematic of an electrostatic actuator. (b) At the snap voltage, V_{snap}, there is an electrostatic instability or a pull-in of the electrostatic actuator, where g_0 is the initial gap spacing and the snap voltage occurs at 1/3 of the initial gap spacing.

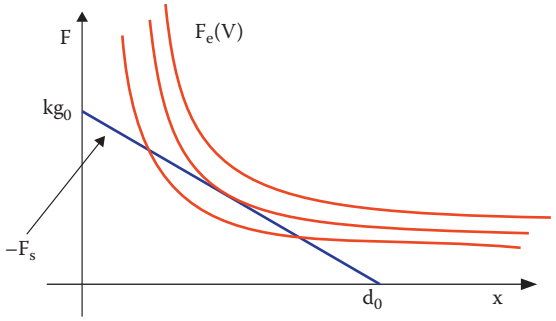


FIGURE 8.10 Positioning of a capacitor plate. The blue line is simply Hooke's law, and the three red lines are the electrostatic force for three different voltages.

This relationship is illustrated in Figure 8.10 for three different voltage values. Notice that the blue line (Hooke's law) intersects the lowest red F_e (V)-curve (electrical force) at two points and touches the next red F_e (V)-curve at only one point. These intersection points and the contact point are now analyzed further.

From Figure 8.10, we recognize that at $g = g_0$ ($z = 0$), when there is no voltage applied, $F_s = 0$. When the electrostatic actuator pull-in has shunted the capacitor plates, the gap $g = 0$ ($z = -g_0$), and the force $F_s = kg_0$.

On the basis of Equation 8.12, we derive the following voltage dependence on z:

$$V = \sqrt{\frac{2kz}{\varepsilon A}}(g_0 - z) \tag{8.13}$$

This is plotted in Figure 8.9b. When the applied voltage, V, is larger than V_{snap} (the snap or pull-in voltage), an unstable situation results (the system loses equilibrium) and the capacitor plates snap together. Let us analyze the conditions for this electrostatic instability in a bit more detail. An equilibrium point is unstable if in the presence of a small perturbation the net force does not tend to return the system back to the equilibrium position. The net force is the sum of the electrostatic and mechanical force, $F_{net} = F_e + F_s$, or, from Equation 8.12:

$$F_{net} = -\frac{\varepsilon AV^2}{2g^2} + k(g_0 - g) \tag{8.14}$$

For a stable equilibrium point, the variation of the net force (δF_{net}) should be of the opposite sign to the perturbation (δg), i.e., $\delta F_{net}/\delta g < 0$. We calculate ΔF_{net} as follows:

$$\delta F_{net} = \frac{\partial F_{net}}{\partial g}\bigg|_V \delta g = \left(\frac{\varepsilon AV^2}{g^3} - k\right)\delta g \tag{8.15}$$

When $\delta g > 0$, the system becomes unstable if $\delta F_{net} > 0$, and the system becomes stable if $\delta F_{net} < 0$, or for a stable equilibrium, one must have the following condition:

$$k > \frac{\varepsilon AV^2}{g^3} \tag{8.16}$$

At the electrostatic instability voltage, pull-in voltage or snap voltage, V_{snap} (the peak of the curve in Figure 8.9b: $\delta V/\delta z = 0$), the net force $F_{net} = 0$, and one obtains the following:

$$k = -\frac{\varepsilon AV_{snap}^2}{g_{snap}^3} \tag{8.17}$$

and:

$$F_{net} = -\frac{kg_{snap}}{2} + k(g_0 - g_{snap}) = 0 \tag{8.18}$$

From Equation 8.18, we derive for the value of g_{snap}:

$$g_{snap} = \frac{2}{3}g_0 \tag{8.19}$$

The value for V_{snap}, the electrostatic instability or snap voltage, is then as follows:

$$V_{snap} = \sqrt{\frac{8kg_0^3}{27\varepsilon A}} \tag{8.20}$$

Representing a normalized displacement as:

$$\zeta = \frac{z}{g_0} = 1 - \frac{g}{g_0} \tag{8.21}$$

And a normalized voltage as:

$$\nu = \frac{V}{V_{snap}} \tag{8.22}$$

The equilibrium condition can be represented as:

$$\frac{4}{27}\frac{\nu^2}{(1-\zeta)^2} = \zeta \tag{8.23}$$

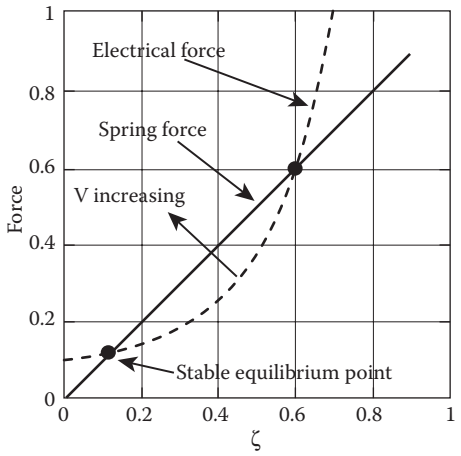


FIGURE 8.11 Force, *F*, as a function of ζ (with ν = 0.8). The spring force is marked by a solid line and the electrical force by an arched line (see also Figure 8.10).

The two intersection points merge when $\zeta = 1/3$ ($g_{snap} = 2/3g_0$) and $\nu = 1$ ($V = V_{snap}$), and this point corresponds to the pull-in point. As shown in Figure 8.10, this is the point where the blue line just touches the red curve. In Figure 8.11, we plot the force, *F*, as a function of ζ (with ν = 0.8). The spring force is marked by a solid line and the electrical force by an arched one.

Fundamental Frequency and Quality Factor of Electrostatic Resonators

Introduction

Two very important characteristics for electrostatic actuators are the fundamental frequency and the quality factor of the moving parts in these systems. The fundamental frequency of a resonating part is the frequency at which a free system vibrates. There are two different common definitions of the *Q* of a resonator. One is via the energy storage, where the *Q* is given as 2π times the ratio of the stored energy to the energy dissipated per oscillation cycle, or equivalently the ratio of the stored energy to the energy dissipated per radian of the oscillation. The other definition is via the resonance bandwidth in which the *Q* is the ratio of the resonance frequency f_0 and the full width at half-maximum (FWHM) bandwidth, Δf, of the resonance. Both definitions are equivalent only in the limit of weakly damped oscillations, i.e., for high *Q* values.

Fundamental Frequency

The fundamental or natural frequency, f_0, of a simple spring-mass system resonator is given by:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{8.24}$$

and:

$$\omega_0 = 2\pi f_0 = \sqrt{\frac{k}{m}} \tag{8.25}$$

with *k* the spring constant (defined as load/deflection), *m* the mass, and ω_0 the natural circular frequency. The spring constant *k* in the two preceding equations scales as *l*, and *m* scales as l^3; it follows that the resonance frequency scales as $f_0 \sim l^{-1}$, confirming that the smaller the resonator the higher its resonant frequency.

For a more complicated resonator, of the type shown in the header of this chapter, the same fundamental equation holds but with the spring constant *k* replaced by *K*, that is, the effective spring constant (also stiffness constant or stiffness matrix), and *m* replaced by *M*, that is, the effective mass or mass matrix of the structure. The stiffness matrix varies with geometry as well as with the induced stresses.

With reference to Table 8.3, we recognize that just as a spring-mass combination has a natural (resonant) frequency determined by the stiffness of the spring and the magnitude of the mass, so too do the values of an inductor and capacitor determine the natural oscillating frequency of an electrical inductance-capacitance (LC) circuit. The larger the inductance, the longer it takes a given moving charge to set up a magnetic field. Since a given charge produces only a small voltage across a large capacitor, the larger the capacitance the longer it will take to charge and discharge that capacitor. The natural oscillating frequency of an LC circuit is given by the following expressions:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{1}{LC}} \tag{8.26}$$

and:

$$\omega_0 = 2\pi f_0 = \sqrt{\frac{1}{LC}} \tag{8.27}$$

Quality Factor

The Q of a resonator, in one of the two prevailing definitions (see introduction above), is given as the ratio of stored over lost energy in one cycle.

$$Q = 2\pi \frac{E}{\Delta E} = 2\pi \frac{\tau}{T} = \tau\omega_0 \tag{8.28}$$

For a mechanical system as sketched in the inset in Table 8.3, the mechanical quality factor is then calculated as follows:

$$Q = \tau\omega_0 = \frac{m}{b}\omega_0 = \frac{\sqrt{km}}{b} \tag{8.29}$$

with b the damping in the system. From the discussion in the introduction above, m and k should be replaced by M and K for more complex resonator structures.

Analogously, for an electrical system, the quality factor is given by the following:

$$Q = \tau\omega_0 = \frac{L}{R}\omega_0 = \frac{1}{R}\sqrt{\frac{L}{C}} \tag{8.30}$$

with L the inductance and C the capacitance (see Table 8.3).

Typical mechanical quality factors for laterally driven polysilicon resonant microstructures (see figure in the header of this chapter) are about 100 or less.[8] A poly-SiC lateral resonator, a material with a high Young's modules and thus a high k, at $<10^{-5}$ Torr, has been micromachined and features Q values >100,000.[13] In a good vacuum, b in Equation 8.29 becomes very small, further improving the factor Q. For NEMS devices, Qs attained today in moderate vacuum are in the range of 10^3 to 10^5. This greatly exceeds typical Qs from electrical resonators. A high Q is linked to small internal dissipation ($D = 1/Q$), imparting, in turn, low operating power levels and high attainable force sensitivity to nanodevices.[14]

Merits of Electrostatic Devices

To analyze the merits of electrostatic actuators, we start by deriving the maximum energy density one can store in such an actuator. In Chapter 7, on scaling, we saw that the maximum electrostatic energy density, U'_m (in J/m^3), between the capacitor plates of an electrostatic actuator is obtained by dividing the maximum potential energy, U_m, by the volume, V_0, of the capacitor gap, $V_0 = w \times v \times d$ (see Figure 7.11):

$$U'_m = \frac{\varepsilon_r \varepsilon_0 E_b^2}{2} \tag{7.36}$$

where E_b represents the highest attainable electric field in the actuator, i.e., the electric breakdown field strength. For an air capacitor, E_b is limited to approximately 3×10^6 Vm^{-1} by the electrical breakdown of air. Thus, the stored energy for an air capacitor is about 40 Jm^{-3}. Electrostatics is a surface force, and the surface force density, F', is given by the following (derivative of Equation 8.4):

$$F' = \frac{\partial F}{\partial A} = \frac{\varepsilon_r \varepsilon_0 E^2}{2} \tag{8.31}$$

where A stands for the surface area. The surface force density equals the energy density in the field. Hence, one wants to use the maximum field possible for the largest possible force. In Chapter 7, we saw that this force scales as l^2, which is favorable in the microdomain where forces with the lowest exponent win out (e.g., electrostatics with l^2 win out over the l^4 scaling of magnetic actuators).

Our reasoning so far involves isotropic electrical fields only. In Chapter 7 we saw that besides favorable scaling, the Paschen effect represents another advantage for microelectrostatic systems. Working on the left side of the Paschen curve (see Figure 7.12) renders very high-field operation possible for a wide variety of electrostatic microdevices without incurring catastrophic sparking. Examples include the electrostatic motors illustrated in Figure 8.6b, the electrostatic switch in Figure 8.13, and the author's volcano ionization sources presented in Example 7.1 at the end of Chapter 7 (Figure 7.14). The field in the nonisotropic Paschen regime scales more nearly as $l^{-1/2}$, whereas the force scales as l^1. With a factor of 10 reduction in size, the inertial force still decreases by a factor of 1000, but the electrostatic force decreases only by a factor of 10. Acceleration and transit times are also higher than for the isotropic system. Superficially, this seems to put magnetic motors, based on volume forces, at a disadvantage compared with electrostatic micromotors. We shall see, however, that this is not the end of the story. For example, magnetic field energy can be made two

orders of magnitude greater than the best one can achieve with electrostatic fields and small air gaps in air or vacuum.[15]

We are interested in using electrostatic force in all types of actuators: not only simple parallel plate capacitors but also linear and rotational motors. As we saw earlier, the force, *F*, is the negative spatial derivative of the energy, *U*, in the system (Equation 8.3). In the case we move the plate of a capacitor with respect to another plate, the electric force will be the spatial derivative, in that direction, of the stored energy *U*. The translational movement of one plate with respect to the other in a linear motor moving in the *x*-direction leads to a force, F_x, as follows:

$$F_x = -\frac{\partial U}{\partial x} = \frac{V^2}{2}\frac{\partial C}{\partial x} = \frac{1^3}{1^1} = 1^2 \qquad (7.37)$$

where *V* is the applied voltage. The electrostatic force for a constant field is thus found to scale as l^2. This is often an advantage, because the mass and, hence, inertial forces scale as l^3. The latter point was clearly illustrated in Figure 8.8.

From Equation 7.37 translational motor action is associated with a change in the capacitance: an increase in capacitance for a motor, a decrease in capacitance for a generator. For a rotational motor a rotational force, F_r, by analogy to Equation 7.37 is defined as follows:

$$F_r = -\frac{\partial U}{\partial \theta} \qquad (8.32)$$

where $\partial\theta$ is the differential angular displacement in radians. From Equations 7.37 and 8.32, F_r can be expressed in terms of the change in capacitance.

$$F_r = -\frac{V^2}{2}\frac{\partial C}{\partial \theta} = -\frac{q^2}{2C^2}\frac{\partial C}{\partial \theta} \qquad (8.33)$$

The second expression on the right is obtained by introducing the relation $V = qC^{-1}$, where *q* is the total charge on the capacitor plates. Detailed calculations of the capacitance of electrostatic motors/generators can be found in Mahadevan[16,17] and in Kumar et al.[18] Torque of a rotational motor is given by the following:

$$T = r \times F_r \qquad (8.34)$$

With *r* the radius and F_r perpendicular, Equation 8.34 simply becomes $T = r \cdot F_r$. The power, *P*, generated by the motor can easily be calculated from the torque as follows:

$$P = T\omega = T2\pi f \qquad (8.35)$$

where ω represents the angular frequency and *f* the frequency at which the motor rotates. From Equation 8.35, one would expect that as electrostatic motors decrease, the inherent increase in frequency could help to offset the decrease in torque. However, static and dynamic frictional forces (surface forces!) come into play and are a major barrier. Despite these problems, surface-micromachined electrostatic motors of the type shown in Figure 8.6b, with rotational speeds of 15,000 rpm and continuous operation for over a week, have proven to be possible at voltages below 300 V.[19] Friction-reduction methods include deposition of a silicon nitride sliding surface; electrostatic, magnetic, or other types of levitation; and replacing sliding contacts with rolling contacts in wobble or harmonic motors.[20,21]

In harmonic motors, the rotor rolls on the inside of the stator without slipping and rotates slightly in the process (Figure 8.12). It is this slight rotation

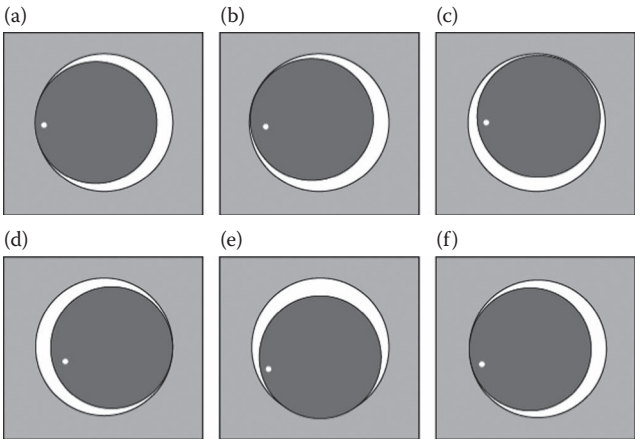


FIGURE 8.12 Wobble or harmonic motors. (a) The rotor is touching the edge of the hole to the left. The white dot marks a point on the rotor; (b)–(f) show the rotor as it progressively rolls around the hole in a clockwise manner. Note the position of the white dot in (f) after one rotation. The rotor has twisted about its axis slightly in a counterclockwise direction. As the rotor rolls repeatedly around the hole, the twisting motion produces an output torque.

of the rotor that produces the output motion of the motor. Applying voltages between the rotor and electrodes in the stator hole electrostatically drives the rotor. Each of the rotations of the rotor about the stator hole produces another small twist of the rotor axis, producing a beneficial gear reduction. Advantages of this type of motor are the rolling motion, which avoids sliding friction, the closeness of the rotor and stator, enabling a higher electrostatic energy density, and the gear reduction, producing a large torque. Harmonic motors do not scale up well, however, and Trimmer compares large wobble motors to an unbalanced washing machine.[22] In the microdomain, inertial effects decrease, and these unbalanced forces are insignificant for microharmonic motors.

To drive a load several hundred micrometers in thickness and width and several millimeters in diameter, the required torque of a motor is of the order of 10^{-5} Nm. However, the torque generated with a typical surface-micromachined electrostatic motor, 100–150 µm in diameter and with an air gap of about 2 µm, is only in the nano-newton-meter or pico-newton-meter range (see Figure 8.6b).[21] Calculation for an outer rotor, surface-micromachined wobble motor,* with copper electroplated structural elements, indicates that to achieve a torque of 10^{-4} to 10^{-5} Nm, the stator radius must be a few millimeters, and the thickness needs to be of the order of 10 µm.[24] As frequently observed, actuators do not scale advantageously in the microdomain. Torque is a volume effect, and small-size devices have a limited range of force available for actuation. In view of this recognition, perhaps an excessive amount of research has gone into surface-micromachined electrostatic motors. These flat surface micromachined micromotors could still find a use, though, as shutters for charged particle or photon beams, memory writing/reading devices, masks with movable parts, etc.—in other words for applications in which the required forces are minimal. Below, we will compare electrostatic motors with the more complicated and typically larger micromachined magnetic motors that, it appears, can perform a broader range of work more readily.

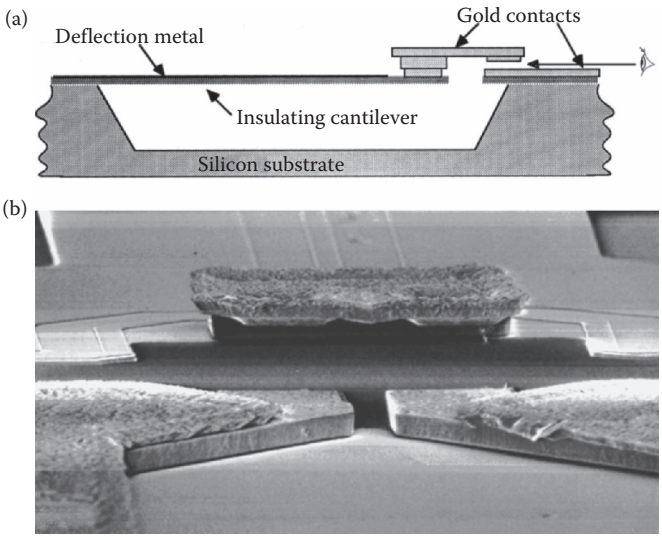


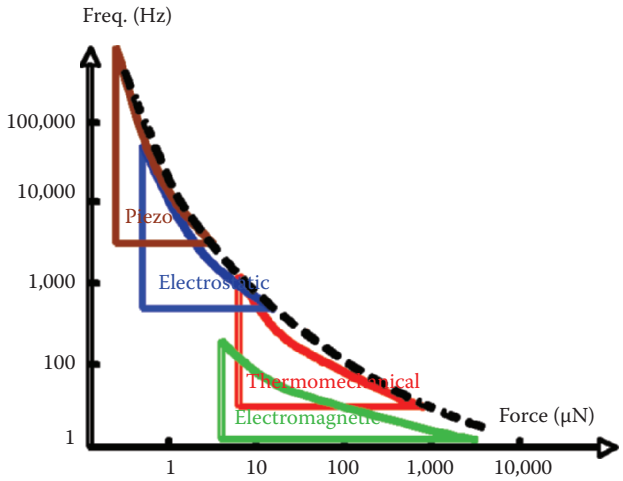
FIGURE 8.13 Electrostatic microswitch. (a) Schematic; (b) Scanning electron microscope micrograph, looking straight into the opened contact. Switch built by the author's research group.

Besides different types of monolithic electrostatic micromotors,[19,25] electrostatic actuators have been used for hybrid mounted micromotors;[26] microvalves;[28] mechanical resonators for use in, for example, gyros and accelerometers;[27] displacement actuators for optical components (e.g., for positioning micromirrors);[28] and switches.[29] An electrostatic microswitch developed by this author is illustrated in Figure 8.13. Upon applying a bias between the cantilever and the substrate, two metal gold pads make physical contact enabling large currents to flow. A mechanical switch as shown has advantages over an electronic switch, as its off-impedance (air gap) is infinite and its on-impedance (metal contact) is very low.

Summary Electrostatic Actuators: Performance and Modeling

Electrostatic fields can exert great forces, but generally only across very short distances (1 µm). The extremely low current consumption associated with electrostatic devices makes for highly efficient actuation and the high speeds (up to 100 kHz) and ease of fabrication of electrostatic devices has enabled many applications (see above). In Figure 8.14, where

* A detailed description of an electrostatic eccentric drive micromotor (wobble motor) can be found in Price et al.[23]



1. Stroke: 1 μm
2. Maximum speed: 100 kHz
3. Intermediate force density
4. Easy to fabricate
5. Voltage: ~100 V
6. Footprint: 10 μm to 1 mm
7. Maximum electric field before breakdown (approximately 3×10^8 Vm^{-1})
8. Maximum electrostatic energy density, U_m' (in J/m^3): 4.0×10^5 Jm^{-3}

FIGURE 8.14 Summary of electrostatic actuators.

we plot frequency versus force, we summarize some of the most salient electrostatic actuator characteristics in comparison to actuators based on other principles.

In Figure 8.15 we generalize the description of the dynamics of an electromechanical system. We employ x as the direction of displacement (instead of z; see above) and K for a combined stiffness constant (instead of k; see above). The differential equation, describing the electrical equivalent circuit shown here (Figure 8.15b), is as follows:

$$L\frac{d}{dt}\left(\frac{dQ}{dt}\right)+R\left(\frac{dQ}{dt}\right)+\frac{1}{C}Q(t)=V_i$$

and with $$\frac{d}{dt}\left(\frac{dQ}{dt}\right)=\ddot{Q}(t) \quad \text{and} \quad \frac{dx}{dt}=\dot{Q}(t)$$

$$L\ddot{Q}(t)+R\dot{Q}(t)+\frac{1}{C}Q(t)=V_i \qquad (8.36)$$

where V_i is the input voltage and V_0 the output voltage. L is inductance, R is resistance, and C is capacitance. For the mechanical equivalent circuit (Figure 8.15a), the differential equation is as follows:

$$m\frac{d}{dt}\left(\frac{dx}{dt}\right)+b\left(\frac{dx}{dt}\right)+Kx(t)=F$$

with $$\frac{d}{dt}\left(\frac{dx}{dt}\right)=\ddot{x}(t) \quad \text{and} \quad \left(\frac{dx}{dt}\right)=\dot{x}(t)$$

$$m\ddot{x}(t)+b\dot{x}(t)+Kx(t)=F \qquad (8.37)$$

where the external force F is the input, the displacement x is the output, b is damping, m is the mass, and K is the stiffness.

The modeling equations gathered above are in the time domain, but often analysis of a system is easier

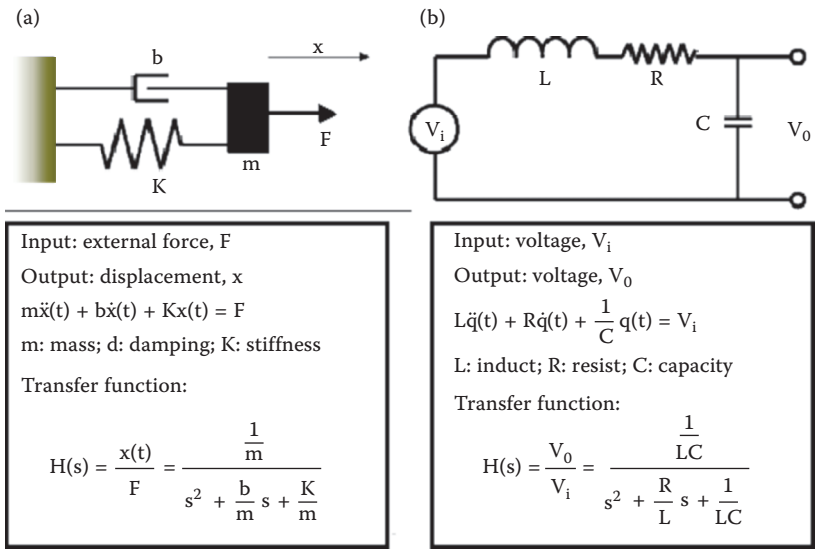


FIGURE 8.15 General description of the dynamics of an electromechanical system.

to perform in the frequency domain. In order to convert these equations to the frequency domain, we apply the Laplace transform to determine the transfer function [$H(s)$] of the system. As a consequence functions with a real dependent variable (such as time) are converted into functions with a complex dependent variable (such as frequency, often represented by s), which are easier to manipulate. The transfer function of the system then is the ratio of the output Laplace transform to the input Laplace transform assuming zero initial conditions. Many important characteristics of dynamic or control systems can be determined from the transfer function. The general procedure to find the transfer function of a linear differential equation from input to output is to take the Laplace transforms of both sides assuming zero conditions and to solve for the ratio of the output Laplace over the input Laplace. For the electrical case the transfer function is as follows:

$$H(s) = \frac{V_0}{V_i} = \frac{\frac{1}{LC}}{s^2 + \frac{R}{L}s + \frac{1}{LC}} \qquad (8.38)$$

And for the mechanical case, it is as follows:

$$H(s) = \frac{x(t)}{F} = \frac{\frac{1}{m}}{s^2 + \frac{b}{m}s + \frac{K}{m}} \qquad (8.39)$$

Piezoelectricity

Introduction

In 1880, Pierre Curie and his brother Paul-Jacques discovered that external forces applied to single crystals of quartz and some other minerals (e.g., Rochelle salt) generate a charge on the surface of these crystals. The charge, they found, is roughly proportional to the applied mechanical stress (force per unit area). This is called the *direct piezoelectric effect*. The effect received its name in 1881 from Wilhelm Hankel, and it remained a curiosity until 1921, when Walter Cady discovered that a quartz resonator could be used for stabilizing electronic oscillators.[30] Piezoelectric materials exhibit the inverse effects as well: an applied voltage generates

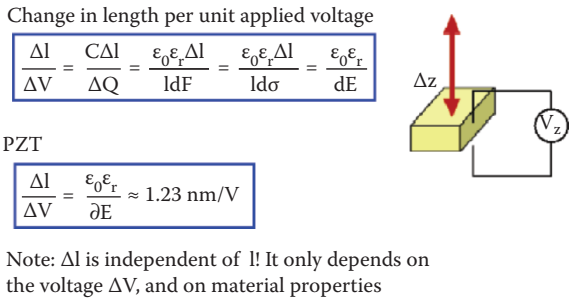


FIGURE 8.16 The inverse piezoelectric effect. Used to build actuators.

a deformation of the crystal (Figure 8.16). Lippman predicted this effect from thermodynamic principles, and the Curie brothers confirmed it experimentally a year later (1881). In the direct case, the crystals work like sensors; in the indirect case, piezoelectrics exhibit an actuation behavior. The mechanical waves generated by piezoelectrics are also called *acoustic waves*. Acoustic wave devices are so named because their detection mechanism is a mechanical, or acoustic, wave. Piezoelectric acoustic wave sensors apply an oscillating electric field to create a mechanical wave, which propagates through the substrate and is then converted back to an electric field for measurement. As the acoustic wave propagates through or on the surface of the piezoelectric material, any changes to the characteristics of the propagation path affect the velocity and/or amplitude of the wave. Changes in velocity can be monitored by measuring the frequency or phase characteristics of the sensor and can then be correlated to the corresponding physical quantity being measured. Virtually all acoustic wave devices and sensors use a piezoelectric material to generate the acoustic waves.

Piezoelectric actuators follow mammalian muscle as the most ubiquitous actuator principle in nature. For example, bones possess piezoelectric properties.[31] Other biological materials that have been found to be piezoelectric include tendon, dentin, ivory, intestine, silk, elastin, wood, and the nucleic acids.

Piezoelectricity is not to be confused with ferroelectricity, discovered in 1920 by J. Valasek. Ferroelectricity is the property of a spontaneous or induced electric dipole moment. In crystals of Rochelle salt ($NaKC_4H_4O_6 \cdot 4H_2O$), the unit cells

have a spontaneous asymmetric charge distribution, and the orientation of this polarization can be switched by applying a field strong enough to drag the unsymmetrically placed ions into equivalent positions in the opposite direction. All ferroelectric materials are piezoelectric, but the converse is not always true. Piezoelectricity relates to the crystalline ionic structure; ferroelectricity relates to electron spin. Above the so-called *Curie temperature*, spontaneous polarization of a ferroelectric is lost, because thermal vibrations randomize the dipole orientations. The transition from a nonferroelectric to a ferroelectric state is analogous to the magnetic ordering transition in a ferromagnet below its Curie temperature. Ferroic materials are multifunctional materials, materials having both sensing and actuation functions. A large number of applications of ferroelectric materials exploit an indirect consequence of ferroelectricity, such as dielectric, piezoelectric, pyroelectric, and electrooptic properties. The biggest use of ferroelectric materials is as dielectrics in capacitors, ferroelectrics for thin films for nonvolatile memories, piezoelectric materials for medical ultrasound and imagining and actuators, and electrooptical materials for data storage and displays. Ferroelectric materials exhibit strong electrostriction.

Electrostriction, covered below, is a secondary coupling in which the strain is proportional to square of the electric field. Electrostriction, a property of all dielectrics but pronounced only for ferroelectrics, is similar to the piezoelectric effect in that it involves an increase in length parallel to an applied electric field. In electrostrictive materials, in contrast to piezoelectric materials, the direction of this small change in geometry does not reverse if the direction of the electrical field is reversed.

Mechanism of Piezoelectricity

A simplified model of piezoelectricity entails the notion of anions (–) and cations (+) moving in opposite directions under the influence of an electric field or a mechanical force. The forces generated by this motion cause lattice deformation for noncentrosymmetric crystals because of the presence of both high- and low-stiffness ionic bonds. The effect for quartz is illustrated in Figure 8.17. If the cell shown here is deformed along the x- or y-axis, the O-ion is displaced, and positive or negative charges are formed. As a result, all piezoelectric materials are necessarily anisotropic; in cases of central symmetry, an applied force does not yield an electric polarization. By applying mechanical deformations to piezoelectric crystals, electric dipoles are generated, and a potential difference develops when changing those mechanical deformations. Silicon is not piezoelectric because it is cubic and covalent rather than

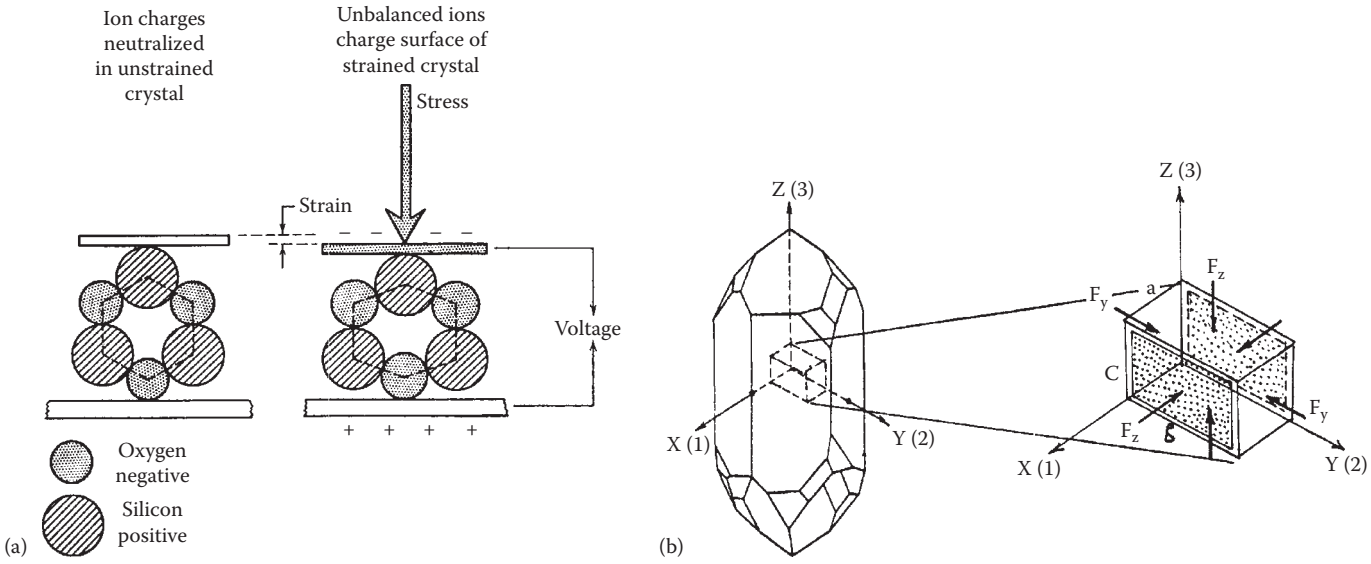


FIGURE 8.17 Piezoelectricity in an ionic crystal such as quartz. (a) Ion position in quartz lattice with and without applied stress. (b) Element cut from quartz crystal under stress.

ionic and noncentrosymmetric. It is important to remember that the potential and the associated currents in piezoelectric materials are a function of the continuously changing mechanical deformation. Therefore, typical and practical uses are in situations involving dynamic strains of an oscillatory nature.[32]

As we saw in Volume I, Chapter 2, the thousands of crystals found in nature can be grouped into 230 space groups on the basis of symmetry elements. Since the relative position of the symmetry elements in a crystal is not important, and only their orientation counts, the 230 space groups can be reduced to just thirty-two point groups. The thirty-two point groups further divide over the seven well-known crystal systems: triclinic, monoclinic, orthorhombic, tertragonal, trigonal, hexagonal, and cubic. Of the thirty-two point groups, twenty-one do not have a center of symmetry (i.e., they are noncentrosymmetric). The latter crystals possess one or more crystallographically unique directional axes, and all, except the 432 point group, show the piezoelectric effect along unique directional axes. Piezoelectric properties are thus present in twenty of the thirty-two different crystallographic point groups, although only a few of them are used. They are found also in amorphous ferroelectric materials. Of the twenty crystallographic classes, only ten display ferroelectric properties.

A more detailed understanding of piezoelectricity is based on an understanding of the piezoelectric equations describing the coupling between electric and mechanical strains in a piezoelectric material. Polarization and stress are vector and tensor properties, respectively, and in general, arbitrary components of each can be related via the piezoelectric effect. For this reason, piezoelectricity is a complicated property, and up to eighteen constants may be required to specify it. When a stress, σ (FA^{-1}), is applied to a slab of material cut from a piezoelectric material such as quartz, the resulting one-dimensional strain, ε, in the elastic range, can be written as (see Volume I, Chapter 4, "Stress-Strain Curve and Elasticity Constants"):

$$\varepsilon = S\sigma \tag{8.40}$$

which represents Hooke's law, where *S* stands for the compliance (S^{-1} = E = Young's modulus). With a potential difference applied across the faces of the piezoelectric, an electric field, *E* (Vm^{-1}), is established and we obtain the following (see Volume I, Equation 3.37):

$$D = \varepsilon_r E = \varepsilon_0 E(1 + \chi) = P \tag{8.41}$$

where D = electric displacement (or electric flux density)
ε_r = relative permittivity of the material (dimensionless)
E = electric field
ε_0 = permittivity of vacuum in $A \times s\ V^{-1} \times m^{-1}$
χ = electrical susceptibility
P = polarization, that is, the electric dipole moment per unit volume of material

Equation 8.40 contains only mechanical quantities, and Equation 8.41 contains only electrical parameters. Piezoelectric materials possess a special interlocking behavior in which electrical charges are produced by straining the material, and mechanical forces are produced by subjecting the material to an electric field.[33] For a one-dimensional piezoelectric material with electric field and stress in the same direction, according to the principle of energy conservation from thermodynamics, we can describe this situation as follows:[34]

$$D = d\sigma + \varepsilon(d)^{\sigma} E \tag{8.42}$$

$$D = e\varepsilon + \varepsilon(d)^{\varepsilon} E \tag{8.43}$$

where *d* represents a piezoelectric constant (charge density/applied stress) expressed in coulombs divided by newtons (C/N), and $\varepsilon(d)\sigma$ is the permittivity at constant stress [we added a "(*d*)" to distinguish the dielectric constant from the strain ε]. To work with *d* coefficients is useful when charge generators are contemplated, for example, in piezoelectric accelerometers. In the second expression, *e* stands for a piezoelectric constant (charge density/applied strain) in Cm^{-1}, and $\varepsilon(d)^{\varepsilon}$ is the dielectric constant at a constant strain. Solving for *E*, we can rewrite Equations 8.42 and 8.43 as follows:

$$E = -g\sigma + \frac{D}{\varepsilon(d)^{\sigma}} \tag{8.44}$$

$$E = -h\varepsilon + \frac{D}{\varepsilon(d)^{\varepsilon}} \tag{8.45}$$

where $g = d/\varepsilon\ (d)^{\sigma}$ is a piezoelectric constant (field/applied stress), also called a *voltage constant*, in V/m/N, and $h = e/\varepsilon\ (d)^{e}$ is a piezoelectric constant (field/applied strain) in V/m. The above set of equations describes the direct piezoelectric effect. The equations for the inverse effect are written as follows:

$$\varepsilon = dE + S^{E}\sigma \tag{8.46}$$

$$\varepsilon = gD + S^{D}\sigma \tag{8.47}$$

where d and g = piezoelectric constants given in mV^{-1} and mC^{-1}, respectively
S^{E} = compliance at constant field
S^{D} = the compliance at constant electric flux density

Solving Equations 8.46 and 8.47 for σ results in the following:

$$\sigma = -eE + E(Y)^{E}\varepsilon \tag{8.48}$$

$$\sigma = -hD + E(Y)^{D}\varepsilon \tag{8.49}$$

where $e = d/S^{E}$ and $h = g/S^{D}$ = piezoelectric constants given in N/V/m and N/C, respectively
$E(Y)^{E}$ = Young's modulus [we added a "(*Y*)" here to distinguish the Young's modulus from the electric field *E*] under constant electrical field
$E(Y)^{D}$ = Young's modulus under constant electric flux density

Equations 8.42 to 8.49 are known as the *piezoelectric constitutive relations*. They are summarized in Table 8.4, together with the definition of the four piezoelectric constants *d*, *e*, *g*, and *h*.

If no external field is imposed, according to Equations 8.41, 8.42, and 8.43, a stress, σ, will lead to the following polarization, *P*, or direct effect:

$$P = d\sigma \tag{8.50}$$

and, according to Equations 8.46 and 8.47, if there is no external stress applied, a field, *E*, will lead to the following strain and converse effect, ε:

$$\varepsilon = dE \tag{8.51}$$

so that the same constant—the piezoelectric coefficient *d*—is used for the direct and reverse effects.

The ratio of the converted energy of one kind (mechanical or electrical) stored at any instant in a piezoelectric to the input energy of the second kind (electrical or mechanical) is defined as the square of the coupling coefficient:

$$k = \sqrt{\frac{\text{mechanical energy stored}}{\text{electrical energy applied}}} = \sqrt{\frac{\text{electrical energy stored}}{\text{mechanical energy applied}}} \tag{8.52}$$

The electromechanical coupling coefficient, *k*, is thus a measure of the interchange of electrical and mechanical energy. It can be shown that *k* equals the geometric mean of the piezoelectric voltage coefficient (*g*) and the piezoelectric stress coefficient (*e*) and is indicative of the ability of a material to both detect and generate mechanical vibrations:[35]

$$k = \sqrt{ge} \tag{8.53}$$

To activate the maximum piezoelectric strain in a given crystal, the orientation dependence of the piezoelectric effect should be carefully considered.

Table 8.4 Piezoelectric Constitutive Equations and Definitions of Piezoelectric Parameters

Piezoelectric Equations	Equation Number in Text	Definitions of Constants	SI Units
$D = d\sigma + \varepsilon(d)\sigma E$	8.42	d = charge density/applied stress	C/N
$D = e\varepsilon + \varepsilon(d)\varepsilon E$	8.43	e = charge density/applied strain	C/m
$E = -g\sigma + D/\varepsilon(d)\sigma$	8.44	g = field/applied stress	V/m/N
$E = -h\varepsilon + D/\varepsilon(d)\varepsilon$	8.45	h = field/applied strain	V/m
$\varepsilon = dE + S^{E}\sigma$	8.46	d = strain/applied field	m/V
$\varepsilon = gD + S^{D}\sigma$	8.47	g = strain/applied charge density	m/C
$\sigma = -eE + E(Y)^{E}\varepsilon$	8.48	e = stress/applied field	N/V/m
$\sigma = -hD + E(Y)^{D}\varepsilon$	8.49	h = stress/applied charge density	N/C

For the general case of a piezoelectric crystal, rather than a one-dimensional piezoelectric as treated above, the piezoelectric constitutive relations must be generalized. The generalized elastic response of a piezoelectric crystal to an electric field (see Equation 8.51) may then be expressed as follows:

$$\varepsilon_k = \sum_{i=1}^{3} \mathbf{d}_{ik}\mathbf{E}_i \tag{8.54}$$

and the polarization in the absence of a field, from Equation 8.50, as follows:

$$\mathbf{P}_i = \sum_{k=1}^{6} \mathbf{d}_{ik}\sigma_k \tag{8.55}$$

In these equations, i = 1, 2, and 3 make up the indices of the components of polarization and k = 1, 2, …, 6 the indices of the components of mechanical stress and strain. In the piezoelectric constants, the first subscript refers to the direction of the field; the second subscript refers to the direction of the stress. The subscripts 1, 2, and 3 indicate the *x*-, *y*-, and *z*-axes, respectively. For example, in d_{33}, i = 3 indicates that the polarizing electrodes are perpendicular to the 3-axis (i.e., *z*-axis), and the piezoelectric-induced strain or applied stress is in the 3-direction (along the *z*-axis) as well. For the coupling coefficient, *k* (Equation 8.53), the same notation is used. The above expressions are equivalent to writing (strain tensor) = (d tensor) × (electric field vector) and (polarization vector) = (d tensor) × (stress tensor), respectively. Piezoelectricity, we saw in Volume I, Chapter 2, on crystallography, is an example of a vector-tensor effect; an electric field (vector) causes a mechanical deformation (tensor). Elastic deformation under the influence of a stress tensor is an example of a tensor-tensor effect.

On the basis of this notation, we can now generalize the constitutive relations as:

$$\mathbf{D}_i = \mathbf{d}_{ik}\sigma_\kappa + \varepsilon(\mathbf{d})_{ik}\sigma\mathbf{E}_i \tag{8.56}$$

and:

$$\varepsilon_k = \mathbf{d}_{ik}\mathbf{E}_i + \mathbf{S}^{E}_{ik}\sigma_k \tag{8.57}$$

where $\mathbf{D}_i$ = electric displacement vector (3 × 1), C/m²
$\mathbf{E}_i$ = electric field vector (3 × 1), V/m
σ_i = stress (force/unit area) tensor (6 × 1), N/m²
ε_k = strain (relative displacement) tensor (6 × 1)
$\mathbf{d}_{ik}$ = a matrix of piezoelectric constants (3 × 6), m/V
$\varepsilon(\mathbf{d})_{ik}$ = permittivity measured at constant stress (3 × 3), As/Vm
$\mathbf{S}^{E}_{ik}$ = elastic compliance matrix when subjected to a constant electric field (6 × 6), m²/N

The other equations in Table 8.4 may be generalized in the same fashion.

From Volume I, Equation 4.83, we recall that the strain tensor has six components: σ_x, σ_y, σ_z for compression or tension and τ_{xy}, τ_{xz}, and τ_{yz} for the shear components. The piezoelectric behavior of a crystal can be completely described if the set of data related to the piezoelectric constants (d_{ik}), elastic compliances (S_{ik}), and permittivities [$\varepsilon(d)^{\sigma}_{ik}$] is given. This set can be introduced with a 9 × 9 matrix in which columns are associated with mechanical stresses and field strengths and rows refer to strains and polarizations:[36]

	σ_x	σ_y	σ_z	τ_{xy}	τ_{xz}	τ_{yz}	E_x	E_y	E_z
ε_x	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}	d_{11}	d_{21}	d_{31}
ε_y	S_{21}	S_{22}	S_{23}	S_{24}	S_{25}	S_{26}	d_{12}	d_{22}	d_{32}
ε_z	S_{31}	S_{32}	S_{33}	S_{34}	S_{35}	S_{36}	d_{13}	d_{23}	d_{33}
γ_{xy}	S_{41}	S_{42}	S_{43}	S_{44}	S_{45}	S_{46}	d_{14}	d_{24}	d_{34}
γ_{xz}	S_{51}	S_{52}	S_{53}	S_{54}	S_{55}	S_{56}	d_{15}	d_{25}	d_{35}
γ_{yz}	S_{61}	S_{62}	S_{63}	S_{64}	S_{65}	S_{66}	d_{16}	d_{26}	d_{36}
P_x	d_{11}	d_{12}	d_{13}	d_{14}	d_{15}	d_{16}	ε_{11}	ε_{12}	ε_{13}
P_y	d_{21}	d_{22}	d_{23}	d_{24}	d_{25}	d_{26}	ε_{21}	ε_{22}	ε_{23}
P_z	d_{31}	d_{32}	d_{33}	d_{34}	d_{35}	d_{36}	ε_{31}	ε_{32}	ε_{33}

(8.58)

This matrix is symmetrical ($S_{ik} = S_{ki}$, $d_{ik} = d_{ki}$, and $\varepsilon_{ik} = \varepsilon_{ki}$) and reduces to forty-five different terms, including twenty-one compliances, six permittivities, and eighteen piezoelectric constants. In the top left of this matrix (rectangular box), we recognize Equation 4.83, but because of the presence of piezoelectricity, the matrix is this much larger. For simplicity, we have replaced $\varepsilon(d)^{\sigma}_{ik}$ by ε_{ik} in the above matrix. A further simplified notation with one subscript is often introduced for dielectric constants: $\varepsilon_{11} = \varepsilon_1$, $\varepsilon_{22} = \varepsilon_2$, and $\varepsilon_{33} = \varepsilon_3$. For the defined orientation of the crystallographic axes, $\varepsilon_{ik} = 0$ for $i \neq k$.

For crystalline quartz, the eighteen remaining piezoelectric constants further reduce to two:

$$d_{ik} = \begin{bmatrix} d_{11} & d_{12} & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{25} & d_{26} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (8.59)$$

with $d_{11} = -d_{12} = -d_{26}/2 = 2.31$ pC/N and $d_{14} = -d_{25} = 0.73$ pC/N. The meaning of d_{14} is as follows: a torsion stress of 1 N/m^2 around the axis 1 (x) of quartz (direction 4, or xz) induces a charge density of 0.73 pC/N in two metal plates on the material in direction 1 (x).

Exercise 8.1:

A piezoelectric ceramic of 1 cm thickness in the 3-direction and 10 cm length in the 1-direction is subjected to a potential of 0.01 volts in the 3-direction, and the resulting change in length is 10^{-11} cm. What is the numerical value and subscript of the d constant? Answer: d_{31} = strain developed/field applied = (10^{-11} cm/10 cm)/(0.01 V/0.01 m) = 10^{-12} mV^{-1}.

Exercise 8.2:

If a voltage V is applied across the thickness of a piezoelectric slab, the displacements ΔL, ΔW, and Δt along the length, width, and thickness directions, respectively, are given by $\Delta L = d_{31} \times V \times L/t$, $\Delta W = d_{31} \times V \times W/t$, and $\Delta t = d_{33} \times V$, where L and W are the length and width of the plate, respectively, and t is the thickness or separation between the electrodes. When a force F is applied along the length, width, or thickness direction of the same slab, a voltage V is measured, given in each of the three cases as $V = d_{31} \times F/\varepsilon \times L$, $V = d_{31} \times F/\varepsilon \times W$, and $V = d_{33} \times F \times t/\varepsilon \times L \times W$.

At frequencies far below the resonance, piezoelectric transducers are fundamentally capacitors. Since, in a capacitor, the voltage V is related to the charge Q via the capacitance C ($Q = CV$), the charge coefficients d_{ik} are related in the same way to the voltage coefficients g_{ik} via ε_{ik}, or:

$$d_{ik} = \varepsilon (d)^{\sigma}_{ik} . g_{ik} . \varepsilon_0 \quad (8.60)$$

with ε_0 being the permittivity of free space. At resonance, the dielectric constant is reduced by a factor of (1 – k), where k is the coupling coefficient of the mode in question.

For an in-depth mathematical treatise of piezoelectricity, we refer to *The Theory of Piezoelectric Shells and Plates*.[37] Other excellent references are by Cady[38] and Zelenka.[39]

Overview of Types of Acoustic Waves

Introduction

Acoustic wave devices involve piezoelectric substrates or the coupling of piezoelectric materials to nonpiezoelectric structures to (e.g., a Si cantilever equipped with a thin film of piezoelectric ZnO). These devices are described by the mode of sound wave propagation through the piezoelectric substrate or through the linked nonpiezoelectric material. Primarily, their velocities and displacement directions distinguish sound wave types. It is important to remember here that acoustic waves are the result of induced particle vibration that depends on the elastic properties of the solid. In Table 8.5 we summarize several, but not all, of the wave modes possible in solids.

Metal transducers launch acoustic waves at ultrasonic frequencies and the type and resonant frequency of such waves are fixed by the piezoelectric crystal orientation, the thickness (t) of the piezoelectric used, and the geometry of the metal transducers. The resulting mechanical waves are called *transverse* or *shear waves* when the particle displacement is perpendicular to the direction of the wave propagation. Shear waves cannot propagate in fluids, and in a solid they typically travel at half the velocity of a longitudinal wave in the same material. In contrast, compressional or longitudinal acoustic waves are defined when particle displacement is parallel to

Table 8.5 Acoustic Wave Types in Solids

Wave Type	Particle Vibrations
Longitudinal	Parallel to wave direction
Transverse (shear)	Perpendicular to wave direction
Surface (Rayleigh)	Elliptical orbit, symmetrical mode
Plate (Lamb)	Component perpendicular to surface (extensional wave)
Plate (Love)	Parallel to plane layer, perpendicular to wave direction

the direction of the acoustic wave travel, and these waves can be transmitted in solids and liquids.

Bulk Acoustic Wave Devices

While bulk acoustic waves (BAWs) propagate through the volume of the piezoelectric substrate, the propagation of surface acoustic waves (SAWs) or Rayleigh waves is restricted to a distance of about one acoustic wavelength from the piezoelectric substrate surface. Rayleigh waves are somewhat similar to water waves in that the motion of particles is both longitudinal and transverse in a plane containing the direction of propagation and the normal to the surface.

A commonly used BAW device is the thickness shear mode (TSM) resonator shown in Figure 8.18. In TSM mode, bulk transverse waves with particle displacement parallel to the surface of the sensor are generated. Such waves, on propagation into a liquid, suffer severe attenuation of the acoustic energy due to the viscous forces in the liquid. The device pictured in Figure 8.18 is more commonly known as a *quartz crystal microbalance* (QCM) and is used, among other things, to measure deposited metal thickness in an evaporation station (see Volume II, Figure 7.2a). It consists of an AT-cut quartz disc with metal electrodes on opposite sides to effect the application of an oscillating electric field. A quartz AT cut is made normal to the *y*-axis but rotated 35.25° from the *z*-axis. In this direction, quartz has a zero temperature coefficient at 40°C, making the AT cut ideal for frequency control applications. Quartz is a hexagonal crystal (see Figure 8.17), and if the electrical field is aligned with the *y*-axis, the resulting deformation is a shear along the neutral *z*-axis. To achieve a shear motion, the cut must be normal to the *y*-axis so that charging electrodes placed on the two faces of the wafer can produce the field. The thickness (*t*) of the quartz

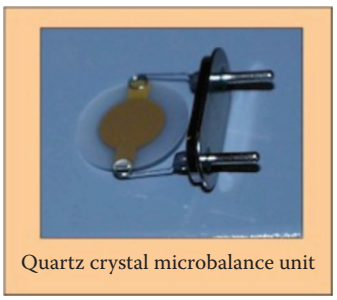


FIGURE 8.18 Quartz crystal microbalance.

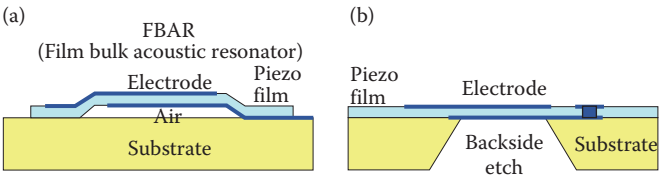


FIGURE 8.19 Two methods to make film bulk acoustic resonators (FBARS). The piezo film is typically aluminum nitride (AlN).

plate determines the wavelengths (λ) of the fundamental (n = 1) and harmonic resonances through the expression $\lambda = 2t/n$ (n = 3, 5, 7). Consequently, the resonant frequency for the fundamental mode is restricted by the thickness *t* of the quartz wafers to the range of approximately 5–20 MHz.

A more recent type of BAW device is a film bulk acoustic resonator (FBAR) that is used in mobile telephones. FBARs feature a $Q > 400$ at 1.9 GHz and often use aluminum nitride (AlN) as the piezoelectric material. The major difference between an FBAR and the QCM of Figure 8.19 is that the AlN thin film piezoelectric is much thinner than the quartz cut. Two types of FBAR constructs are shown in Figure 8.19. Figure 8.19a involves the etching of a sacrificial support layer below the piezofilm film. Figure 8.19b consists of etching the substrate from the back of the wafer to the front surface creating a so-called *pothole structure*.

Surface Acoustic Waves

It was Lord Rayleigh, in 1887, who discovered the surface acoustic wave (SAW) mode of sound propagation and predicted their properties. Named for their discoverer, Rayleigh waves have a longitudinal and a vertical shear component that couple with a medium in contact with the device's surface. Rayleigh waves are also of interest to geologists since these surface waves are excited by earthquakes.

A SAW device, as illustrated in Figure 8.20, comprises a thick ST-cut quartz plate and two interdigital metal transducers (IDTs).* In a typical "delay line" configuration, input and output IDTs generate or receive Rayleigh waves on the same side of a piezoelectric substrate. It was White and Voltmer[40]

* An interdigital transducer (IDT), or interdigitated transducer, is a device which consists of two interlocking comb-shaped metallic coatings.

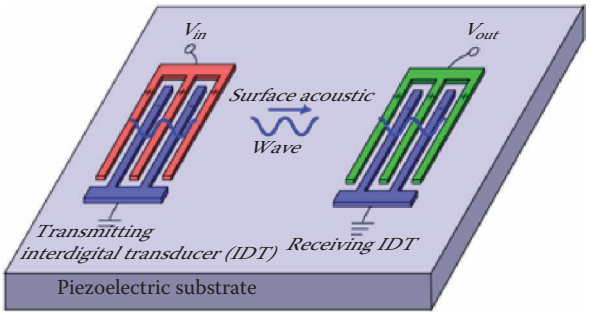


FIGURE 8.20 Surface-acoustic wave device; see also the hygrometer in Figure 8.26.

who pioneered the marriage of piezoelectricity with microelectronics through the demonstration that photolithographic techniques could be employed to deposit IDTs on piezoelectric substrates to excite Rayleigh waves. Rayleigh waves have a velocity that is ~5 orders of magnitude less than the corresponding electromagnetic wave, making Rayleigh surface waves among the slowest to propagate in solids. The wave amplitudes are typically ~10 Å, and the wavelengths range from 1 to 100 μm. The frequency of a SAW device is determined by the "finger" spacing of the IDTs ranging from 30 to 500 MHz (so they can be more than ten times faster than a QCM).

The surface particles in a Rayleigh wave move elliptically, resulting in a wave consisting of both shear and compressional components. The latter provokes an important attenuation effect in liquids that prevents the application of SAW devices in such media. An exception to this loss of sound energy occurs for devices using waves that propagate at a velocity lower than the sound velocity in the liquid. Regardless of the displacement components, such modes do not radiate coherently and are thus relatively undamped by liquids [see next section, on flexural plate wave (FPW) devices].

Flexural Plate Waves

A final important category of surface-launched acoustic wave sensors is those composed of very thin plate devices, where the thickness of a piezoelectric substrate is reduced to a dimension on the order of the acoustic wavelength itself. These devices use a thin piezoelectric substrate, or plate, as an acoustic waveguide that confines the acoustic energy between the upper and lower surfaces of the plate. The devices that generate such waves are termed *FPW devices*, as illustrated in Figure 8.21, and the waves they generate are called *Lamb waves*. Although surface-excited, the generated Lamb waves travel through the bulk of the piezoelectric membrane, and since both surfaces undergo displacement, detection may occur on either side. This is a very important advantage, as one side may feature the IDTs that must be isolated from, say, conducting liquids or corrosive gases, while the other side can be used as a sensing surface. Lamb waves generated in a plate of finite thickness can be of both symmetric and antisymmetric modes when referring to the median plane of the plate. Since the piezoelectric plate is so thin, the operating frequencies of FPW devices are substantially lower than for other surface-launched acoustic wave devices (2–7 MHz). As a consequence of the minimal thickness of the plate, the wave velocity is lower than the compressional velocity of sound in liquids. Accordingly, compressional waves are not coupled in liquids, resulting in less attenuation of acoustic energy. This is a consequence of Snell's law: the acoustic wave is trapped in the plate much as a light wave is trapped in an optical waveguide. This means that FPWs can even be used in liquids.

A practical FPW device, as shown in Figure 8.21, has a design similar to that of a SAW device. The basic elements of an FPW device consist of a composite plate of low-stress thin substrate material (e.g., 0.8–4-μm silicon nitride), a thin zinc oxide film (0.3–1.5 μm) as piezoelectric, and a pair of aluminum IDTs on the ZnO layer. IDTs are positioned in a way to form a delay-line configuration. A metal ground plane (0.2–05 μm thick) is sandwiched between the silicon nitride and piezoelectric layer. A chemical sensitive film may be deposited on the opposite side of the membrane. The IDTs launch and receive the plate waves (e.g., the 0th-order antisymmetric Lamb waves*) and, together with an amplifier, form a feed-

* Lamb waves are associated with Rayleigh waves on a plate whose thickness is smaller than the acoustic wavelength. They can be considered as two Rayleigh waves propagating on both sides of the thin plate. Two kinds of waves can propagate through the plate independently, namely the symmetric and the antisymmetric waves. The A_0 wave is the antisymmetric Lamb wave with the lowest phase velocity. This makes the wave suitable for liquid sensing because the A_0 wave does not excite compressional waves in a loading liquid and thus reduces scattering energy if its phase velocity is lower than the sound velocity of liquid. The A_0 mode Lamb wave sensor is also called the *FPW sensor.*

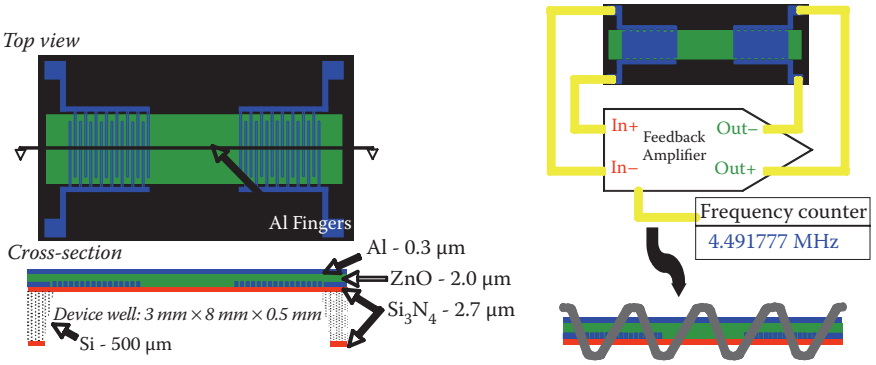


FIGURE 8.21 Construction and operation of a flexural pate wave device. (Courtesy of Dr. Dick White, University of California, Berkeley.)

back oscillator whose output frequency may provide a measure of the mass per unit area of the membrane, including a chemically sensitive film deposited on it. Summarizing the differences between an FPW device and a SAW device: 1) FPW devices operate at lower frequencies (a few megahertz as opposed to hundreds of megahertz for SAW sensors); 2) the IDTs are deposited on the opposite side of the FPW sensing material, and thus the sensor can be easier applied to liquids because the IDT electrodes are insulated from the liquid; and 3) the plate thickness is much smaller than the acoustic wavelength. FPW devices offer high mass sensitivity at much lower operating frequency than SAW devices.

Piezoelectric actuators generally produce very strong forces and very small motions that can be amplified, for example, by making the piezoelectric material part of a bimorph or a stack (Figure 8.22). Stacked structures attain larger displacements but are more difficult to implement in a miniaturized device. Piezo materials operate at high speed and return to a neutral position when the power is switched off. Alternating currents produce oscillations in a piezoelectric, and operation at the specimen's fundamental resonant frequency produces the largest elongation and at the highest power efficiency.

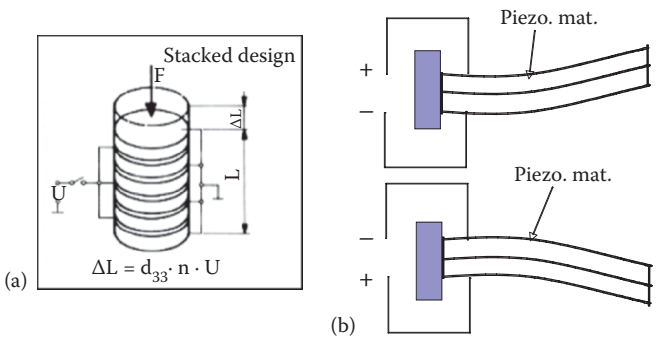


FIGURE 8.22 Piezoelectric stack actuator (a) and piezoelectric bimorph actuator (b).

Overview of Piezoelectric Materials

The first extensively used piezoelectric materials were naturally occurring crystals —quartz and Rochelle salt ($NaKC_4H_4O_6 \cdot 4H_2O$). Quartz has great physical and chemical stability but only a small coupling coefficient, k. Applications of quartz includes oscillators and resonators (see, for example, the quartz angular rate sensor based on a tuning fork in Figure 5.1). Rochelle salt has a great piezoelectric effect, but it is soluble in water and has poor temperature characteristics. It has been used in various transducers such as phonograph pickups.

In synthetic piezoelectrics, ceramics formed by many tightly compacted crystallites are most popular. Many piezoelectric materials, such as barium titanate ($BaTiO_3$), lead titanate ($PbTiO_3$), lead zirconate titanate (PZT), lead lanthanum zirconate titanate (PLZT), lead magnesium niobate (PMN), potassium niobate ($KNbO_3$), potassium sodium niobate ($K_xNa_{1-x}NbO_3$), and potassium tantalate niobate [$K(Ta_xNb_{1-x})O_3$], have a perovskite structure, that is, a structure of the type ABO_3. These ceramics are all ferroelectrics. As we saw in Volume I, Chapter 2, this absence of a center of symmetry in the crystal structure gives rise to spontaneous polarization. The crystal structure is cubic above the Curie temperature and tetragonal as it cools down. To align the

dipoles in the compacted monocrystals in the same direction (i.e., to polarize or pole them), they are subjected to a strong electric field during their fabrication process. Above the Curie point, the dipole directions in ferroelectric materials have random orientations. To align the dipoles, fields of 10 $kVcm^{-1}$ are common at temperatures slightly above the Curie temperature. The ceramic is then cooled while maintaining the field. When the field is removed, the crystallites cannot reorder in random form because of the mechanical stresses accumulated, resulting in a permanent electric polarization.

Barium titanate was the first material to be developed as a piezoceramic (see also Volume I, Figure 2.10, where the use of this material as a pyroelectric was highlighted). It was discovered independently, in the United States (Waigner and Salomon, 1942), the Soviet Union (Wul and Goldman, 1945), and Japan (Miyake and Ueda, 1946). Applications include detection of mechanical vibration, construction of actuators, and generation of acoustic and ultrasonic vibrations.

PZT is a solid solution of $PbTiO_3$ (tetragonal, with six poling directions) and $PbZrO_3$ (rhombohedral, with eight poling directions). This piezoceramic comes in various compositions such as PZT-4, PZT-5, PZT-5A, and PZT-5H. It was discovered in 1955 by Takagi, Shirane, and Sawagachi in Japan and Jaffe in the United States. The material has a high Curie temperature and is used in devices such as ceramic filters and piezoelectric igniters. Important for MEMS, PZT can be fashioned in thin film form by RF sputtering and rapid thermal annealing and subsequent poling. Alternatively, a sol-gel process is used to fabricate these films. One obtains a piezoelectric material with a relative dielectric constant of more than 1000 this way.[41] Abe and Reed[42] prefer sputtering and argue that the sol-gel technique is less attractive for MEMS applications, given that a 1-µm-thick film would require repeated spin coatings and pyrolysis steps. They prepared PZT films by sputtering from a composite target and found substrate heating during deposition to be crucial to transform the films into the perovskite phase during subsequent annealing. Piezoelectric thin films, such as PZT, zinc oxide (ZnO), and aluminum nitride (AlN), have the cost advantage over crystalline materials. Among these three, the electromechanical coupling effect of PZT is 3 to 9 times larger and the dielectric constant 100 times larger than that of AlN and ZnO, which makes PZT potentially the best candidate for thin film acoustic sensors. However, the polycrystalline structure of PZT and the required heat treatment during the coating process complicate the realization. A problem with all these polycrystalline materials in general relates to their temperature sensitivity and aging (loss of piezo properties) when approaching the Curie temperature.

Polymers such as polyvinylidene fluoride (PVDF), lacking central symmetry, also display piezoelectric properties. This piezopolymer is also pyroelectric, capable of converting changes in temperature to electrical output (see below). Compared with quartz and ceramics, piezofilm is obviously more pliant and lighter in weight. In addition, it is rugged, quite inert, and low cost. This engineering material could provide a sensing solution for vibration sensors, acceleration and shock sensors, passive infrared (IR) sensors, solid state switches, and acoustic and ultrasonic sensors. It should be noted that the peak reversible stress and strain developed by PVDF approximate 3 MPa and 0.1%, respectively. The latter values are rather small, and as for piezoelectric ceramics and magnetostrictive materials (see below), mechanical amplifiers are required to produce the larger displacements needed for many applications. A 200-fold mechanical amplification is required to make PVDF's peak strain similar to that of mammalian muscle.[43] Table 8.6[44,45] summarizes some of the physical properties making PVDF an interesting sensor material.

AlN is a wide-bandgap semiconductor and an excellent candidate for the integration of acoustic wave devices on chips with silicon-based electronics. AlN is a stiff and light material that exhibits excellent piezoelectric properties, including an electromechanical coupling coefficient of 0.088 and a high in-plane acoustic velocity of ~5700 m/s. As a piezoelectric ceramic, aluminum nitride is compatible with standard silicon MEMS processing and has found widespread use in FBARs (see Figure 8.19) and SAW devices (Figure 8.20) AlN may also prove to be useful for the integration of mechanical devices and in the field of packaging of bio-MEMS

Table 8.6 Typical Properties of Piezofilm (Uniaxially Oriented Film)

Film Thickness (μm)	10–100
Piezoelectric strain constants (pC/N) d_{31} d_{33}	 28 –35
Young's modulus (10^9/m^2) E_{33}	 5.4
Poisson ratio ν_{21} ν_{31} ν_{32}	 0.25 0.57 0.45
Pyroelectric coefficient (10^{-6} C/m^{2}K)	–30
Mass density (10^3 kg/m^3)	1.78
Temperature range (°C)	–40 to 145
Breakdown voltage (V/μm)	80
Maximum operating voltage (V/μm)	30
Yield strength (10^6 N/m^2)	45–55
Permittivity (10^{-12} F/m) at 10kHz	106–113
Capacitance (pF/cm^2) at 10 kHz	380
Electromechanical coupling factor, 10 Hz k_{31} k_{32}	 13% 1.7%
Piezoelectric stress constant (10^{-3} m/mC/m^2) g_{31} g_{33}	 216 –339
Volume resistivity (Ω · m)	$>10^{13}$

Sources: AMP. 1993. *Piezo film sensors*. Valley Forge, PA: AMP;[44] and Nalwa, H. S., ed. 1995. *Ferroelectric polymers: Chemistry, physics, and applications.* New York: Marcel Dekker.[45]

devices because of its superior thermal conductivity 130–140 (W/mK, room temperature to 100°C) and nontoxicity. The material is often grown by DC magnetron sputtering or by LPCVD on a (001) sapphire substrate (heated to 300–400°C) and can also be epitaxially grown on <111> silicon. Even atomic layer deposition (ALD) has been used to deposit AlN.

The last piezoelectric material we cover here is ZnO, a very versatile piezoelectric that has been used in cantilever-beam accelerometers, gas sensors, FPW devices, IR detectors, anemometers, tactile sensor arrays, etc. A typical change in length per electric field unit (perpendicular to the field E) for ZnO equals 5×10^{-10} C/N compared with 100–200×10^{-10} C/N for PZT-type thin film ceramics. Techniques that have been used for zinc-oxide thin film deposition include RF and dc sputtering, chemical vapor deposition, ion plating, reactive magnetron sputtering, and planar magnetron sputtering (in Polla and Muller[46] and references therein). For thin film deposition of ZnO, planar magnetron sputtering appears to give the best piezoelectric and pyroelectric characteristics; using this technique highly oriented ZnO films have been deposited on SiO_2, polycrystalline silicon, and bare silicon substrates. The compatibility of ZnO with silicon processing is demonstrated, for example, in the commercially available acoustic sensor from Honeywell Corporation, shown in Figure 8.27. The microphone chip illustrated here incorporates ZnO deposited on a silicon substrate, including signal-conditioning circuitry. Muller[47] found the best thin film crystallinity at a sputtering power of 200 W with a 10 mTorr[46] ambient gas mixture consisting of an equal mix of oxygen and argon. The distance between the substrate and target measures about 4 cm, with the substrate temperature maintained at 230°C during deposition. Tjhen et al.[48] characterized the thin film properties of sputtered ZnO and found the electrical properties to be sensitive to substrate material, stress, and surface condition.

The considerations in designing sensors with AlN, ZnO, or PZT films include the following:

1. Value of electromechanical coupling
2. Good adhesion to the substrate
3. Resistance to environmental effects (e.g., humidity, temperature)
4. VLSI process compatibility
5. Cost effectiveness

To further improve the mechanical properties of piezoelectric sensors, piezoelectric composite materials are sometimes used. These are heterogeneous systems consisting of two or more different phases, of which at least one shows piezoelectric properties. Hirata et al.,[49] for example, used LIGA to make arrays of PZT rods with diameters below 20 μm and an aspect ratio of over 5. The PZT rods are embedded in a polymer matrix to make a PZT/polymer composite. The PZT slurry, average particle size 0.4 μm, is injected into a LIGA resist mold (for information on slurry casting of piezoceramics see Preu et al.[50] and Lubitz[51]). After solidification, the resist is removed by plasma etching, and the array of

Table 8.7 Comparison of Some Important Piezoelectric Materials

Material	Piezoelectric Constant d (pC/N)			Dielectric Constant	Young's Modulus E (GPa)	Curie Temperature (°C)	Maximum k (Coupling Coefficient)
AlN	–0.48 (d_{15})	–0.58 (d_{31})	1.55 (d_{33})	9.4	$E_{11} = 410 \pm 10$ $E_{12} = 149 \pm 10$ $E_{13} = 99 \pm 4$ $E_{33} = 389 \pm 10$ $E_{44} = 125 \pm 5$	Not known (>1150 °C)	
Quartz	2.31 (d_{11})	0.73 (d_{14})		$\varepsilon_1 = 4.52$ $\varepsilon_3 = 4.63$	107	550	0.1
PZT (depending on composition)	80–593 (d_{33})	–94 to –274 (d_{31})	494–784 (d_{15})	$\varepsilon_3 = 425–1900$	53	193–490	0.69–0.75
PVDF (Kynar)	23 (d_{31})	4 (d_{32})	–35 (d_{33})	12	3	>MP (150)	0.2
ZnO	–12 (d_{15})	12 (d_{33})	(–4.7) (d_{31})	$\varepsilon_3 = 8.2$	123		
$BaTiO_3$	78 (d_{31})	190 (d_{33})		1,700			

Sources: Based on Pallas-Areny, R., and J.G. Webster. 1991. *Sensors and signal conditioning.* New York: Wiley;[34] Khazan, A.D. 1994. *Transducers and their elements.* Englewood Cliffs, NJ: PTR Prentice Hall;[36] and Flynn, A.M., L.S. Tavrow, S.F. Bart, R.A. Brooks, D.J. Ehrlich, K.R. Udayakumar, and L.E. Cross. 1992. Piezoelectric micromotors. *J Microelectromech Syst* 1:44.[41]

ceramic posts is cast in an epoxy resin. Resolution of acoustic-imaging transducers, for example, for medical applications, should gain resolution as the PZT rods undergo further miniaturization as a result of minimized cross-talk in the array.

Typical piezoelectric materials are compared in Table 8.7. For a survey of inorganic ferroelectrics, refer to *Ferroelectric Transducers and Sensors*.[35] For a detailed review of ferroelectric polymers, refer to *Ferroelectric Polymers: Chemistry, Physics, and Applications*.[45]

Piezoelectric Applications

Overview

When the characteristics of an acoustic path change, a change in acoustic output results. As the acoustic wave propagates through or on the surface of a piezoelectric material, changes to the characteristics of the propagation path affect the velocity and/or amplitude of the wave. Those changes in velocity can be monitored by measuring the frequency or the phase characteristics of the piezoelectric device and can then be correlated to the corresponding physical quantity being measured. In principle, then, all acoustic wave devices are sensors in that they are sensitive to many different physical parameters. The range of phenomena that can be detected by acoustic devices can be further expanded dramatically by coating the piezoelectric material or a structure coupled to it with materials that undergo changes in their mass, elasticity, or conductivity upon exposure to some physical or chemical stimulus. These sensors then may become pressure, torque, shock, and force detectors if under an applied stress the dynamics of the propagating medium change. They become mass, or gravimetric, sensors when particles are allowed to contact the propagation medium, changing the stress on it. They are turned into vapor sensors when a coating is applied that absorbs only specific chemical vapors. These devices work by measuring the mass of the absorbed vapor. A heating element under the chemical film can also be used to desorb chemicals from the device. If the coating absorbs specific biological chemicals in liquids, the detectors become biosensors.

Bulk ceramic piezoelectric devices, say on the basis of PZTs, have been widely used for decades, but thin film applications are more recent arrivals. Telecommunications constitutes the largest market today, accounting for about 3 billion acoustic wave filters annually, primarily for mobile cell phones. This use involves SAW devices, and lately, FBARs. These components are used as bandpass filters. Some of the new emerging applications for acoustic wave devices include all types of sensors for automotive applications (accelerometers and torque and tire pressure sensors), medical applications (chemical

sensors), and industrial and commercial applications (vapor, humidity, temperature, and mass sensors). Some of these sensors are capable of being passively and wirelessly interrogated (no sensor power source required). The use of piezoelectrics in actuators is also a growth area and includes micromotors, ignitors, bimorphs, impact printer heads, speakers, microphones, precision positioning stages, microscale surgical tools,[52] pumps, inchworm stepping motors, and tip-positioning stages for scanning tunneling microscope systems.

The small strains, usually less than 0.1%, and high stresses (e.g., 35 MPa) generated by piezoelectric devices have spawned a diverse range of actuator applications, as noted above (see also Figure 8.23 below; see Figure 1.11 for a micropiezopump from Animas-Debiotech's). Bulk piezoelectric transducers have also been used for a long time in medical ultrasound, in both passive and active modes.[53] In the passive mode, the transducer acts as a sound receiver only; it converts mechanical sound energy into an electrical signal. In the active mode (the converse effect), the transducer is an active sound transmitter; electrical energy is transformed into mechanical sound energy. Mostly, piezoelectrics are used in a pulse echo mode in which the transducers are used to simultaneously perform active and passive functions. A sound wave is propagated into the surrounding medium, and a faint echo is received after a small time gap due to the acoustic mismatch between the interface materials. This principle is used for ultrasonic imaging applications. When used in the human body with a sound velocity of 1500 ms^{-1}, the resolution varies from 1 mm to 50 μm for the frequency range of 1.5–3 MHz. The maximum frequency is dictated by the attenuation in the body, that is, 0.5 dB/cm MHz.

In the following we detail four of the applications mentioned above, namely, micromotors, pumps, acoustic-based chemical sensors, and acoustic imaging.

Piezoelectric Motors

Piezoelectric actuators, as Flynn et al.[41] point out, offer significant advantages over both electrostatic (see above) and electromagnetic (see below) actuators. With motors, for example, the greater the energy density that can be stored in the gap between rotor and stator, the greater the potential for converting

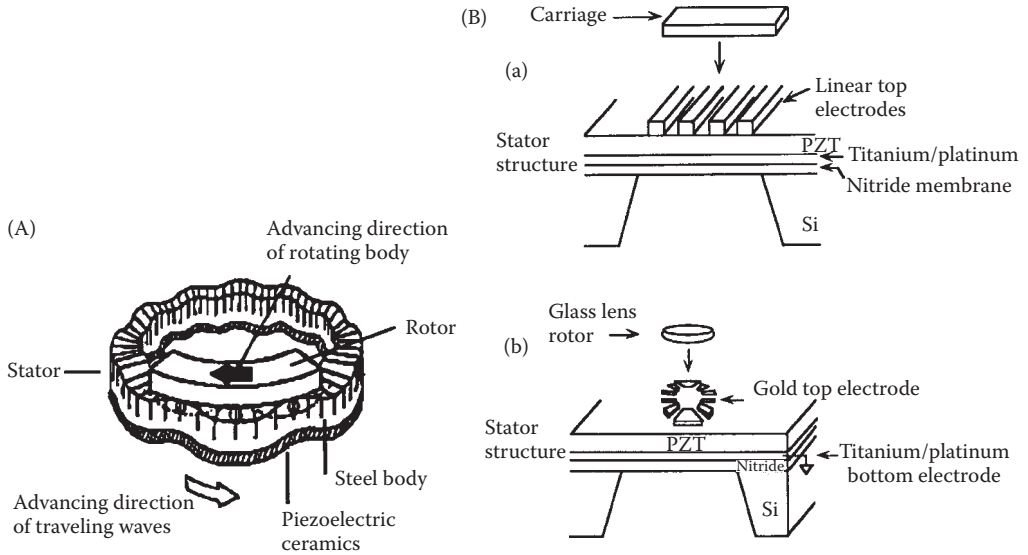


FIGURE 8.23 Piezoelectric motors. (A) Schematic of commercially available Panasonic ultrasonic motor: an electrically induced wave of mechanical deformation travels through a piezoelectric medium in the stator, moving the rotor body along through friction. (B) Thin-film lead zirconate titanate (PZT) motors. (a) Linear motor: the PZT stator is patterned onto a silicon membrane. The stator can deflect more because the nitride membrane is thin; titanium and platinum form the ground electrode, and the linear gold stripes are the top electrodes. A carriage is deposited on the stator by hand. (b) Rotary motor: identical to the linear motor except that the top electrodes are patterned in a circle. (After Flynn, A.M., L.S. Tavrow, S.F. Bart, R.A. Brooks, D.J. Ehrlich, K.R. Udayakumar, and L.E. Cross. 1992. Piezoelectric micromotors. *J Microelectromech Syst* 1:44.[41])

that energy to torque or useful work. In what follows, we compare the energy density one can store in the rotor/stator gap for various motor mechanisms. For electrostatic motors, the maximum energy density storable in the air gap, is determined by E_b, the maximum electric field before breakdown (approximately 3×10^8 Vm^{-1} for 1-μm gaps). From Chapter 7 we recall:

$$U'_m = \frac{\varepsilon_r \varepsilon_0 E_b^2}{2} \tag{7.36}$$

where $\varepsilon_r = \varepsilon_{air}$ represents the permittivity of air (almost equal to that of free space). This results in an energy density of 4.0×10^5 Jm^{-3}. With a piezoelectric motor made from ferroelectric material such as PZT, the energy density again is determined by the maximum electric field before breakdown. Thin film PZT can withstand high electric fields, about the same as an air gap ($E_b = 3 \times 10^8$ Vm^{-1}), but the dielectric constant is 3 orders of magnitude larger ($\varepsilon_{PZT} = 1300\ \varepsilon_{air}$) than air:

$$U'_m = \frac{\varepsilon_{pzt} \varepsilon_0 E_b^2}{2} \tag{8.61}$$

In principle then, the energy density also should be 3 orders of magnitude larger. In case of magnetic motors, the magnetic energy density, U'_m, stored in an air gap is given by:

$$U'_m = B^2/2\mu_0 \tag{8.62}$$

where B is the magnitude of the magnetic flux density in the gap and μ_0 the magnetic permeability of free space (see further below Equation 8.80). With magnetic actuators, the gap energy density, by pushing B into saturation (about 1.5 T), leads to 950,000 Jm^{-3}. The energy density stored in the rotor/stator gap for magnetic, electrostatic, piezoelectric, and SMA actuation (see below) are compared in Table 8.8.

Obviously, piezoelectric ultrasonic motors have an energy advantage. Other advantages, summarized from Flynn et al.,[41] include the following:

- Low voltages: no air gap is needed; mechanical forces are generated by applying a voltage directly across the piezoelectric film. With a 0.3-μm thin film, only a few volts are required, as opposed to hundreds of volts needed in air gap electrostatic motors.
- Geardown: motors can be fabricated without the need of a gearbox. Electrostatic wobble motors are also able to produce an inherent gear reduction but do not have the high dielectric advantage.
- No levitation: with electrostatic motors, levitation and flatness are very important in obtaining good sliding motion of the rotor around the bearing. The piezoelectric motor depends on friction so that no levitation is required, and it can be freely sized.
- Axial coupling: electrostatic motors require axial symmetry around the bearing. Since height is difficult to obtain with most nontraditional micromachining techniques (except for LIGA and pseudo-LIGA methods), a limited area is available for energy transduction. With the piezoelectric traveling-wave motors, linear or rotary motors can be built. As the stator is flat, with the rotor sitting on top of it, planar technologies are very capable of creating extra area to couple power out.
- Rotor material: the rotor can be of any material. A very important factor is that the conductivity of fluids being pumped with such motors does not affect the device, whereas an

Table 8.8 Comparing Energy Densities for Magnetic, Shape Memory Alloy, Electrostatic, and Piezoelectric Actuation

Principle	Maximum Work Energy Density	Equation	Special Drive Condition
Magnetic	9.5×10^5 Jm^{-3}	½ B^2/μ_0	1.5 T
Shape memory alloy	10.4×10^6 Jm^{-3} (from stress-strain isotherms)		1.4 W mm^{-3}
Electrostatic	4×10^5 Jm^{-3}	½ $\varepsilon_0 \varepsilon_{air} Eb^2$	3×10^8 Vm^{-1},1-μm gap in air
Piezoelectric	5.2×10^7 Jm^{-3}	½ $\varepsilon_0 \varepsilon_{PZT} Eb^2$	3×10^8 Vm^{-1},1-μm-thick lead zirconate titanate film

electrostatic motor would be shunted by a conductive liquid.

- Holding torque: because they are based on friction, piezoelectric ultrasonic motors can maintain holding torque even in the absence of applied power.

A schematic of a commercially available piezoelectric motor from Panasonic, which uses an ultrasonic traveling wave, is shown in Figure 8.23A. Two bulk ceramic PZT layers, each segmented with alternating poled regions, are placed on top of one another. For a given polarity of the applied bias, one segment contracts and the neighboring segment expands. With the two ceramic pieces put on top of each other with their segmented areas out of phase, any point on the stator then moves with the rotor being pulled along the stator surface through frictional coupling.

Thin film PZT linear and rotary motors by Flynn et al.[41] and Udayakumar et al.[54] are illustrated in Figure 8.23B. The stators in both cases are microfabricated using lithography techniques. Carriages and rotors may be made out of any type of material (in the rotational case, a glass lens was used for a rotor). The work demonstrates that, for a 5-V excitation, 1.6×10^{-12} N/m/V^2 normalized torque could be achieved, as opposed to 1.4×10^{-15} N/m/V^2 for electrostatic motors operating at 100 V.

The piezoelectric effect scales down with the bulk of the material; therefore, miniaturization opportunities are limited, and hybrid-type microactuators, where larger pieces of piezoelectric material are glued onto a substrate, are sometimes more reasonable. This has been the approach followed, for example, in the fabrication of piezoelectric micropumps (see below and Figure 1.11),[55,56] linear stepper motor,[57] and microvalves.[58]

Piezoelectric Pumping

In a tube vibrating with acoustic energy, an axially directed, acoustic-streaming force is generated along the inner surface.[59] This driving force can be generated along the entire length of a narrow tube to create a distributed pump that moves fluids without an externally applied pressure. The pump surface area and the driving force per unit volume of enclosed fluid increase as the tube diameter decreases

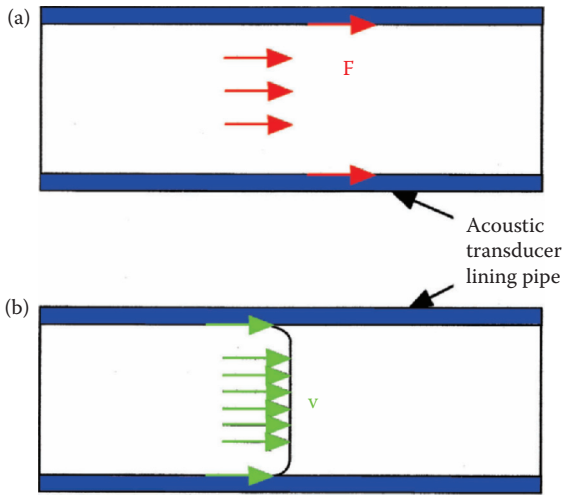


FIGURE 8.24 Acoustic streaming in a fluid inside a pipe. (a) Streaming force generated by an axially directed, traveling acoustic wave that grazes the inside walls of the tube. (b) The resulting steady state velocity distribution (blunt flow).

(pumping force $F \sim l^{-1}$). The axially directed, steady state force acting on the fluid is proportional to the square of the acoustic-wave amplitude and diminishes exponentially with distance from the wall. This force is sketched in Figure 8.24a. Because the force reaches a maximum near the walls of the tube, blunt flow is established in the fluid, as shown in Figure 8.24b, establishing an attractive technology to choose for miniaturization.

Unique properties of acoustic streaming pumps include the ability to pump without directly contacting the fluid so that a wide variety of liquids and gases may be pumped, and a simple compact structure with zero internal dead-volume. Tens of individual acoustic pumps could be operated in coordination to make an active network of interconnecting channels. One implementation of a piezopump is based on a flexural plate wave (FPW) delay line (see also Figure 8.21). An FPW device consists of a thin (~1 μm thick) membrane supported on all sides by a silicon chip. The membrane is coated with a piezoelectric film (say, ZnO) and aluminum interdigital transducers (IDTs). An RF voltage is applied to one IDT, producing mechanical stress in the piezoelectric layer, which generates flexural acoustic waves in the membrane (see Figure 8.25). The first IDT thus provides the electric field necessary to displace the substrate and form an acoustic wave. This wave propagates

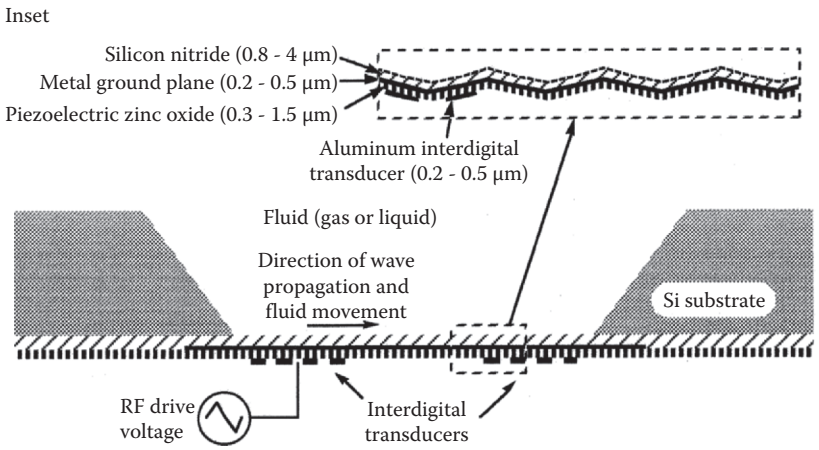


FIGURE 8.25 Micromachined flexural plate wave pump. The RF drive voltage applied to the piezoelectric interdigital transducers induces the flexural wave motion shown in the inset. Fluid motion is induced near the surface by acoustic stress. (Courtesy of Dr. R. White, University of California, Berkeley.)

through the substrate, where it is converted back to an electric field at the second IDT. Magnified, these waves look like the ripples in a flag waving in the breeze, as sketched in the inset of Figure 8.25. This wave motion induces acoustic streaming in the fluid next to the membrane. Wave propagation and fluid velocity both operate from left to right as shown in the figure. The structure generates large acoustic-wave amplitudes and is particularly effective for producing streaming; it has been proven to move air at speeds of 30 mm/s and water at 0.3 mm/s with only 5 V of RF drive.[60,61] Moroney et al. both translated and rotated small polysilicon blocks (2 μm thick and 50 μm square to 250 by 500 μm) using traveling ultrasonic flexural waves.[60] Acoustic streaming has also been established with the related SAW delay line.[62] As shown in Figure 8.20, a SAW device is fabricated by depositing interdigitated electrodes on the surface of a piezoelectric (typically $LiNbO_3$, $LiTaO_3$, and more recently also PZT thin films). An elastic wave generated at the input IDT travels along the surface and is detected by the output IDT. Besides their use for delay lines and filters in mobile phones, television and microwave communications, SAW devices have been used for many chemical sensor applications (see next) and in streaming experiments as discussed here.[63] Miyazaki et al., using a thin metal pipe flattened and bonded to a piezoelectric plate, produced a head pressure of about 10 mmH_2O and a flow rate of 0.02 cm^3/s at 40 V, peak to peak, applied to the piezoelectric ceramic plate.[62]

Acoustic Sensing of Mass

Introduction

In this section we say a few words about chemical and biological sensing by weighing with acoustic devices. Loading it directly onto a piezoelectric material can sense the mass of an analyte; alternatively, a thin layer of piezoelectric material might drive another resonator structure on which the mass accumulates as in microcantilever beams and resonator-comb drives. The main difficulty with mass sensors is achieving selectivity. In these applications, the adsorption of chemicals of interest is usually accomplished through polymer layers as sensitive coatings. One approach to improve selectivity further is to fabricate an array of sensors, each having a different partially selective film and using a pattern recognition algorithm to identify the species present (see artificial nose in Chapter 10). Another approach is to couple a front end to the sensor, such as a gas chromatograph, which separates the species and enables identification based on time of arrival at the sensor. These devices typically offer a signal-to-noise (S/N) ratio on the order of 100, with limits of detection (LOD) around 10 ppm. Mass sensitive piezotransducers used for chemical sensing are categorized by operating principle as bulk quartz resonators (BAWs), FBARs, SAW devices, FPW devices, cantilever beam resonators (see Example 8.1), and comb drive resonators.

As we saw above, all acoustic sensors will function well in gaseous or vacuum environments, but only a subset operates efficiently in contact with liquids. The shear horizontal wave does not radiate appreciable energy into liquids, allowing liquid operation without excessive damping.

Mass-Sensing BAW Devices

The mass sensitivity of a quartz crystal microbalance (QCM) or a BAW device, as derived by Sauerbrey[64] in 1959, is given by:

$$\Delta f = -\frac{2f_0^2}{A\nu_q\rho_q}\Delta m = -\frac{Cf_0^2}{A}\Delta m \qquad (8.63)$$

with:

$$C = \frac{2}{\nu_q\rho_q}$$

where Δf = frequency shift due to the added mass in hertz
C = a constant
f_0 = fundamental frequency of the oscillator in Hz
$\Delta m/A$ = surface mass loading in kg/m^2
ν_q = shear wave velocity in quartz (= 3340 ms^{-1})
ρ_q = density of quartz crystal (= 2648 kgm^{-3})

Assumptions in the derivation of this equation are that the sensor film is thin and rigidly coupled to the resonator surface, the density and acoustic properties of the film are identical to that of the quartz resonator (for very thin coatings, the influence of the coating layer on the acoustic properties is assumed to be negligible), the film is uniform and evenly coated onto the mass-sensitive electrode surface, no mechanical stress is applied to the crystal by the film, and no interfacial slip occurs at the quartz/film boundary. The equation does not apply to non-Newtonian fluids, viscoelastic films, thick films, and liquid environment. In essence, this treats any added layer on the device as being composed of quartz; that is, the new film is considered as extending the acoustic wavelength of the device. The equation predicts that the sensitivity of the BAW sensor is proportional to the square of the natural frequency. In practice, though, the natural frequency is limited since it cannot increase without decreasing the thickness of the crystal. A practical lower thickness of about 150 µm imposes a frequency upper limit of approximately 10 MHz. This frequency determines the maximum sensitivity, which based on this equation is calculated to be 0.23 kHz/µg.

In order to compare the various gravimetric sensing devices reviewed here, it is convenient to define the mass sensitivity of a sensor as follows:

$$S_m = \lim_{\Delta m \to 0} \frac{1}{f}\frac{\Delta f}{\Delta m} = \frac{1}{f}\frac{df}{dm} \qquad (8.64)$$

where Δm and dm are normalized to the active sensor area of the device. The sensitivity is thus the fractional change of the resonant frequency of the structure with addition of mass to the sensor. Another important characterization is the sensors's minimum detection limits. If the limit of mass resolution is defined as a mass sensitivity resulting in a frequency shift three times larger than the oscillator fluctuation (which is say 0.1 Hz), the minimum detectable mass density (MDMD) may be derived by rearranging Equation 8.64 as follows:

$$\Delta m_{min} = \frac{1}{S_m}\frac{\Delta f_{min}}{f} \qquad (8.65)$$

where Δm_{min} and Δf_{min} (say 0.3 Hz) are the MDMD and minimum detectable frequency change, respectively.

Based on Equation 8.64, for a thickness shear mode (TSM) device (a QCM), the theoretical sensitivity, S_m, varies as $1/\rho_q d$, where d is the device thickness. In the case of a 1-cm by 1-cm device operating at 6 MHz and with a typical thickness (d), an MDMD of 5 ng/cm^2 is calculated (Equation 8.65). This result is shown in Table 8.9.

Thin-film bulk acoustic resonators combine very high frequencies (GHz) with very thin deposited piezoelectrics (~1 µm or less). White and coworkers made an aerosol monitor using an FBAR device (see Figure 8.19) and calculated a theoretical mass limit of detection of 0.001 ng for it.[65] An FBAR is not only much more sensitive than a QCM, it is also significantly smaller.

Mass-Sensing SAW Devices

In the previous section, we saw that thinner BAW devices operate at higher frequencies and increase the mass sensitivity. However, thinning down

Table 8.9 Gravimetric Sensitivities of Acoustic Sensors

Sensor	Device Description	Theoretical Mass Sensitivity, S_m (cm^2/g)	Typical Operating frequency (MHz)	Theoretical Mass Limit of Detection (ng)	Typical Size
Thickness shear mode or quartz crystal microbalance	AT-cut quartz	$-1/\rho d$	6	5	1×1 cm
Surface acoustic wave	ST-cut quartz	$-K(\sigma)/\rho\lambda$	20–500	0.1	2×6 mm
Flexural plate wave	ZnO, SiN, Al	$-1/2\rho d$	2–5	0.002	1×6 mm
Film bulk acoustic resonator	AlN, SiN, Al, Au	$-1/\rho d$	1000–2000	0.001	100×100 μm
Microcantilever with mass loading distributed evenly over the cantilever surface		$S_m \sim 1/\rho d$	5–0.02	0.02	Micrometer-size (see Figure 8.77, where t is used for thickness)

Sources: Ballantine, D.S., R.M. White, S.L. Martin, A.J. Ricco, E.T. Zellers, G.C. Frye, and H. Wohltjen. 1997. *Acoustic wave sensors: Theory, design, and physico-chemical applications*. San Diego, CA: Academic Press;[63] and Kobrin, P., C. Seabury, A. Harker, and R. O'Toole. 1999. Thin-film resonant chemical sensor with resonant acoustic isolator. US Patent 5936150.[66]

quartz plates beyond the normal range results in fragile devices that are difficult to manufacture and to handle. For this reason, SAW devices have been investigated for gravimetry, typically operating at 100 MHz. Chemical vapor sensors based on SAW devices were first reported in 1979.[67] Most of them rely on the mass sensitivity of the detector, in conjunction with a chemically selective coating that absorbs the vapors of interest and results in an increased mass loading of the device. One disadvantage of the SAW devices is that Rayleigh waves are surface-normal waves, making them poorly suited for liquid sensing. When a SAW sensor is contacted by a liquid, the resulting compressional waves cause an excessive attenuation of the surface wave.

Often one works with two SAW devices, one of which is used as a reference, effectively minimizing the effects of temperature variations. As we saw above, SAWs are sound waves that travel parallel to the surface with their displacement amplitude decaying into the material so that they are confined to within roughly one wavelength of the surface. Interdigitated electrodes deposited on the piezoelectric material generate the acoustic waves and receive the signal. The pitch of the transducer defines the wavelength. SAW detectors, as shown in Figures 8.20 and 8.26, typically have a higher natural frequency than BAW detectors (30–500 MHz vs. a few MHz), and since signal strength is proportional to the fundamental frequency squared (Equation 8.63), they are more sensitive than quartz crystal microbalances. For the detailed derivation of the sensitivity S_m and MDMD for a SAW device, consult Ballantine et al.[63] They calculated $S_m = -K(\sigma)/\rho\lambda$, where λ is the wavelength and $K(\sigma)$ is a function of the Poisson ratio, σ. As reflected in Table 8.9, the mass sensitivity of a SAW device might be up to 50 times higher than that for a QCM, while a SAW device is typically somewhat smaller in size.

The superior sensitivity of a SAW device over a QCM may also be understood by recognizing that the sensitivity of the acoustic mass sensor is proportional to the amount of acoustic energy in the propagation path that is being perturbed. Bulk acoustic wave sensors typically disperse the energy from the sensing surface through the whole bulk of the material all the way to the other surface. This distribution of energy minimizes the energy density on the surface, which is where the sensing is done. SAW sensors, conversely, focus their energy on the surface, tending to make them more sensitive. However, SAW detectors are susceptible to changes in stiffness, dielectric constant, and relative permittivity relative to the surface mass, which can degrade sensitivity. BAW detectors have the principal advantage of negligible influences from conductivity and permittivity.

For both TSM and SAW sensors, to obtain higher sensitivity, one must increase the operating frequency.

SAW Example Application: Dew Point/Humidity Sensor

If a SAW sensor is temperature controlled and exposed to the ambient atmosphere, water will

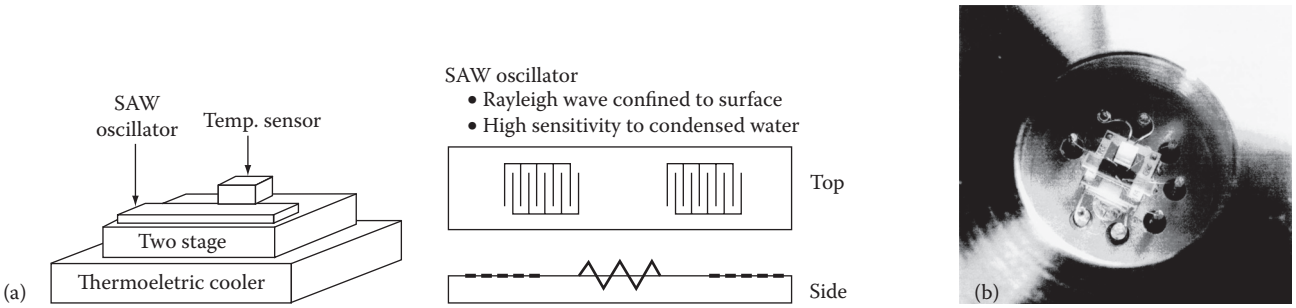


FIGURE 8.26 Surface acoustic wave (SAW) hygrometer. (a) schematic of the direct dew point sensor.[68] (Courtesy of Dr. M. Hoenk, Jet Propulsion Laboratory, Pasadena, CA.) (b) photograph of SAW hygrometer. The SAW device (long, rectangular bar) is mounted next to a platinum temperature sensor (square white package), on top of a miniature two-stage thermoelectric cooler. The device size is approximately 1 × 1 cm.

condense on it at the dew point temperature, making it an effective dew point sensor. Current commercial instruments for high-precision dew point measurements are based on optical techniques, which have cost, contamination, accuracy, sensitivity, and long-term stability issues. At Jet Propulsion Laboratory, as part of a Micro Weather Station development program, a new microhygrometer operating on accurate dew point measuring principles was demonstrated. A schematic illustrating the SAW hygrometer is shown in Figure 8.26. This device precisely measures the dew point/frost point temperature of ambient air. A quartz SAW microsensor detects the presence of thin layers of water on its surface as its temperature is varied by a miniature thermoelectric cooler. A fast digital feedback loop is used to control the sensor's temperature such that the amount of water on the sensor is held constant. This temperature is the dew point or frost point. This approach is functionally similar to the conventional chilled mirror hygrometers. However, the SAW moisture sensor provides much higher sensitivity and faster response for small amounts of water than does the optical detectors used in conventional hygrometers. This advantage, coupled with a high-performance digital feedback loop and precise measurements of the sensor temperature, enables accurate measurements with a response over an order of magnitude faster than conventional airborne hygrometers.

Mass-Sensing FPW Devices

FPW devices with sorbent polymer films respond to vapors in a manner similar to SAW devices coated with the same polymer. The FPW vapor sensor, however, is slightly smaller (see Table 8.9) and offers lower absolute noise levels and hence lower vapor detection limits. Earlier we listed other advantages associated with using an FPW or Lamb wave device for mass sensing. Lamb waves can propagate along a substrate even when immersed in water. Even though particle displacement is in the direction normal to the substrate, the low velocity of the acoustic waves preclude them from leaking into the environment. The accompanying electronics necessary for Lamb wave device applications also work at lower frequencies than those of SAW devices and are thus easier to design and cheaper to implement. Most importantly, as shown in Table 8.9, the mass sensitivity of a Lamb wave sensor is considerably greater than that of a SAW or Rayleigh wave sensor. The sensitivity calculated by Ballantine et al. for an FPW is given as $S_m \sim -1/2\rho d$.[63] In an FPW device, higher sensitivity is reached by reducing the plate thickness, d (this may be made less than a micrometer!). Mass sensitivity increases as membranes thickness decreases, and frequency decreases at the same time. This scaling law is in contrast to SAW devices, where mass sensitivity increases as frequency increases.

Acoustic Imaging

There is substantial interest in two-dimensional arrays of acoustic transducers for three-dimensional acoustic imaging. Such arrays may be used in handheld diver's sonar, medical ultrasound imaging, and

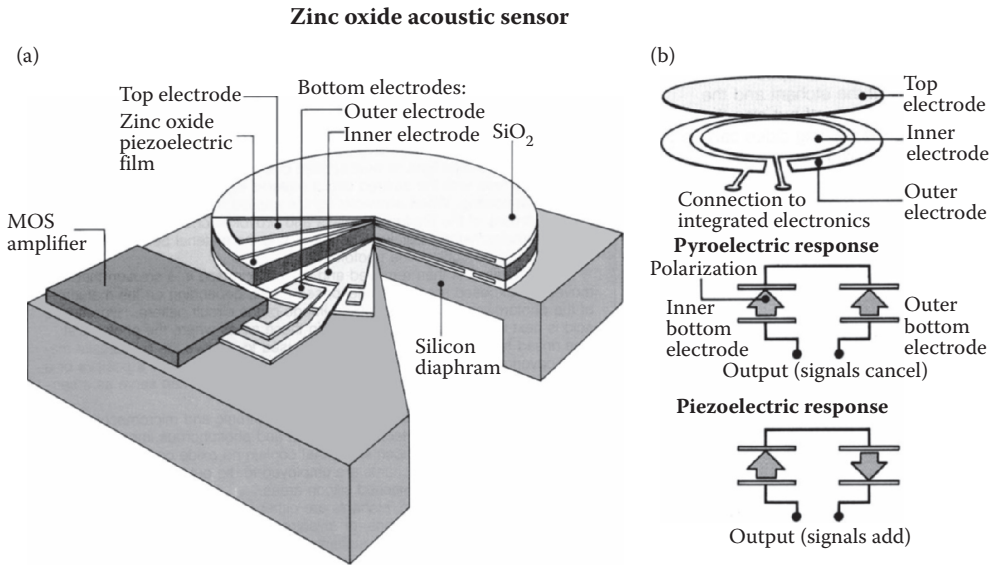


FIGURE 8.27 Zinc oxide acoustic sensor. (a) Two parallel plate electrodes in Honeywell's acoustic sensor act as capacitors. (b) Voltages due to temperature variations (pyroelectric effect) on zinc oxide film cancel, while those due to pressure add, doubling the output. (After Allen, R. 1984. Sensors in silicon. *High Technology,* September: 43–81.[70])

nondestructive testing. Bernstein et al.[69] microfabricated a 16 × 16 array of PZT pixels (3 μm thick and 0.4 × 0.4 mm each) for imaging purposes. The PZT pixels were deposited by sol-gel technology with polyimide in between the pixels. In a novel approach, the ferroelectric film was poled laterally (that is, in the plane of the deposited film) so that much larger interelectrode distances became possible. Instead of using metal electrodes on opposite sides of the thin PZT film, metal electrodes with a 30-μm spacing on the top side of the film resulted in a 30-fold increase in voltage sensitivity. With a stress applied in the same direction as the poling, there was a further gain of a factor of 2 due to a higher ferroelectric coupling constant in that direction. In total, a 30 dB improved sensitivity, compared with conventional polarizing across the thickness of the PZT film, was measured.[69] A commercially available acoustic sensor from Honeywell Corporation is shown in Figure 8.27.[70] The microphone chip illustrated incorporates ZnO deposited on a silicon substrate. A unique annular electrode design permits the parasitic signal in the zinc oxide film to be canceled because of temperature variations while doubling the output because of pressure (see Figure 8.27b).

Cast or spun-on and poled polymeric films of PVDF can also be employed to integrate piezoelectrics inexpensively into silicon micromachines. Gluing of commercial PVDF film with a urethane adhesive also works. Kolesar and Dyson,[71] for example, realized a two-dimensional electrically multiplexed tactile sensor this way. A 40-μm-thick PVDF film was glued onto an 8 × 8 array of taxel electrodes (400 × 400 μm each) electrically coupled to a set of sixty-four MOSFET amplifiers located around the periphery of this array. The response of the tactile sensor was found to be linear for loads between 0.008 and 1.35 N. Crude imaging of the applied loads also was demonstrated.

Piezoelectricity Summary

Some pros for using piezoelectricity as an actuator mechanism are high-energy density/bandwidth and lower operating voltages than electrostatics. Some cons are fabrication complexity and reproducibility of the actuator characteristics, and at the microscale, piezoelectrics provide very limited strain/displacement.

In Table 8.10 (see also Figure 8.14) we summarize some of the salient points associated with piezoelectric actuators.

Dielectric Induction

In the case of dielectric induction effects on particles in liquids and on fluids themselves, one speaks about

Table 8.10 Salient Features of Piezoelectric Actuators

1. Stroke: small, 10–2 μm
2. Speed: >100 kHz (see Figure 8.14)
3. Force density: low (see Figure 8.14)
4. Complex to fabricate
5. Voltage: ~10–100 V
6. Footprint: 1–10 μm
7. Maximum energy density, U_m (in J/m³): 52 × 106 Jm^{-3}

dielectrophoresis and EHD, respectively. The word *dielectrophoresis* implies the movement of particles through induction, while in *electrohydrodynamics* the emphasis is on the movement of the fluid itself. In dielectrophoresis, the induced charges can be attractive or repulsive with respect to the inducing field, depending on the dielectric constants of the particle and the medium in which the particle is embedded. This principle can be used for sorting, moving, and separating small objects such as biological cells.[72–74] The phenomenon is detailed in Volume I, Chapter 6, on fluidics.

In EHD pumping, fluid forces are generated by the interaction of electric fields and charges in the fluid. In such pumps electrodes are regularly spaced along a microchannel and each electrode is phase shifted with respect to the next. In contrast to forces generated by mechanical pumping using an impeller or bellows, EHD pumping requires no moving parts.[75–77] The interaction of electrical fields with induced electrical charge in the fluid yields a force that then transfers momentum to the fluid.[78] The EHD effect has also been put to use in mixing of fluids of different conductivities. Choi mixed DI water and saline solution in a 10-pL volume this way.[79] Surface charges induced at the interface of the two liquid streams react with the applied electric fields to generate shear forces leading to rapid mixing of the streams passing the electrodes. Reasonable electric fields can only be built up with acceptable voltage levels within microstructures; therefore, this principle becomes more and more effective with decreasing dimensions. A requirement for the continued existence of free charge is the presence of a spatial gradient in the conductivity or permittivity. Free charge injected into a region without gradient will relax in a time characterized by the relaxation time of the charge. One way to accomplish a conductivity gradient in the bulk of a slightly conductive fluid is by imposing a temperature gradient. A limitation of EHD is its reduced effectiveness with conductive fluids, such as in biological environments.

In an induction motor, localized charges are induced in the rotor by the electrostatic field generated by the stator electrodes. Since a finite relaxation time is needed for this excess charge to be diffused, a driving torque is generated. An induction motor is driven by a traveling wave of voltage and uses charge relaxation rather than physically salient poles in a variable-capacitance motor as shown in Figure 8.6b.[80] Consequently, the rotor can be a smooth uniform plate and may even be a fluid.[81] In the case of the rotor blades, fabrication involves some rather tricky planarization steps. The conductivity of the induction motor materials, on the other hand, strongly affects performance, so fabrication difficulties associated with variable-capacitance motors with rotor blades are traded for those associated with conductivity control in induction motors.[82]

Electrostriction

Electrostriction is a property of all dielectric materials caused by the presence of randomly aligned electrical domains within the material. With an electric field applied, the opposite sides of the domains become differently charged and attract each other, reducing material thickness in the direction of the applied field (and increasing thickness in the orthogonal directions in accordance with the Poisson's ratio. The resulting strain (ratio of deformation to the original dimension) is proportional to the square of the polarization. Reversal of the electric field does not reverse the direction of the deformation. In other words, electrostrictive materials—basically all dielectrics—develop mechanical deformations when subjected to an external electrical field; that is, they show an increase in length parallel to the field. Even apparently symmetric crystals of nonconducting materials exhibit this effect, but it is usually smaller than the piezoelectric effect.[83] The phenomenon is attributed to the rotation of small electrical domains in the material upon exposure to a field. In the absence of the field,

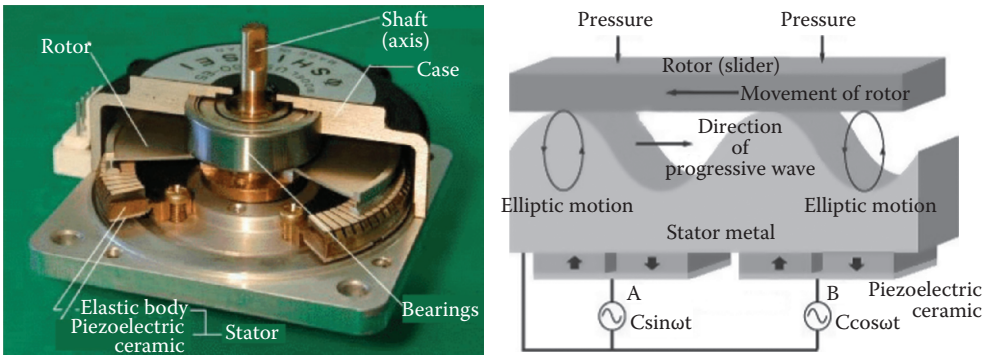


FIGURE 8.28 A ceramic piezoelectrostrictor in an ultrasonic motor. (From http://www.shinsei-motor.com/English/index.html.)

the domains are oriented randomly. The alignment of the electrical domains parallel to the electrical field results in the development of a deformation field in the electrostrictive material. Usually, the effect is small, but a few materials, such as certain titanates and zirconates, show a large effect, in which case they are ferroelectric and exhibit properties analogous to ferromagnetic materials. In both groups of materials, spontaneous polarization occurs within small regions or domains. The electric dipoles in ferroelectric materials or the elementary magnets in ferromagnetic materials are in parallel alignment within each domain. Like piezoelectrics, electrostrictors are often based on the perovskite crystal structure. Two examples of electrostrictive materials are barium titanate and lead magnesium niobate (PMN).[32,33] In the case of PMN, electrostriction gives rise to a mechanical deformation comparable to that induced by the piezoelectric effect in PZT. Fripp and Hagood compared the relative merits of piezoceramics and electrostrictors for vibration suppression.[83] Uchino made astronomy mirrors, the shape of which is continuously adjustable by bonding silvered glass to a series of electrostrictive layers that enable one to make minute adjustments to the mirror's shape in response to an applied field.[84]

A ceramic piezoelectrostrictor is used in the motor shown in Figure 8.28. This ultrasonic motor uses ultrasonic vibration (20 kHz or above) as the driving source and comprises a stator (a piezoelectrostrictor ceramic glued to a metal as the elastic body) and a rotor (the dynamic body). The piezoelectrostrictor ceramic is glued onto one side of the metal stator. The rotor is pressed against the opposite stator metal surface. Comb tooth-shaped grooves are created on the metal stator surface facing the rotor. The rotor is pressed tightly against this comb-toothed side of the stator metal surface. The ceramic element generates ultrasonic vibration when a specific high-frequency voltage is applied and the piezoelectric ceramic itself is expanding and contracting. The ultrasonic wave, i.e., swelling and falling of the stator surface, travels along the surface of the stator in a single direction. The ultrasonic vibration that travels in this manner is called the *progressive wave*. The comb tooth-shaped grooves on the stator surface are devised to make the amplitude of the elliptic motion large and to reduce abrasion. As the progressive wave travels and undulates through the contact, an elliptic motion is generated. The locus of the elliptic motion points in the opposite direction of the progressive wave traveling on the stator surface.

Pyroelectricity

The spontaneous polarization in ferrorelectrics is temperature sensitive. This is called the *pyroelectric effect*, which was discovered in tourmaline [(Na,Ca)(Li,Mg,Al)(Al,Fe,Mn)$_6$ (BO_3)$_3$ (Si_6O_{18})(OH_4)], the so-called *Ceylon magnet* (class 3m) by Teophrast in 314 BC and first recorded in the scientific literature in 1824 by Brewster.[32,85] The word is derived from the Greek for "heat electricity" and is the development of electric polarization in classes of noncentrosymmetric crystals subjected to a temperature change. Electrical polarity in a biological material is produced by a change in temperature. Pyroelectricity is probably a basic physical property of all living organisms. First discovered in 1966 in

tendon and bone, it has since been shown to exist in most animal and plant tissues and in individual cells. Pyroelectricity appears to play a fundamental part in the growth processes (morphogenesis) and in physiological functions (such as sensory perception) of organisms. From Volume I, Chapter 2, we know that there are twenty-one classes lacking a center of symmetry: 1, 2, m, 2mm, 222, 3, 3mm, 32, 4, –4, 4mm, –42m, 422, 6, –6, 6mm, 62m, 622, 23, 43m, and 432. Among materials electrically polarized because of crystal structure, pyroelectric materials are such that they become electrically polarized with changes in temperature, and the amount of electric charge produced varies according to temperature change. The phenomenon is present in crystallized dielectrics with one or more polar axes of symmetry. In polar crystals, the centroids of positive and negative charges in a unit cell do not coincide. As a result, nonzero polarization, *P*, of the crystal occurs. Pyroelectricity is thus only possible in crystals lacking a center of symmetry and also have a unique polar axis: 1, 2, m, 2mm, 3, 3mm, 4, 4mm, 6, and 6mm. The polarization, *P*, of the crystal changes with temperature, and this is the result of a structure change with temperature. While all pyroelectric materials are piezoelectric, the opposite is not true; for example, quartz is piezoelectric but not pyroelectric. Both zinc oxide, with its wurtzite crystal structure, and PZT, with a ferroelectric perovskite structure, feature an intrinsic dipole moment and show pyroelectricity even in thin film form. In Volume I, Figure 2.10, we already illustrated the pyroelectric effect in $BaTiO_3$. The strongest pyroelectric effect for a thin film is measured with the sol-gel processed tetragonal $PbTiO_3$.[48,86] The pyroelectric constant, *p*, of these materials is defined as the differential change in polarization with temperature in the case of uniform heating, constant stress (σ), and low electric field (*E*) in the crystal (see also row 1 in Volume I, Table 2.1):

$$p = \left(\frac{\partial P}{\partial T}\right)_{\sigma,E=0} \tag{8.66}$$

p is expressed in C/cm^2 K. Thin films of ZnO, PZT, and $PbTiO_3$ all show pyroelectric coefficients similar to the bulk material: $0.95–1.05 \times 10^{-9}$ C/cm^2 K for ZnO, $50–70 \times 10^{-9}$ for PZT, and $95–125 \times 10^{-9}$ for $PbTiO_3$.[48,86]

In the pyroelectric IR (for 10 μm) sensor configuration shown in Figure 8.29a, two sensing elements are connected in electrical opposition and one is shielded from the radiation source, to cancel out the effect of any drift in ambient temperature. A germanium window is used because it is opaque to visible light but transparent at wavelengths near 10 μm. In Figure 8.29b, a sensor IR array that was developed in the 1980s is shown. Detector elements made of lead strontium titanate (PST) or barium strontium titanate (BST) were commonly used. These materials have Curie temperatures close to room temperature. The devices operate at temperatures above the Curie level (in the so-called *paraelectric phase*), and an applied electric field induces the pyroelectric effect. Laser-assisted chemical etching produces an array with a large number of detector pixels. The elements on that array are soldered (or "bump bonded") to a silicon multiplexer. Contemporary versions have as many as 384 × 288 pixels, and a noise-equivalent temperature difference (NETD) is a measure of the detector-noise-limited sensitivity of single-band IR systems) as low as 75 mK. Micromachined thin-film elements with even lower NETDs are under development.

Single crystals of triglycine sulfate (TSG), $LiTaO_3$, and $(Sr,Ba)Nb_2O_6$ are widely used for heat-sensing applications. Using $LiTaO_3$, Hsieh et al.[87] produced sensitive, wide-range anemometers. In all thermal anemometers, the goal is to measure heat loss. Usually, this is accomplished by measuring the resistance of a temperature-sensitive resistor; but with a pyroelectric sensor, heat loss generates a current directly. Gas flows of 0.1 mL/min to over 20 L/min were detected. It should be remembered that only changing temperatures can be detected this way. For a crystal maintained at a constant temperature, the charges will decay rapidly.

Using a thin (<5 μm), freestanding membrane of poly(vinylidene fluoride-trifluoroethylene) [P(VDF/TrFE)] copolymer in a Si frame, Zavracky et al. made a pyroelectric IR detector array.[88] The freestanding film is provided with top and bottom semitransparent Cr/Au electrodes, and incident IR light produces a voltage between these electrodes. The advantage of

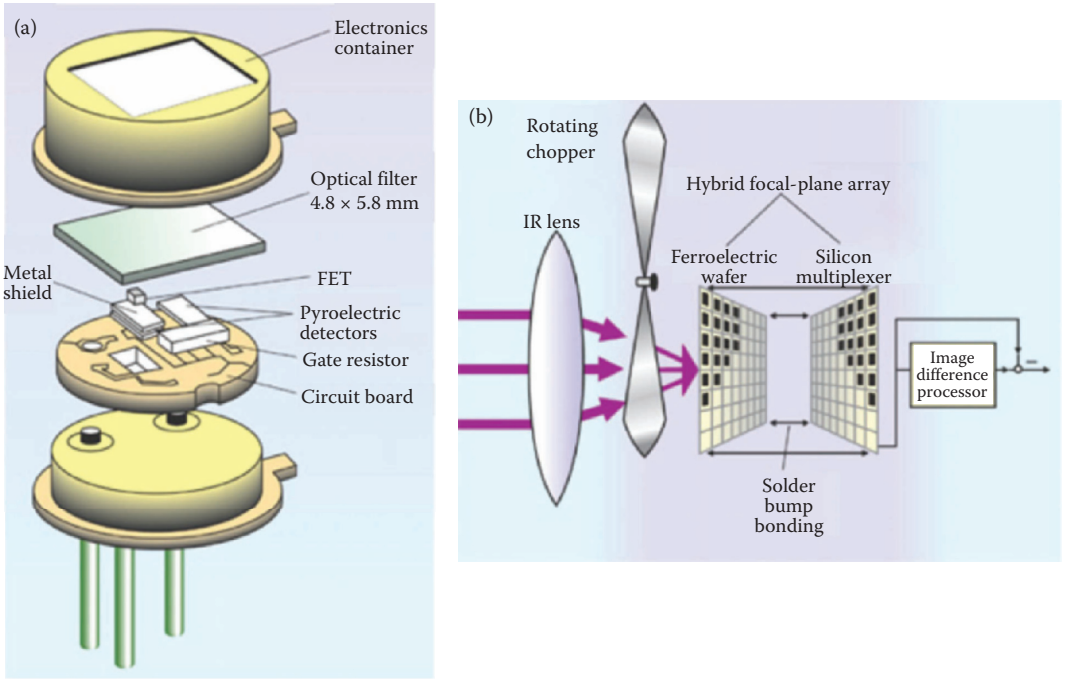


FIGURE 8.29 Single- and multiple-element pyroelectric detectors. (a) In this single-element configuration intended to monitor radiation with a wavelength near 10 μm, two lead titanate pyroelectric elements are used. One detector element is exposed to incoming radiation, another shielded under a metal strip. Both crystals are electrically connected with opposite polarities, to cancel out the effect of any drift in ambient temperature. (b) A ferroelectric-hybrid focal-plane device comprises a lens, usually made of germanium to block visible light, and an array of individual ferroelectric elements, each of which is bonded by tiny balls of solder to elements from a silicon multiplexer. The incident infrared radiation must be periodically blocked, here by a chopper, to ensure that a temperature variation is measured. (From http://www.physicstoday.org/vol-58/iss-8/p31.html.)

this polymeric pyroelectric is that it is well suited for microfabrication, since it can be applied to a substrate by spin on from a solvent-based precursor. Complementary metal-oxide-silicon (CMOS) devices fabricated in the Si frame can be used for amplification and signal processing. To prevent thermal conduction from the sensing film to the substrate, which limits sensitivity, the thin polymer film may be supported by a thin silicon nitride film.

Pyroelectric materials are used in remote control devices, night vision, and thermal sensors.

Thermoelectricity*

Introduction

Thermoelectricity covers three separate phenomena: the Seebeck effect, the Peltier effect, and the Thomson effect. The Seebeck phenomenon, discovered by the Estonian-German physicist Seebeck in 1821, is perhaps the best known, since it is used in thermocouples for the measurement of temperature differences. The effect is that a voltage is created in the presence of a temperature difference between two different metals or semiconductors. The measured voltage is proportional to the temperature difference, and the proportionality constant is known as the *Seebeck coefficient*, *S*.

Independently, the French physicist Peltier discovered the reverse effect, i.e., a current flow across a junction results into heating or cooling. In the absence of a magnetic field, the Seebeck and Peltier effects are the two most important thermoelectric effects.

The Thomson effect is the heating or cooling of a current-carrying conductor with a temperature gradient. Any current-carrying conductor (except for a superconductor) with a temperature difference between two points either absorbs or emits heat,

* Thermoelectric actuators could equally well be discussed under thermal actuators.

depending on the material. The Thomson effect was predicted and experimentally observed by William Thomson (Lord Kelvin) in 1851. The Seebeck effect constitutes a combination of the Peltier and Thomson effects. The Thomson effect is unique among the three main thermoelectric phenomena because it features the only thermoelectric coefficient directly measurable on an individual material. The Peltier and Seebeck coefficients can be determined only for pairs of materials. The Joule heat generated when a voltage difference is applied across a resistor is a loss mechanism due to nonidealities; it is related to thermoelectricity, but it is not generally called a *thermoelectric effect*. The Peltier-Seebeck and Thomson effects are reversible, whereas Joule heating, because of the second law of thermodynamics, cannot be reversible.

Mechanism behind Thermoelectricity

An applied temperature difference between two different metals or semiconductors causes charge carriers, electrons or holes, to diffuse from the hot side to the cold side, similar to the way a classical gas expands when heated. The diffusing charges are scattered by impurities, imperfections, and lattice vibrations (phonons), and since scattering is energy dependent, the hot and cold carriers diffuse at different rates. This creates a higher density of carriers at one end of the material, and the separation of charges creates an electric field. This electric field, however, opposes the uneven scattering of carriers, and an equilibrium is reached with the net number of carriers diffusing in one direction canceled by the net number of carriers moving in the opposite direction. This means the thermopower of a material depends strongly on impurities, imperfections, and structural changes (e.g., different dimensionalities, two-dimensional, and one-dimensional), and the thermopower of a material is a collection of several effects.

The build-up voltage causes a continuous current to flow in the conductors if they form a complete loop. This way, in a thermocouple (TC), a voltage is measured at the leads as the sum of the voltages across both legs of the device (see Figure 8.30). The thermocouple voltage is of the order of several microvolts per degree difference.

The Seebeck coefficient, S (in units of V/K), also known as the *thermopower* or *thermoelectric power*, measures the voltage or electric field induced in response to a temperature difference. If the temperature difference, ΔT, between the two ends of a material is small, then the thermopower of a material is defined as follows:

$$S = \frac{\Delta V}{\Delta T} \tag{8.67}$$

and a thermoelectric voltage ΔV is seen at the terminals. The thermopower S is an important material parameter and determines the efficiency of a thermoelectric material. A larger induced thermoelectric voltage for a given temperature gradient leads to a larger thermoelectric efficiency.

For the thermocouple in Figure 8.30, the measured electromotive force (EMF) is calculated as follows:

$$V = \int_{T_1}^{T_2} [S_A(T) - S_B(T)]dT \tag{8.68}$$

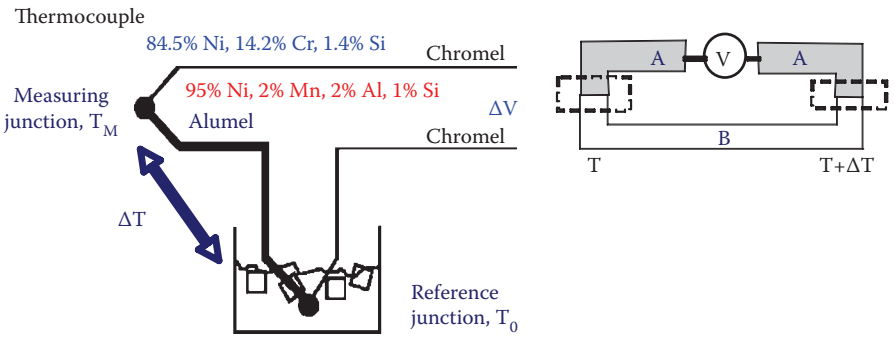


FIGURE 8.30 Thermocouple operation. Dissimilar metal junctions (at *A*/*B* and *B*/*A*) with the reference junction at T_0 (say, an ice bath) and the measuring junction at T_M, lead to a voltage reading *V*. (From a talk by Chien-Neng Liao, NTHU.)

where S_A and S_B are the Seebeck coefficients of the metals *A* and *B*, and $\Delta T = T_1 - T_2$ is the temperature differential between the two junctions. The Seebeck coefficients are nonlinear and depend on the conductors' absolute temperature, material, and molecular structure. A thermocouple can be used to measure a temperature difference directly, or to measure an absolute temperature, by setting one end to a known temperature. Several thermocouples in series are called a *thermopile*. The thermocouple can also be used to extract energy in applications where heat and temperature differences are available "for free" (see energy harvesting in Chapter 9). Thin-film thermocouples have been used, for example, to power a watch on the basis of the temperature difference between its cool front face and the warm skin contact. A disadvantage of this approach is the very low efficiency of the conversion.

Typically, a metal displays a small thermopower because it comes with half-filled bands. In contrast, semiconductors can be doped with an excess amount of electrons or holes and thus can have larger positive or negative values of the thermopower depending on the charge of the excess carriers.

The Peltier effect is illustrated in Figure 8.31. When charge carriers flow from one material into another in which their average transport energy is higher, they absorb thermal energy from the lattice, and this cools the junction between the two materials. One can think of this as the electron gas expanding in the energy space as they enter the second material. The thermoelectric process is reversible, so if the direction of the current is reversed, energy is released to the lattice, and there is heating at the junction.

The definition of the Peltier coefficient, Π_{AB}, is given as the ratio of heat absorption (J_Q) over current flow (J_E) (see Figure 8.31).

$$\Pi_{AB} = \frac{J_Q}{J_E} \tag{8.69}$$

As mentioned above, there is no direct experimental method to determine an absolute Seebeck coefficient or absolute Peltier coefficient for an individual material. However, in 1854, Thomson found two relationships, now called the *Thomson* or *Kelvin relationships*, relating the three thermoelectric coefficients. Therefore, only one can be considered unique. The absolute temperature, *T*; the Peltier coefficient, Π; and the Seebeck coefficient, *S*, are related by the first Thomson relation:

$$\Pi_{AB} = ST \tag{8.70}$$

In turn, *S* and *T* are related to the Thomson coefficient, μ, by the second Thomson relation as follows:

$$\mu = T\frac{dS}{dT} \tag{8.71}$$

The efficiency of a conventional thermoelectric cooler is controlled by a parameter *Z*, given by:

$$Z = \frac{S^2\sigma}{\kappa} \tag{8.72}$$

where σ and κ are the electrical and total thermal conductivities (with both lattice and electrons contributions) of the material, respectively, and *S* is the Seebeck coefficient (in μV/K). More commonly a *dimensionless figure of merit ZT* is defined by multiplying Z with the average temperature T. Larger values of ZT indicate a greater thermodynamic efficiency and *ZT* is a convenient figure for comparing the

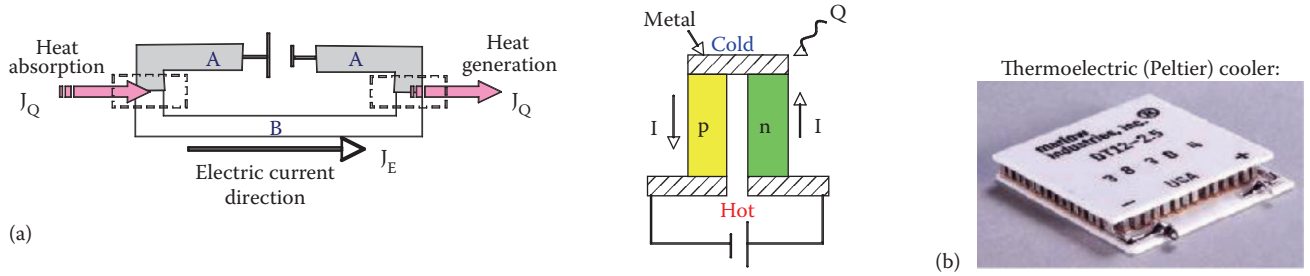


FIGURE 8.31 In a Peltier device, current flow across a junction of dissimilar metals causes a heat flux, thus cooling one side and heating the other. (a) Illustration of the Peltier effect. (b) Picture of a Peltier cooler. (From a talk by Chien-Neng Liao, NTHU.)

efficiency of devices using different materials. One can show that the efficiency for cooling, represented by Φ, and that of heating, symbolized by η, both increase with an increase of:

$$(1 + ZT)^{\frac{1}{2}} = \left(1 + \frac{S^2\sigma}{\kappa}T\right)^{\frac{1}{2}} \qquad (8.73)$$

Consequently, requirements for a good thermoelectric (TE) material are a high figure of merit (*ZT*), i.e., a high Seebeck coefficient (*S*), a high electrical conductivity (σ), and a low thermal conductivity (κ). For good thermoelectric materials one has a *ZT* of ~1, and this remained the upper limit for many years. Thermoelectric compounds include bismuth telluride (Bi_2Te_3), used for refrigeration, bismuth antimony (BiSb), lead telluride (PbTe), and silicon germanium ($Si_{1-x}Ge_x$), used for power generation. These materials today are used for smaller cooling/heating applications; they are not highly efficient but are environmentally friendly and very reliable. On a larger scale, thermoelectric refrigeration would be very interesting as it involves no toxic chlorofluorocarbons (CFCs) and no moving parts and is quick and reliable. The biggest problem is that *ZT* has to be improved from ~0.9 to 3 or higher to become more commercially attractive [*ZT* needs to improve to >1.3 at 300 K for a major impact in electronics cooling and to around 2.5 for a revolutionary impact in air conditioning and power from waste heat (energy scavenging)]. The pioneering work of Hicks and Dresselhaus in 1993 renewed interest in thermoelectrics.[89] Dresselhaus et al. showed that lower-dimensional structures such as nanowires based on bismuth, BiSb, Bi_2Te_3, and SiGe nanotechnology could lead to a higher *ZT*. This group demonstrated that electrons in low-dimensional semiconductors such as quantum wells and wires exhibit an improved thermoelectric power factor. This is due to the fact that electron motion perpendicular to the potential barrier is quantized creating sharp features in the electronic density of states (DOS). The higher density of states at the Fermi level increases σ, and the increased phonon boundary scattering lowers κ (see also Volume I, Chapter 3 and Chapter 3, this volume). Both factors contribute to make *ZT* higher. In Figure 8.32, we show the thermoelectric figure-of-merit *ZT* of a Bi quantum well and quantum wire as a function of size (see also Chapter 3, this volume).

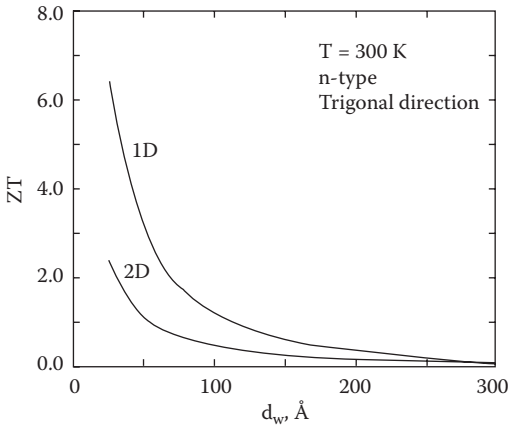


FIGURE 8.32 Thermoelectric figure of merit *ZT* of bismuth quantum well and quantum wire as a function of dimension.[89]

Because of the massive amount of processing know-how and the relatively large Seebeck coefficient, polysilicon is an attractive material for thermocouples. The Seebeck coefficient for 0.38-μm-thick n-polysilicon films with a sheet resistance of 30 Ω/□ is –100 μV/K; for n-polysilicon with a sheet resistance of 2600 Ω/□, it is –450 μV/K. For p-polysilicon with a 400 Ω/□ sheet resistance, it is 270 μV/K. Silicon might also start playing a more and more important role in heating and cooling applications. In 2008, a team from Caltech and a team from University of California, Berkeley, separately reported that they were able to increase silicon's ability to convert heat into electric current by as much as 100 times.[90,91] Bulk silicon is only an average material for thermoelectric conversion, because its thermal conductivity is too high; heat travels across it so well that it is difficult to create the necessary temperature gradient. Bulk silicon at room temperature has a *ZT* of only 0.01, but the Berkeley team increased that to 0.4, and the Caltech team increased it to 0.6. This puts silicon nanowires almost on par with bismuth telluride (see above). Making thermoelectric devices out of silicon, which is much more abundant, cheaper, and easier to handle than bismuth telluride, could help create a large market for these devices.

An excellent 2006 review on nanoscale thermal transport and microrefrigerators on a chip is by Dr. Ali Shakouri.[92]

Electrets

A permanently charged dielectric, an electret, forms an elegant base to improve upon electrostatic actuation by setting up a permanent electric field, thereby reducing or eliminating the need for a large applied bias. There are many different methods to form electrets from high-impedance dielectrics, such as the corona, the liquid contact, and the electron beam methods.[93,94] An electret-based micromachine may incorporate a capacitor charged because of the presence of the electret between the capacitor plates while one of the plates is sensitive to an external force (e.g., pressure). Micromachined electret microphones have been built this way and exhibit an open-circuit sensitivity of about 2.5 mV/μbar at 1 kHz.[95] By charging up SiO_2 with a 300-V corona, electrets in these subminiature microphones exhibit a charge decay time constant amounting to more than 100 years of expected operation. Similarly, an electret-based pressure sensor with a permanently charged polymer foil (commercially available Teflon FEP) as the electret has been manufactured. The foil was charged by electron beam exposure. A maximum sensitivity of 10mA/A/100 mmHg (i.e., the measured relative change of drain current), about ten times higher than the sensitivity of piezoresistive pressure sensors with comparable dimensions ($1 \times 2 \times 0.3$ mm), was determined.

Wolffenbuttel et al.[96] point out that an insulating, electret-implanted rotor circumvents two major limitations of electrostatic motors, namely, the relatively large voltages and the friction between the rotor and the stator during rotation. Replacing the polysilicon rotor in a motor as shown in Figure 8.6b allows propulsion at smaller drive voltages as well as electrostatic levitation and lateral alignment of the rotor, making it virtually contactless. Temperature and humidity sensitivity constitute two of the biggest drawbacks of employing electrets in actuators.

Earlier, we pointed out that bone has piezoelectric properties; Mascarenhas shows that bone can also be made into an electret,[97,98] and Fukada et al.[99] report that plastic electrets applied to bone produce alterations in bone growth.

Electrorheological Fluids

Introduction: Field-Responsive Fluids

Field-responsive fluids refer to colloidal dispersions whose rheological properties are modified under the influence of an external field, typically electric or magnetic. Upon application of such fields, induced dipoles in the dispersion cause the particles to form columns that align with the applied field. Field-responsive fluids are commonly classified as electrorheological (ER), magnetorheological (MR), and ferrofluids (FF). FF are a variant of MR fluids, having as main requirement particle size below 20 nm.

The idea of field-responsive fluids dates back to the eighteenth century. As early as 1779, Gowan Knight attempted to obtain a magnetic fluid by suspending iron fillings in water but the fillings would settle too quickly. In 1896, Duff[100] observed a reversible 0.5% change of viscosity in castor oil and glycerin when applying an electric field of 27 KV/cm. In 1947 Winslow discovered a similar but larger effect when dispersing fine dielectric powders in a nonconducting fluid[101] that he called *electroviscous*.[102] Later, this term was changed to *electrorheological*. A year later, Rabinow[103] published his discovery of a similar effect, but applying a magnetic field instead, giving birth to *magnetorheological fluids*. Since then, a considerable amount of research has been conducted on the topic. Many patents were filed in the 1950s, but the lack of commercial devices and probably the fact that most suspensions suffered from particle sedimentation dampened the initial enthusiasm. In the 1980s, research revived with ER fluids as the main focus since they were thought to be closer to commercial applications. This initial hunch proved to be wrong and MR fluids became the center of attention in the 1990s and eventually led to the application of MR fluids in commercial dampers, loudspeakers, and dynamic seals. Although the interest on ER fluids is still large, they have not yet been applied in commercial devices. The reasons for the switch from ER to MR devices are that the MR effect is an order of magnitude stronger than that of ER fluids and that the required magnetic field levels are easier to achieve.

Given the fact that it is easy to obtain large electric and magnetic fields in the microdomain, research on adapting ER and MR fluids to MEMS should follow soon.

Here we only discuss electrorheological fluids. MR fluids and FF are covered below, under magnetic actuators.

Electrorheological Fluids

ER fluids are a class of colloidal dispersions, usually dielectric particles in a nonconducting fluid, that exhibit large reversible changes in their viscosity or flow rate; that is, in their rheological behavior when subjected to external electrical fields (typically of 1–4 KV/mm). These changes in rheological behavior are typically manifested by a pseudo-phase change from the liquid to a viscous, semisolid, gel-like state and result in a dramatic increase in flow resistance (the viscosity might increase by several orders of magnitude). The total power required is rather low, and the response of ER fluids to an electrical stimulus is typically in the millisecond range. The electrical field induces dipoles in the particulate phase, that is, the fine dielectric powder suspended in nonconducting fluid, and forges an interaction between the individual particles. The induced dipole then aligns with the field to minimize dipole-dipole interaction energy. Eventually, columnar structures like those presented in Figure 8.33 are formed.[104,105]

Willis Winslow, who discovered and patented[101] the electrorheological phenomenon in 1947, worked with silica gel in mineral oil. More recent formulations are based on suspensions of particles of polymers or ceramic materials in silicone oil. Typical products proposed with these actuators include clutches and transmission systems, hydraulic valves, and vibration isolation systems such as engine mounts and shock absorbers. The bulk conductance of the carrier fluid can present a difficulty to power usage; true nonconducting fluids work best. Problems with separation of fluid and particles due to evaporation, sedimentation, centrifugal forces, and electrophoresis can lead to device failures. Wear due to abrasion from particles can also cause device failure. Electrorheological fluids are very temperature sensitive, and both boiling and freezing are major issues. Furthermore, ER fluids are sensitive to impurities since polarization mechanisms might be changed by them.

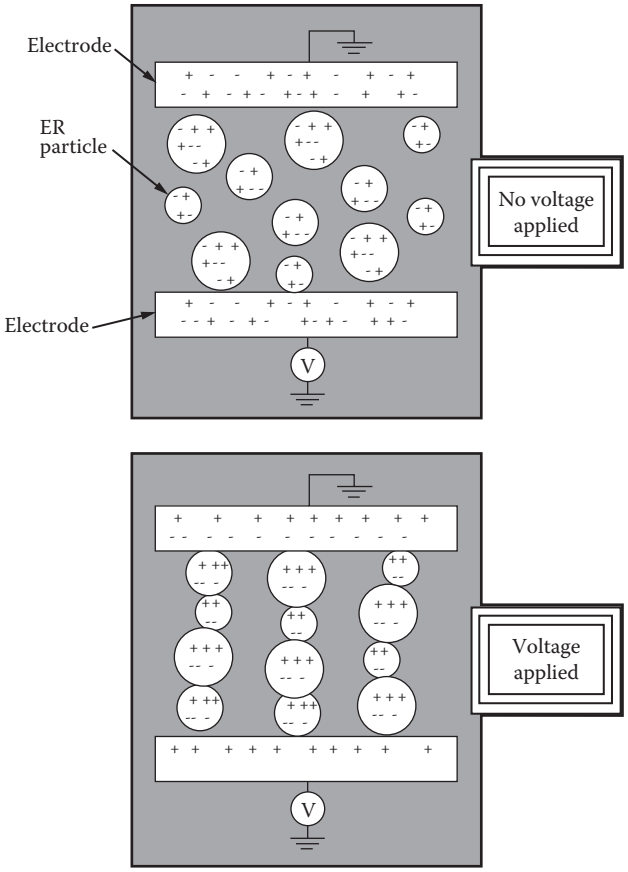


FIGURE 8.33 The electrorheological phenomenon. (After Gandhi, M.V., and B.S. Thompson. 1992. *Smart materials and structures*. London: Chapman & Hall.[32])

In 2003 a "giant electrorheological effect" was reported by Wen et al.,[106] who demonstrated an ER effect of the order of magnitude of an MR one. Electrorheological devices have not yet reached the commercial market, while MR fluids are already used in some applications (see "Magnetorheological Fluids" below).

Magnetic Actuators

Introduction

In the macroworld, electromagnetic (EM) forces dominate the development of actuators such as conventional motors. On the other hand, large electrostatic motors such as the 1889 capacitive motor made by Zipernowsky[107] are rarities rather than commodities. Because of the three-dimensional nature of magnets and solenoids, EM systems are difficult to micromachine using planar integrated

circuit (IC) processes. Electrostatic motors are simpler and more compatible with IC fabrication, and electrostatics often scale more favorably into the microdomain. These factors explain the current preponderance of electrostatic actuators in micromachining. We shall learn from a detailed comparison of electrostatic versus magnetic actuation that both present advantages and disadvantages and that both have severe problems when high-power output is demanded from the miniaturized systems in which they are embedded.

Principle behind Magnetic Actuation

A good starting point to understand magnetic actuators is Ampère's circuital law, used to calculate the magnetic field strength **H** (in A/m) around a current-carrying wire (Volume I, Figure 5.25). That law, we saw in Volume I, Chapter 5, states that the line integral of an electric current density *J* running through a straight wire generates a magnetic field strength **H** (in A/m) around that wire. The integral of **H** along the circular path *C* is equal to the current *I*, which in turn can be obtained as the integral of the current density **J** across the cross-section of the wire *S*:

$$\oint_c \mathbf{H}\cdot d\mathbf{l} = \int_s \mathbf{J}\cdot d\mathbf{s} \qquad \text{(Volume I, Equation 5.28)}$$

When applying this to a solenoid inductor as shown in Figure 8.34 one obtains the following:

$$\oint_c \mathbf{H}\cdot d\mathbf{l} = NI \tag{8.74}$$

where *N* is the number of windings in the solenoid. The electric inductance *L* (in *H*) of the solenoid equals $N\Phi/I$ (see Volume I, Table 5.1), where Φ is the magnetic flux. With $\Phi = BS = \mu HS$ (with μ the magnetic permeability and *B* the magnetic induction field) and, using Equation 8.74, we obtain:

$$\Phi = \mu\frac{NI}{1}S \tag{8.75}$$

where *l* is the length of the solenoid, so that for the inductance *L* of the solenoid we get the following:

$$L = \mu\frac{N^2}{1}S \tag{8.76}$$

We can now calculate the stored magnetic energy (U_m) and energy density (U'_m) of the solenoid. The electromotive force, or V_{EMF}, generated by the coil is given by (see also first term on the left in Equation 8.36):

$$V_{EMF} = -L\frac{dI}{dt} = -N\frac{d\Phi}{dt} \tag{8.77}$$

So that for the stored energy:

$$U_m = \int -(V_{EMF})dq = \int_0^I L\left(\frac{dI}{dt}\right)dq = \frac{1}{2}LI^2 \tag{8.78}$$

or:

$$U_m = \frac{1}{2}\mu\left(\frac{N}{1}\right)^2 S1\frac{B^2}{\mu^2(N/1)^2} = \frac{B^2}{2\mu}S1 \tag{8.79}$$

and for the energy density we divide by the volume (V = Sl) to obtain the magnetic energy density, U'_m, as follows:

$$U'_m = B^2/2\mu \tag{8.80}$$

which, in the case of an air gap, was introduced earlier as Equation 8.62, where *B* is the magnitude of the magnetic flux density in the gap and μ_0 the magnetic permeability of free space. Earlier, we also mentioned that for magnetic actuators in air, the gap energy density by pushing *B* into saturation (about 1.5 T) leads to 950,000 Jm^{-3}. The energy density stored in the rotor/stator gap for magnetic, electrostatic, piezoelectric, and SMA actuation (see below) was compared in Table 8.8.

By analogy with the electrostatic energy density from Equation 8.2:

$$U' = \frac{1}{2}\varepsilon E^2 = \frac{1}{2}\mathbf{D}\cdot\mathbf{E} \tag{8.2}$$

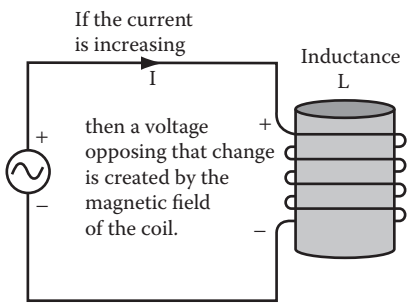


FIGURE 8.34 Solenoid inductor.

the magnetic energy density can be written out as follows:

$$U'_m = \frac{B^2}{2\mu} = \frac{1}{2}\mathbf{H}\cdot\mathbf{B} \tag{8.81}$$

And by analogy to the electrostatic surface force F, given earlier as follows:

$$F = -\nabla U \tag{8.3}$$

we calculate a translational magnetic force, F_m, as (see Figure 8.35, where translation is in the y direction):

$$F_m = -\nabla U_m \tag{8.82}$$

with:

$$U_m = \frac{1}{2}LI^2 = \frac{1}{2}\mu N^2 l S I^2 \tag{8.79}$$

we obtain for a small movement, dy:

$$dU_m = \frac{1}{2}N^2 S I^2\,[\mu(1-dy) + \mu_0 dy] \tag{8.83}$$

Or, with $\mu = \mu_0(1+\chi)$, we calculate the following for the magnetic restoring force along the y-axis:

$$F_m = \frac{dU_m}{dy} = -\frac{1}{2}N^2 S I^2 \mu_0 \chi \tag{8.84}$$

To garner some magnetic actuator scaling insights, we rewrite Equation 5.28 of Volume I as follows:

$$\oint_c \mathbf{B}\cdot d\mathbf{l} = \mu_0 I = \mu_0 \int_s \mathbf{J}\cdot d\mathbf{s} \tag{8.85}$$

This equation describes the creation of a magnetic field by a current, with S defining the area of the surface bounded by C, and J the current density. We use this expression for an example calculation of scaling of the solenoid shown in Figure 8.36.[108] As the system becomes smaller, area ds decreases more rapidly than dl, and B decreases unless the current density J is increased. We refer to the appendix in Trimmer[109] for a more detailed analysis. His results indicate that different assumptions about current density and heat transfer lead to different scaling laws. When the current density is kept constant and the size of the system is decreased, the force between two EM wires or coils scales as (l^4). In other words, a size reduction of 10 means a magnetic force reduction of 10,000. Electrostatic force scales as l^2, and from Chapter 7, on scaling, we remember that the lower the coefficient in l the better the scaling behavior into the microdomain. So from this vantage point, electrostatic actuators scale better than EM ones. The scaling results of the interaction between a coil and a permanent magnet are somewhat better: (l^3). To improve the situation further, we can increase the current density. The resulting heat, as we will see below, is more effectively removed from microstructures, thus avoiding overheating and improving the scaling. Constant heat flow per unit area of the windings results in l^3 ($l^{2.5}$ with a permanent magnet). A constant temperature difference between windings and environment yields l^2 (and also l^2 for the permanent magnet case). For the last magnetic

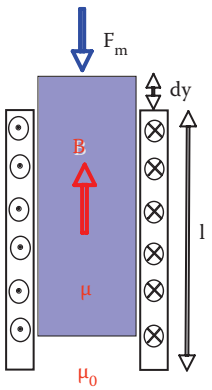


FIGURE 8.35 Magnetic force in a translational motion.

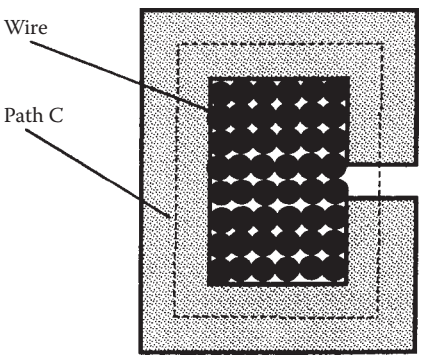


FIGURE 8.36 The line integral of B·dl over a closed curve C is equal to μ_0, the magnetic permeability of free space times the current through this loop C. As the system becomes smaller, the area decreases more rapidly than the length of the dashed line. [After Trimmer, T. 1990. In *Proceedings, Integrated Micro-Motion Systems: Micromachining, Control, and Applications (3rd Toyota Conference),* 1–15. Aichi, Japan.[108]]

case, where the force scales as (l^2), the power that must be dissipated per unit volume scales as (l^{-1}), or, when the scale is decreased by a factor of 10, 10 times as much power must be dissipated. The use of superconductors could eliminate this problem (see below).

Examples of Magnetic Actuators

Introduction

The magnetic actuation examples we consider here are magnetic levitation for making valves, magnetic micromotors, and a bead array counter (BARC).

Levitation Valve

Friction and wear in miniaturized systems primarily relate to the surface contact between solids, in particular in bearing surfaces supporting the load of the micromachinery. They scale as l^2 and become increasingly important in the microdomain (see also Chapter 7). The friction in early electrostatic micromotors was actually found to be comparable to the friction of brake materials on cast iron.[110] Levitation eliminates wear and friction, two factors affecting reliability and control. Levitation is obtained from magnetic,[111] electrostatic,[112,113] and fluidic forces.[114] Here, we consider magnetic levitation for making a valve.

Levitation with magnetic devices, or *maglev*, can be achieved by various methods, such as with permanent magnets, electromagnets (including superconducting magnets), and diamagnetic bodies. Working with permanent magnets leads to better scaling behavior than working with electromagnets (see above). Wagner et al.[115] levitated a small, rare-earth permanent magnet out of the plane of a silicon substrate equipped with a planar coil as shown in Figure 8.37b. This moving structure may be used to close off a hole in a valve seat, making for a magnetically actuated valve. The permanent magnet was glued onto a silicon micromachined thin plate suspended by a thin silicon spring made from suspending beams parallel to the substrate edge. The monolithically integrated seventeen-turn planar coil was used to generate a magnetic field force, which forces the magnet on its spring suspended platform

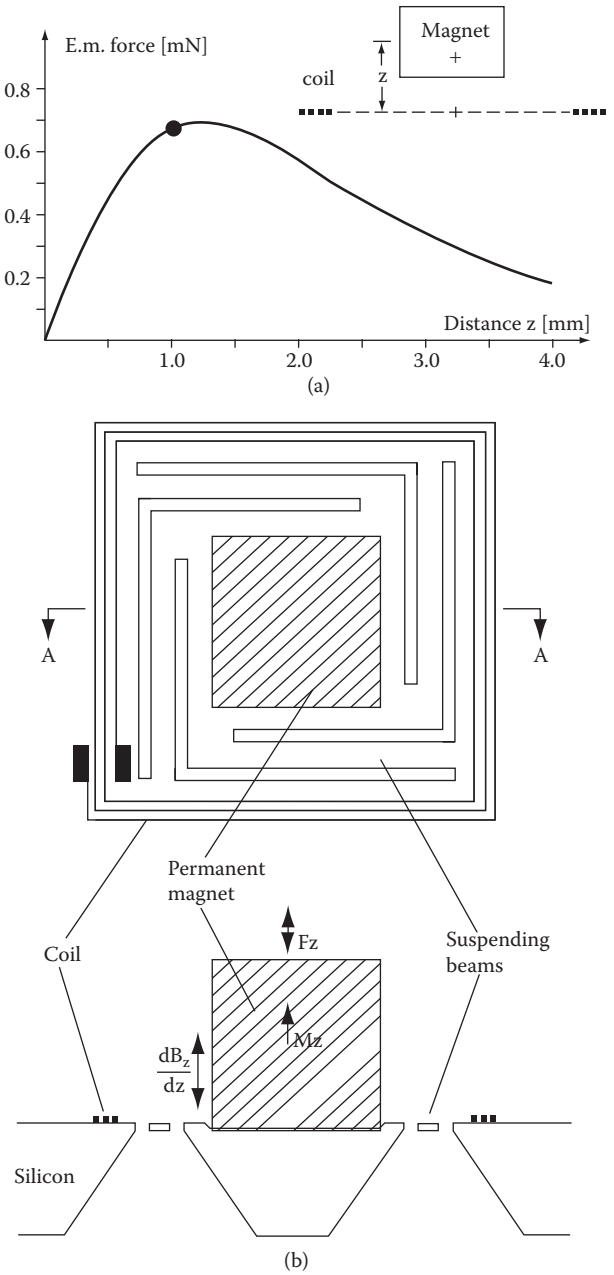


FIGURE 8.37 Magnetic actuator. (a) Electromagnetic force on a magnet. The dot indicates the magnet position in the fabricated device. (b) Schematic of a vertical magnetic microactuator with integrated planar coil and hybrid-mounted permanent magnet. [After Benecke, W. 1991. Silicon-microactuators: Activation mechanisms and scaling problems. In *1991 International Conference on Solid-State Sensors and Actuators (Transducers '91)*, 46–50. San Francisco, CA.[116]]

to move vertically (z-direction).[116] Using planar IC technologies to manufacture the coil ensures an optimum heat flow within the device (see below under Thermal Actuators). If μ_z, the magnetization of the permanent magnet, is independent of the magnetic

induction field, B, the vertical force acting on the permanent magnet is given by the following:

$$F_{m,z} = \mu_z \int \frac{dB_z}{dz}\, dV \tag{8.86}$$

where B_z represents the vertical component of the magnetic field produced by the planar coil. The magnetic force is proportional to the volume of the magnet. Thin magnetic layers will, in general, be insufficient to generate high forces. Problems with commercially available NdFeB permanent magnets include the fact that these magnets, like other rare-earth-based magnets, incorporate on their side (nonpole) surfaces a magnetically reversed layer of about 20 µm thick. This surface demagnetization limits the miniaturization of these permanent magnets to a few hundred micrometers; the smallest commercially available size is about 0.3 mm. The deflection of the highlighted actuator is given by Hooke's law as follows:

$$\Delta z = \frac{F_{m,z}}{k} \tag{8.87}$$

with k symbolizing the spring constant. The magnitude of the magnetic field depends on the current density achieved in the windings generating the field, and this current in turn is limited by heating of the solenoid metal (gold, in this case). Sheet resistance of thin-film conductors, electromigration, and thermal and geometric constraints all conspire to reduce the achievable magnetic energy density. Wagner et al.[115,117] assume constant current and very efficient power dissipation and derive a magnetic force with quadratic scaling and a deflection exhibiting linear scaling. To obtain constant current, this group has the number of coil turns scaling linearly with the cross-section of the coil wire. Quadratic force scaling, also found for electrostatics, lends itself well to the microrange. In practice, the magnetic force in the example increases rapidly at first and then decreases slowly with z, exhibiting a maximum value of 0.68 mN at a distance of 1.25 mm from the coil plane (see Figure 8.37a).[116] Increasing the volume of the permanent magnet thus works only up to a point, as the magnetic field decreases slowly with z. For a driving current range between –300 and +300 mA, an elevation of the permanent magnet of 143 µm is achieved, resulting in a mean slope of 24 µm per 100 mA. This example indicates that at least for linear actuators, magnetic actuation can be designed to scale as well as electrostatics and that these actuators are not too hard to implement when allowing for some final manual assembly.

Along the same lines, a meso (intermediate) scale, electromagnetically actuated, normally closed valve was realized using low-temperature cofired ceramic (LTCC) tape technology by Gongora-Rubio et al.[118] In this hybrid device, LTCC green tape technology is used for the electromagnet [several layers of planar silver spirals on alumina (Dupont 951 series)] and for the fluid manifold (alumina) fabricated at the same time. Anisotropically etched Si is employed for the planar rectangular spring (similar to the Benecke spring in Figure 8.37b) and a high-energy SmCo for the mini-permanent magnet. Using a 900-Gauss SmCo (1-mm diameter) magnet, a 200-µm deflection of the rectangular planar Si spring was obtained. The square spiral spring is covered with a polysiloxane film; the rectangular spring butts up against a polysiloxane valve seat on the ceramic coil stack and is lifted of the valve seat when powered up.

Magnetic Micromotors

Electromagnetic (EM) microactuators with glued magnets of millimeter dimensions have been used for different types of vertical, torsional, and multiaxial actuation. Here we consider EM micromotors and compare them with electrostatic micromotors.

Making micromotors has fascinated the MEMS community from day one. To witness, on November 28, 1960, William McLellan collected $1000 from Richard Feynman for having made the first operating electric motor only 1/64 inch in size.[119] The handmade motor was a more or less standard, two-phase, permanent-magnet motor.

Actuators in general, we remarked earlier, often do not scale advantageously in the microdomain. The torque of both electrostatic and electromagnetic motors depends on the shearing stress on the rotor integrated over its whole area, and force production is proportional to changes in the stored energy in the gap between rotor and stator (see Equation 8.82). The amount of force that may be generated per unit

substrate area is thus proportional to the height of the motor. Large-aspect ratio rotor structures are therefore desired for both types of motors (or actuators in general for that matter). Torque is a volume effect, and small, flat-surfaced micromachined micromotors have a only a small force available for actuation. For most practical applications, motors with torques larger than 10^{-6} to 10^{-7} Nm are required,* but, as we saw above, the torque generated with a typical surface-micromachined electrostatic motor is only in the nano-newton-meter or pico-newton-meter range (see Figure 8.6b for a typical surface micromachined motor). In practice, the performance with respect to torque and speed of a traditional miniature magnetic motor with a 1-mm diameter to 2-mm-long permanent magnetic rotor is incomparably better than the surface micromachined motors described here and in Volume II, Chapter 7.[120]

With LIGA (see Volume II, Chapter 10) or a combination of LIGA (or pseudo-LIGA) and precision engineering, large aspect ratio magnetic motors can be built that do produce torques that are in the 10^{-6} to 10^{-7} Nm range. As an example, in Figure 8.38 we show one of the best-performing micromotors, fabricated using the sacrificial LIGA process in combination with traditional precision machining. Essentially, a permanent magnetic rotor follows a rotating magnetic field in the stator coils. Both rotor and stator in this micromotor are made from pure nickel. Such a motor with a diameter of 285 μm operates at speeds slightly above 30,000 rpm, with a current excitation level of 600 mA. No evidence of deterioration was evident even after 50 million rotations.[121] The motor manufacture illustrated here shows how LIGA micromachining and precision engineering constitute complementary techniques for producing individual parts that are then assembled afterward. Only components with the smallest features are produced by LIGA; other parts are produced by traditional precision mechanical methods. The cost of producing monolithic magnetic LIGA micromotors urges one to investigate less expensive LIGA-like technologies or

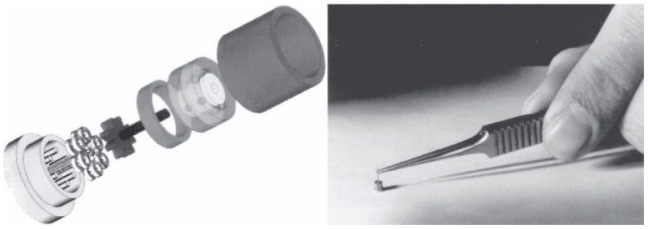


FIGURE 8.38 Large-aspect-ratio magnetic motor. (Courtesy of IMM, Germany.)

hybrid approaches (LIGA combined with traditional machining) as more accessible and adequate alternatives.

A University of Wisconsin team[122] demonstrated a series of magnetic LIGA motors, typically consisting of a plated nickel rotor that is free to rotate about a fixed shaft. Some of the motors made by this group can also be run electrostatically. Given their dramatically increased torque, these motors are again much better than their similar sized surface micromachined cousins. Using a pseudo-LIGA approach, Ahn et al. made a meander-type induction coil involving polyimide and electroplated high-permeability Ni (81%)-Fe (19%).[123] An EM motor based on this approach rotated at speeds up to 500 rpm with applied currents of 300–500 mA and a driving voltage of less than 1 V. The torque predicted for a stator current of 500 mA was 1.2 μN.

High-aspect-ratio techniques, we pointed out, also benefit torques in electrostatic micromotors. Wallrabe et al.,[124] for example, present design rules and tests of electrostatic LIGA micromotors. Minimum driving voltages needed were measured to be about 60 V and optimized designs are expected to deliver torques in the micro-newton-meter range. Suppose now that we use the same high-aspect-ratio machining techniques for both magnetic and electric micromotors, which one is then expected to be the best? For this we turn our attention again to the maximum energy density achievable in the stator-rotor gap. In conventional motors using iron, magnetic induction is limited to 1.5 T because of saturation, producing an energy density of about 9×10^5 Jm^{-3}. This is already more than twice the achievable electrostatic energy in a 1-μm gap (see Table 8.8). Moreover, it is important to remember that, in the case of electrostatic motors, there is little room for improvement, since the electrostatic

* As a reference point, to drive a cylindrical load of several hundred microns in diameter and several mm long, the required torque is of the order of 10^{-5} Nm.

energy density assumed above is close to what is achievable in vacuum. In contrast, there is plenty of room for improvement in magnetic energy density. Thin ferromagnetic films have yielded 2-T fields, and substantially higher fields are generated, at much larger expense, using superconductors (see below).[125] For example, using small-bore 10–15-T superconducting magnets, the achieved energy density of $4–9 \times 10^7$ J m^{-3} is roughly two orders of magnitude larger than the highest possible electric field density. For now, though, the 10–15-T superconducting magnets fall outside the realm of micromachining.

The efficiency—the ratio of power used in performing a desired task to the total amount of power consumed for a current-driven magnetic motor—will be worse than for a voltage-driven electrostatic device. However, Busch-Vishniac argues, since friction in small gaps, rather than electrical losses, are the major losses in microactuators, mechanical losses render nearly all types equally inefficient.[125] Moreover, since the gap in magnetics may be larger, friction is actually easier to avoid in magnetic actuators. One could add that a wider separation gap would also make the actuator less sensitive to dust and humidity.

In addition to the problem of minuscule amounts of power being produced by minuscule motors of both types, that of heat dissipation in a micromotor below a certain size needs to be solved. Busch-Vishniac calculated that a generic problem of heat dissipation exists if microfabricated microactuators continue to shrink in size without a concomitant reduction in the amount of power used in the system. For instance, a motor dissipating 100 mW should not have a rotor smaller than 20 μm, as the temperature might rise above 250°C, resulting in thermal breakdown of the metals in the microstructure. Classical models for heat convection were used to reach this conclusion. A more accurate model would result in a reduction of the heat lost by convection as the effective viscosity in the gap is greater than that predicted by the classical model (see section on heat transfer in Volume I, Chapter 6). In other words, the results generated with the classical model should be viewed as a conservative approximation; real microdevices might run hotter yet.

The more immediate challenge for both types of motors, however, is in making them in batch fashion. MEMS activities in micromachined motors today have generated plenty of new insights into phenomena such as stiction, friction, and wear of various materials but have not yet produced viable, batch-produced micromotors that can compete with a proven product such as the Panasonic piezoelectric device pictured in Figure 8.23a.

The above analysis seems to put magnetic actuation at an advantage over electrostatic actuation and even suggests that micromachined actuators are not always possible or useful. We compare magnetic and electrostatic actuators in general in a separate section, "Comparing Magnetic with Electrostatic Actuation."

The Bead Array Counter

In the bead array counter (BARC), we describe the use of a magnetic field gradient to levitate and remove paramagnetic microbeads that are not specifically bound to the surface of an array of giant magnetoresistive (GMR) sensors.[126] Recent developments in thin-film magnetic technology has resulted in films that exhibit a large change in resistance in response to a magnetic field. This phenomenon is known as GMR to distinguish it from conventional anisotropic magnetoresistance (AMR). AMR resistors exhibit a change of resistance of <3%, but GMR materials achieve a change of as much as 10–20%. GMR films have two or more magnetic layers separated by a nonmagnetic layer. Because of spin-dependent scattering of the conduction electrons, the resistance is maximum when the magnetic moments of the layers are antiparallel and minimum when they are parallel.

The BARC uses molecular recognition, such as DNA hybridization, magnetic microbeads, and magnetoresistive sensors (such as giant magnetoresistive sensors or GMR sensors), and the application of a magnetic field to detect and identify binding of a molecule with a molecular receptor. In one particular embodiment, DNA probes are patterned directly above an array of GMR sensors, as shown in Figure 8.39. Sample containing complementary DNA hybridizes with the probes on the surface. Magnetic or paramagnetic beads are labeled with a binding material (e.g., streptavidin),

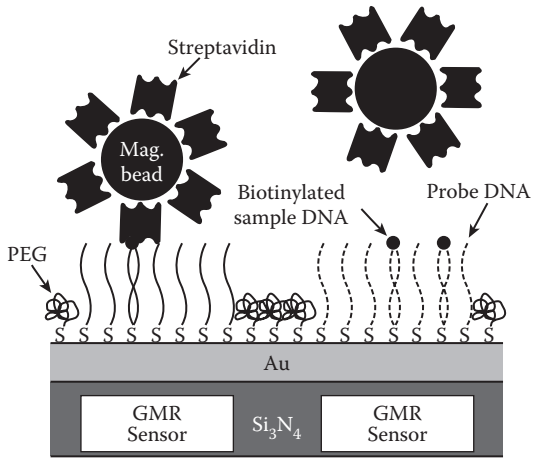


FIGURE 8.39 Schematic diagram of the bead array counter (BARC) chip surface chemistry and hybridization assay. Thiolated DNA probes are patterned onto a gold layer directly above the giant magnetoresistive (GMR) sensors on the BARC chip. Biotinylated sample DNA is then added, and it hybridizes with the DNA probes on the surface when the complementary sequence is present. Unbound sample DNA is washed away. Streptavidin-coated magnetic beads are injected over the chip surface, binding to biotinylated sample DNA hybridized on the BARC chip. Beads that are not specifically bound are removed by applying a magnetic field. Bound beads are detected by the GMR sensors. (Based on Edelstein, R.L., C.R. Tamanaha, P.E. Sheenan, M.M. Miller, D.R. Baselt, L.J. Whitman, and R.J. Colton. 2000. The BARC biosensor applied to the detection of biological warfare agents. *Biosens Bioelectron* 14:805.[126])

and they bind to biotinylated sample DNA hybridized on the BARC chip. Nonspecifically adherent microbeads are removed by means of a controlled magnetic field—applying the magnet removes >95% of these microbeads. A change in the output of the magnetic field sensors indicates the presence of magnetic particles bound to sensors and thereby indicates the presence and concentration of target molecules in the sample. The detection of the presence of beads requires a uniform magnetic field, but removing the beads requires a nonuniform large field gradient. The latter is accomplished by moving a portion of the electromagnetic core closer to the chip with a plunger.

Comparing Magnetic with Electrostatic Actuation

The debate about the relative merits of electrostatic versus magnetic actuation, especially for driving micromotors, has preoccupied many MEMS researchers over the last two decades.[119,127] Fujita and coworkers, for example, have argued that electrostatic actuation is preferred,[128–130] pointing to the following attributes of surface micromachined electrostatic actuators:

1. Thin insulating layers such as SiO_2 or Si_3N_4 exhibit breakdown strengths as high as 2 MV/cm. The power density in this field is 7×10^5 J/m^3; this value is equal to the power density of a 1.3-T magnetic field. The contracting pressure induced by this field is 1.3 MPa. A voltage of about 100 V is sufficient to generate the strong fields mentioned.
2. The electrostatic force is a surface force exhibiting a favorable scaling law. The actuation is simple, as it involves only a pair of electrodes separated by an insulator.
3. The electrostatic actuator is driven by voltage, and voltage switching is far easier and faster than current switching (as in EM actuators). Energy loss through Joule heating is also lower.
4. Weight and power consumption are low.

However, earlier, the comparison between electrostatic and magnetic micromotors demonstrated that many factors besides scaling need to be considered when deciding upon a certain type of actuation principle. Whereas in some cases, the magnetic power might scale disadvantageously into the microscale, the absolute magnetic forces achievable are much larger.

While there is some validity, especially in the case of micromotors, to the arguments favoring magnetic actuators, the criticism of micromachined electrostatic actuators is too harsh where linear micromachined electrostatic devices such as resonators are concerned. In the latter application, power output is of little concern, and these devices have found a major application, for example, in the analog devices accelerometer described in Volume II, Example 7.1. Similarly, for most optical applications such as digital mirror technology from TI (Volume II, Example 7.2), the amount of power required to deflect a laser beam with a micromachined mirror is minimal.

Table 8.11 Salient Points of Electromagnetic Actuators

1. Maximum stroke: 10^{-1} μm
2. Maximum speed: 10 kHz (see Figure 8.14)
3. Intermediate force density: 10^{-2}
4. Complex to fabricate
6. Large footprint: 1–10 mm
7. Special drive conditions: 1.5 T
8. Maximum electrostatic energy density, U'_m (in J/m^3): 9.5×10^5 Jm^{-3}

For many years, micromagnetics have found major use in Hall chips and thin-film heads for magnetic discs (see Figure 1.1d). Besides magnetic micromotors, magnetic valves, and BRAC chips, described above, micromagnetic efforts continue on making denser thin-film magnetic heads;[131] high-performance magnetographic printing heads;[132] electronic switching components;[133] different types of vertical, torsional, and multiaxial actuators;[117] and contactless magnetic transmission of force to ferrofluids (FF) (see "Ferrofluids" below).[115] The pros of magnetic actuation include high force and stroke for the size of the actuator, the ability to attract and repel, multiaxis capability, and the linear current response. The most outstanding advantage of magnetic actuators is the long range of the force. With dimensions above 1 mm and for larger forces, actuators based on permanent magnets become a good choice (see Table 8.11). The cons might be listed as large actuator footprint (approximately a centimeter), requirement of external field sources, complex microscale manufacturing/assembly processes, high-power consumption, and interactions of the magnetic field with nearby system components. Incorporating microfabrication of a permanent magnet into common MEMS fabrication processes remains an active field of research (see also Figure 8.14).

To conclude this comparison of electromagnetic versus electrostatic actuators, we summarize the key points in Table 8.12. Our conclusion regarding the status of microactuators is that electrostatics is useful in dry environments and over limited distances and that electromagnetics is still difficult to collapse into integrated structures.

Magnetostriction

Introduction

The magnetostrictive effect is the contraction or expansion of a material under the influence of a magnetic field and the inverse effect of changes in magnetization due to a stress in the material. It is the magnetic equivalent of electrostriction (see above). This bidirectional effect between the magnetic and mechanical states of a magnetostrictive or piezomagnetic material presents one with another transduction capability that can be used for building actuators and sensors. It was the Scotsman James Joule (1818–1889) who, in 1842, found that when a nickel rod is magnetized it contracts very slightly (the effect is also called the *Joule effect*). A very early use of electrostrictive materials (1861) was for the

Table 8.12 Comparison of Electrostatic versus Magnetic Actuators

	Electrostatic	Magnetic
Field energy density	4×10^5 Jm^{-3}(max)	4×10^7 Jm^{-3} (10 T)
Scaling	(l^2)	(l^3) (constant current)
Gap contamination sensitivity	Very sensitive to humidity and attracts dust	Fairly insensitive
Integrated circuit compatibility	Good	Not very good
Range	Short range	Long range
Power efficiency	Very good in the absence of mechanical friction	More power consuming even in the absence of mechanical friction
Implementation of levitation schemes	Possible	Easier implemented because larger gaps are possible
Miniaturization	Excellent	Difficult
Complexity	Low	High
Control	Voltage switching control is faster, easier to make, and more efficient	Current switching control is less efficient, more complex, and less efficient

generation of sound and ultrasound, and one of the first telephone earpieces was magnetostrictive. The magnetostrictive effect is actually familiar to all of us in the humming of transformers (expanding and contracting of the coils).

How Does Magnetostriction Work?

The magnetostrictive coefficient, ε, is the fractional change in length ($\Delta l/l$) as the magnetization of the material increases from zero to its saturation value ($\Delta l/l = \varepsilon$) (see Figure 8.40). Its value is typically in the parts-per-million range (e.g., 40×10^{-6}). It can be positive or negative in a direction parallel to the magnetic field and is independent of the direction of that field, indicating a square-law type of relationship, such as:

$$\varepsilon = \frac{\Delta 1}{1} = cB_0^2 \qquad (8.88)$$

where ε = static strain
B_0 = flux density
c (m^4Wb^{-2}) = a material constant

For a small flux density ($B \ll B_0$), taking the derivative of Equation 8.88, one derives the basic magnetostrictive strain equation:

$$\varepsilon = \beta B \qquad (8.89)$$

where $\beta = 2cB_0$ defines a magnetostrictive strain constant with dimensions of m^2Wb^{-1}. As the magnetic field, *B*, increases, the length change saturates and increasing the field strength further brings about no additional change. As in the case of piezoelectricity, the inverse effect exists as well; a changing flux density results when modulating the stress on a piezomagnet. This reverse effect is called the *Villari effect*.

The magnetostriction phenomenon is attributed to the rotations of small magnetic domains in the material, randomly orientated in the absence of a magnetic field. The orienting of these small domains by the imposition of a magnetic field results in the development of a strain field (Figure 8.41).

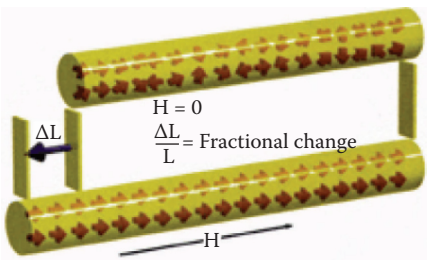


FIGURE 8.40 Magnetostrictive materials: the Joule effect.

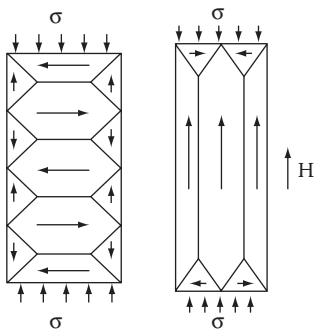


FIGURE 8.41 Magnetostriction. With a compressive load applied to a magnetostrictive material, the domain structure is oriented perpendicular to the applied force. As a magnetic field is introduced, the domain structure rotates, producing the maximum possible strain in the material.

As the intensity of the field increases, more and more domains line up with the field until saturation is reached. Magnetostrictive actuation, like piezoelectric actuation, is important when large forces must be obtained over small distances. Applications for magnetostrictive devices include high-force linear motors, positioners for adaptive optics, active vibration or noise control systems, medical and industrial ultrasonics, pumps, and sonar. Both find applications as sonar sensors, vibration dampers, and tunable-compliance materials.

The Wiedemann effect describes what happens when an axial magnetic field is applied to a magnetostrictive wire, with a current passing through it, i.e., a twisting occurs at the location of the axial magnetic field (Figure 8.42). The twisting is caused by interaction of the axial magnetic field, from an

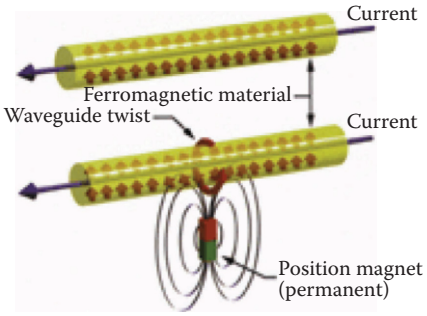


FIGURE 8.42 The Wiedemann effect: the twisting due to an axial magnetic field applied to a ferromagnetic material carrying a current.

Table 8.13 Magnetostrictive Materials with Their Magnetostrictive Coefficients

Material	Saturation Magnetostriction (μm/m)
Nickel	−28
49Co, 49Fe, 2V	−65
Iron	+5
50Ni, 50Fe	+28
87Fe,13Al	+30
95Ni, 5Fe	−35
Cobalt	−50
$CoFe_2O_4$	−250

external permanent magnet, with the magnetic field along the wire, due to the current in the wire. The inverse effect, that of creation of an axial magnetic field by a magnetostrictive material when subjected to a torque, is known as the *Matteucci effect*. Together, these effects are used in torque magnetostrictive sensors.

Magnetostrictive Materials

The amount of magnetostriction in base elements and simple alloys is small, on the order of 10^{-6} m/m. The magnetostrictive effect is exhibited by the transitional metals including iron, cobalt, and nickel and their alloys. The magnetostrictive coefficients of some magnetostrictive materials are shown in Table 8.13. The highest room temperature magnetostriction of a pure element is that of Co, which saturates at 60 microstrains.

"Giant magnetostriction" materials are materials in which the magnetostrictive coefficient exceeds 1000 μm/m (that is about two hundred times larger than that of iron). These materials are today's choice for magnetostrictive sensors and actuators. They were developed in the 1960s for underwater radar; a typical material is Terfenol-D.[104] Terfenol-D is a commercially available magnetostrictive material incorporating the rare-earth elements terbium and dysprosium ($Tb_xDy_{1-x}Fe_y$, with x between 0.27 and 0.3 and y between 1.9 and 1.98) (Etrema Company). This material offers strains up to 0.002—at liquid nitrogen temperatures (−196°C), terbium, dysprosium, and compounds of the two can exhibits strains of 0.01. The Terfenol-D effect is often referred to as the *giant magnetostrictive effect*. It works in the opposite sense of iron magnetostriction; the material expands rather than contracts. An alloy of iron with samarium also displays a giant magnetostrictive effect but, like iron, in the shrinking sense (see below and Honda et al.[134]).

Sputter-deposited magnetostrictive films present an interesting opportunity for actuation in micromachines where contactless, high-frequency operation is desired. Terbium-iron alloys are typical magnetostrictive materials that can be deposited by sputtering. Quandt et al., using an RF-sputtered TbDyFe film (10 μm thick) on a silicon cantilever 2 cm long and 50 μm thick, could deflect the silicon cantilever by more than 200 μm in an external field of only 30 mT.[135] The cantilever was operated at a frequency of 500 Hz, and no degradation could be observed after more than 10^7 operations. The same group is exploring the use of these films for valves and pumps. Honda et al. discovered that amorphous Tb-Fe films exhibit positive magnetostriction, while amorphous Sm-Fe thin films exhibit negative magnetostriction (see above), and this group built magnetostrictive bimorph cantilevers and traveling machines as illustrated in Figure 8.43a and 8.48b.[134] The actuation behavior of the magnetostrictive bimorph cantilever is as follows: a 1-cm-long polyimide beam (thicknesses of 7.5, 50, and 125 μm were experimented with) is sandwiched between an upper Tb-Fe film and a bottom Sm-Fe film, each 1 μm thick. When a magnetic field is applied along the cantilever length direction, the Tb-Fe film expands and the Sm-Fe film contracts, bending the polyimide beam downward. With a magnetic field along the width direction, the Tb-Fe film contracts and the Sm-Fe film expands, deflecting the beam upward. One version of this group's traveling machine, shown in Figure 8.43b, consists of a 7.5-μm-thick polyimide film equipped with magnetostrictive bimorph actuator layers and with two legs inclined so that it travels in one direction. With an alternating magnetic field of 100 Oe at 50 Hz applied along the machine length direction, it vibrates and travels at an average speed of 0.5 mm/s in the arrow direction.

A major application of magnetostriction is in noncontact torque sensing. An integrated silicon micromachined sensor head for torque and force measurements was first proposed in 1994 by

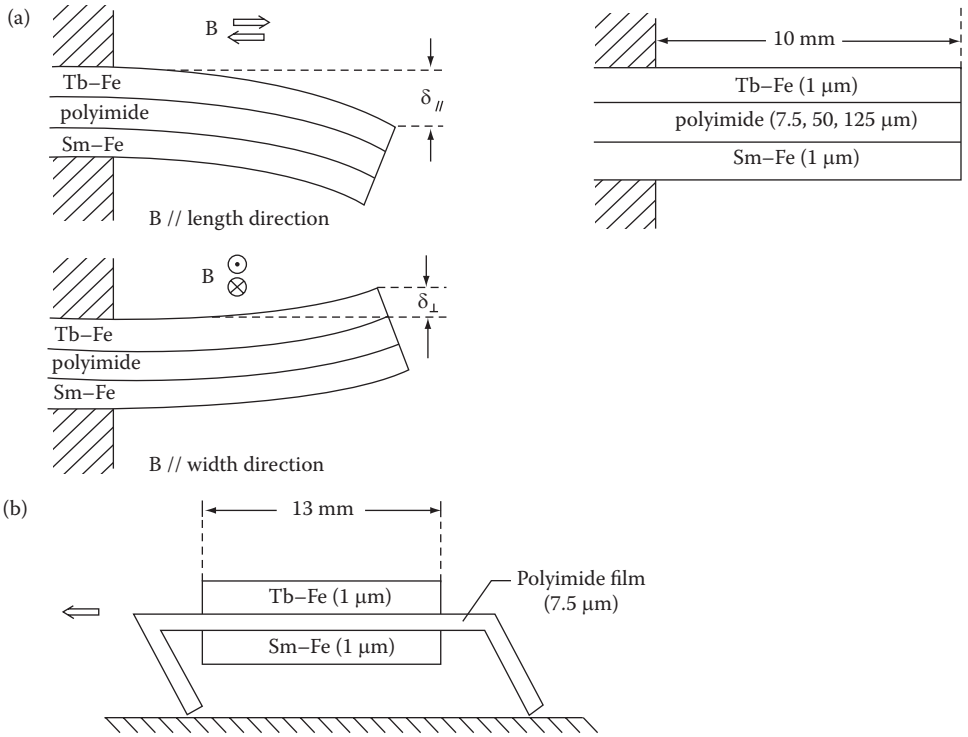


FIGURE 8.43 Magnetostrictive actuators. (a) Actuation behavior of a magnetostrictive bimorph cantilever. (b) Schematic view of traveling machine. [After Honda, T., K.I. Arai, and M. Yamaguchi. 1994. Fabrication of actuators using magnetostrictive thin films. In *IEEE International Workshop on Micro Electro Mechanical Systems (MEMS '94),* 51–56. Oiso, Japan.[134]]

Rombach and Langheinrich.[136] It is based on a magnetic yoke with an exciting and receiving coil that detects the change of permeability of a $Co_{75}Si_{15}B_{10}$ ribbon fixed to the shaft surface of an automotive vehicle. The $Co_{75}Si_{15}B_{10}$ is an amorphous metal alloy rather than the crystalline alloys discussed so far. The most common of these materials is Metglas 2605SC, an amorphous alloy of iron, with boron, silicon, and carbon [for more on these so-called *bulk metallic glasses* (BMGs), see Chapter 5]. These materials exhibit a smaller magnetostrictive distortion at saturation than do the crystalline iron-terbium alloys, and generate smaller forces, but they provide a larger response for small applied fields and vice versa—small mechanical distortions create a larger change in magnetic properties. These amorphous magnetostrictors are used in sensitive strain gauges and in accelerometers as well.[105,137]

Comparing Magnetostricion with Piezoelectrics

The power requirements for giant magnetostriction are greater than for piezoelectric materials, but the actuation offers a larger displacement, and the ratio of mass per unit stress is greater than with a PZT actuator.[138] Magnetostrictive materials have other advantages; they are metal alloys rather than brittle ceramics, and they do not exhibit dielectric breakdown under high fields as do piezoelectric and ferroelectric materials.[105] There are obviously relative advantages and disadvantages to the two technologies; some of these are listed in Table 8.14.

Magnetohydrodynamic Pumping

Magnetohydrodynamic (MHD) pumping is widely used in nuclear power plant cooling systems and has been applied for pumping and sensing of conductive gases and liquid metals.[139] It has not been exploited much in the microdomain.

MHD pumping, like EHD pumping, involves no moving parts, has a continuous flow, and is compatible with solutions containing biological specimens—both phenomena are attractive as scales go down. A MEMS MHD pump is schematically illustrated in Figure 8.44. The effect is based on the Lorentz force, F_L (see Volume I,

Table 8.14 Comparison of Magnetostrictive and Piezoelectric Materials

Parameter	Magnetostrictive	Piezoelectric
Frequency range (ultrasonic range is from 20 to 200 kHz)	<20 kHz More limited because of size	20–200 kHz Most versatile
Audible noise	More noise	Less noisy; higher-frequency piezoelectric equipment is easier brought under Occupational Safety and Health Administration (OSHA) limits by appropriate design and acoustic shielding
Power requirements	Larger	Lower
Displacement	Larger	Lower
Transducer reliability	Both are mature, highly reliable, engineered devices	
Generator reliability	2-year warranty	1-year warranty
Effect of aging; the electrical activity of piezoelectric materials slowly degrades as a half-life function from the time they are "poled" material during manufacture	No depolarization problem	In the case of piezoelectric materials, the reduction in activity after 100 days brings the activity level to within 1% of the ultimate level it will achieve over the next 17 years of aging
Energy efficiency	<50%: magnetic systems are usually less than 50% efficient because of the energy lost in heating of the coils and the effects of magnetic hysteresis	>70%: piezoelectric transducers are extremely efficient because of the direct conversion of electrical to mechanical energy in a single step
Material strengths	Magnetostrictors are metal alloys rather than brittle ceramics, and they do not exhibit dielectric breakdown under high fields as do piezoelectric and ferroelectric materials	Brittle ceramics that do exhibit dielectric breakdown

Equation 3.64), which is the force generated by a current-carrying conductor in a magnetic field. This magnetic force, in the case of a fluid, can be written out as follows:

$$F_L = \sigma \mathbf{EB} \tag{8.90}$$

where σ = conductivity of the fluid ($\Omega^{-1}m^{-1}$)
B = permanent magnetic field (Gauss)
E = electric field applied to the pump (V/m), which is E = V/L

Heng et al., using LIGA, microfabricated an MHD-type micropump. When operating in the DC mode, they observed significant bubble generation due to electrolysis of the fluid being pumped. The same group later worked on an AC-driven nozzle/diffuser type MHD to eliminate or reduce bubble formation.[140] In an AC MHD micropump, a sinusoidal AC electrical current and a perpendicular sinusoidal AC magnetic field pass through the electrolyte solution and transverse to a microchannel. The time-averaged

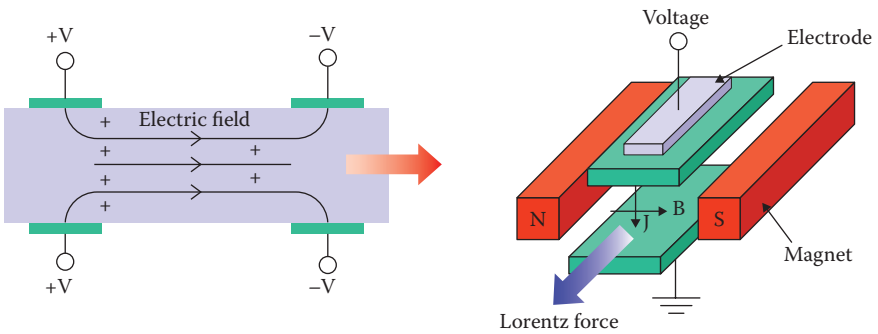


FIGURE 8.44 Schematic of a microelectromechanical system magnetohydrodynamic pump.

Lorentz force in this arrangement is given by the following equation:

$$F_L = IBw\int_0^{2\pi} \sin\omega t \sin(\omega t + \phi)d\omega t \qquad (8.91)$$

where I is the current amplitude and w the width of the microchannel. The ability to control the phase allows for controlling both the flow speed and the flow direction. Such an AC MHD micropump was demonstrated by Lemoff et al.[141] The same group also microfabricated an AC MHD fluidic switch in silicon with a switching speed faster than 0.033 s.[142] The switch is made by integrating two AC MHD pumps into different arms of a fluidic Y-channel. With a 1 M NaCl solution, a flow velocity of a 0.5 to 0.6 mm/s was registered in a 300-μm-deep, 1-mm-wide channel.

Magnetorheological Fluids, Ferrofluids, and Ferrowaxes

Magnetorheological Fluids

Magnetorheological (MR) fluids are colloidal dispersions of mesoscale size (1–10 μm) ferromagnetic particles suspended in polar or nonpolar carrier fluids. The particles are typically high-purity iron powders, but ceramic and alloys are used as well. These particles are multidomain and exhibit low levels of magnetic coercivity. As expected, the clumping of the suspended particles is caused by magnetic attraction.

When unperturbed by a magnetic field, MR fluids appear rather liquid. Upon applying a magnetic field, the viscosity of the fluid increases with the magnetic field magnitude in a nonlinear fashion. Upon reaching very high fields, the MR fluid becomes solid-like (Figure 8.45). As in the case of ER fluids, MR fluids

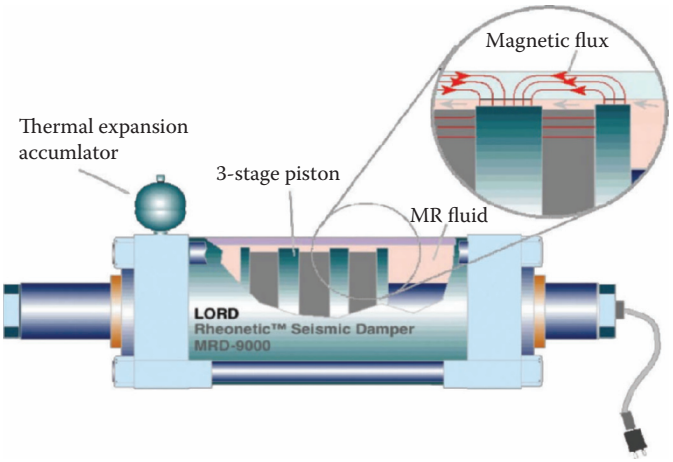


FIGURE 8.46 Magnetorheological fluid damper. Stroke: ±8 cm. Maximum input power: <50 watts. Maximum force (nominal): 200,000 N.

reverse to their original state upon removal of the external field. Overall response time of an MR system is of the order of milliseconds.[143]

MR fluids were discovered and patented in 1948 by Jacob Rabinow,[103] who was then with the National Bureau of Standards (NBS) [now the National Institute of Standards and Technology (NIST)]. After this invention, MR fluids basically disappeared from both the academic and industrial scene until the late 1980s. It is believed that the reason for this decline of interest in MR fluids in the early 1950s was that sedimentation and agglomeration of the magnetic particles were deemed unsolvable. This explains why knowledge of MR fluids is not as mature as that of ER fluids (see above), where the research remained more or less active after their initial discovery. Surprisingly, commercial success came first to MR fluids.

The main applications of MR fluids have been in the development of active vibration control and in the transfer of torque or force systems. In Figure 8.46, an MR fluid damper used in active building

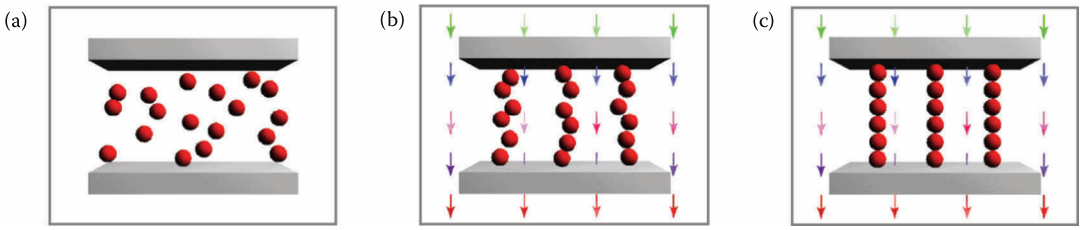


FIGURE 8.45 Magnetorheological fluids: (a) No magnetic field. (b) Intermediate magnetic field. (c) High magnetic field.

control (earthquakes) is shown. Another excellent application of MR fluids is the MagneRide by Delphi Corporation. A well-marketed magnetorheological suspension system, it was first used by GM in the Chevrolet Corvette C5, Cadillac Seville, and STS and is now also used in the Ferrari 599, the Audi TT, and the Audi R8. For more information, see http://delphi.com/manufacturers/auto/other/ride/magneride.

A major goal for the future in MR fluids is the more complete characterization of their performance depending on temperature, time, and shearing stress cycles. A hurdle in the path toward commercialization is the lack of suppliers of high-purity iron powder that makes the cost of MR fluid relatively high, even in large volumes.[143]

Ferrofluids

There are some notable efforts in nanoparticle-based MR fluids. Ferrofluids (FF) are a variant of MR fluids and are differentiated only by the size of the suspended particles used. In FF, one deals with particles smaller than 20 nm in diameter, whereas MR fluids are designed with particles in the micrometer size range (1–10 μm). The magnetic behavior of nanostructured materials has been investigated since the early days of nanotechnology. Some very significant results were first obtained by Herzer and Warlimont in 1992,[143] and an excellent review of the field by Herzer appeared in 1997.[144] When a ferromagnetic material is magnetized in one direction, it will not relax back to zero magnetization when the imposed magnetizing field is removed. It must be driven back to zero by a field in the opposite direction. If an alternating magnetic field is applied to such material, its magnetization will thus trace out a loop called a *hysteresis loop*. It is well known that the microstructure, especially grain size, determines the hysteresis loop of a ferromagnetic material.[144] It is found that the coercivity, H_c, the intensity of the magnetic field needed to reduce the magnetization of a ferromagnetic material to zero after it has reached saturation, increases with decreasing grain size down to values of about 40 nm for a pure metal such as iron or about 700 nm for hard magnets based on rare earth metals.[144] The reason for the H_c increasing proportional to $1/d$ is that as particles decrease in size, the magnetic domain size coincides more and more with the grain size. In this size range, multidomain (M-D) materials change to single-domain (S-D) materials. In the single-domain region, a change of magnetization can be produced only by rotation of the magnetization of a single grain or crystallite as a whole, which requires strong magnetic fields, depending on the shape and anisotropy of the particles. As the size of the magnetic particles is further scaled down, a transformation from ferromagnetic to superparamagnetic behavior occurs and the curve of H_c versus grain size d exhibits a peak at d_s (see Figure 8.47).[144] At particle size d_s, the material behaves as a purely single-domain material, and at d_p, the transition to a superparamagnetic is complete.

Typical particle sizes for the ferrro- to superparamagnetic (SP) phase transformation are between 10 and 20 nm for oxides and 1–3 nm for metals. This range corresponds to the band of lines drawn through the peak in the H_c-d curve. In the nanoparticle regime, between d_s and d_p, the coercivity exhibits a d^6 dependence.[144] In the superparamagnetic state of the material, the room temperature thermal energy, kT, where k is the Boltzmann constant, overcomes the magnetostatic energy of the particle (= domain), resulting in minimal hysteresis. In this case, even when the temperature is below the Curie or Neel temperature (and hence the thermal energy is not sufficient to overcome the coupling forces between neighboring atoms), the thermal energy is sufficient to change the direction of magnetization of the

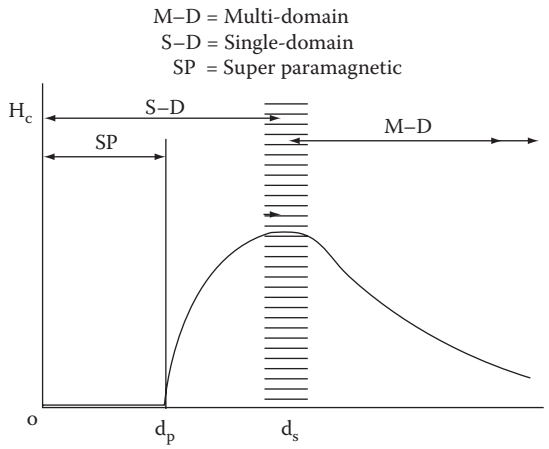


FIGURE 8.47 Coercivity (H_c) versus particle size. (Based on Herzer, G. 1997. In *Handbook of magnetic materials,* ed. K.H.J. Bushow, 417. Elsevier Science B.V.[144])

entire crystallite. The resulting fluctuations in the direction of magnetization cause the magnetic field to average to zero at d_p. Thus, the material behaves in a manner similar to paramagnetism, except that instead of each individual atom being independently influenced by an external magnetic field, the magnetic moment of the entire crystallite tends to align with the magnetic field. Superparamagnetic particles are attracted to a magnetic field but retain no residual magnetism after the field is removed (no hysteresis). These superparamagnetic nanoparticles feature a high-saturation magnetization, high relative permeability, low coercivity, high resistivity, and low hysteresis loss, and all of this benefits a large number of applications,* including in FF.

Superparamagnetic particles can be dispersed in a liquid, e.g., oil, without perceptible agglomeration if treated properly (see below and colloidal stability in Volume I, Chapter 7). Because of the nanometer sized particles in FF, sedimentation is prevented by Brownian motion, and agglomeration is avoided through the use of surfactants or ionic shells that coat the particles. If no coating is used, attractive van der Waals forces would cause the agglomeration of the particles. A typical FF constitutes a stable colloidal suspension of single-domain magnetic nanoparticles dispersed in a liquid carrier in a volume fraction of typically 2–10% (Figure 8.48). As explained above, MR fluids exhibit different viscosities depending on the strength of the magnetic field applied; FF, on the other hand, maintain their viscosity regardless of the magnitude of the magnetic field. Where MR fluids solidify in strong magnetic fields, FF always remain liquid.

In their natural state, with no field applied, FF are subjected to three forces: 1) repulsive forces due to the implemented coating, 2) attractive van der Waals forces, and 3) magnetic dipole-dipole interactions, which can be either repulsive or attractive. A true FF maintains equilibrium between the forces mentioned even in the presence of strong magnetic fields. A stable FF will not conglomerate for several years, until eventually the surfactants or the coatings on the particles break down and gradually cause the irreversible conglomeration of the particles into bigger clusters. At this point, the behavior of the FF tends to simulate that of an MR fluid.

FIGURE 8.48 Ferrofluid fixed with a magnetic field. The bottle is filled with ferrofluid and shows the magnetic force of the ferrofluid acting on the permanent magnet (Institute for Machine Tools and Factory Management, IWF, Germany).

Dispersed nanoparticles in FF are typically magnetite (Fe_3O_4) or maghemite (γ-Fe_2O_3), although cobalt, iron, and manganese zinc ferrite Zn_xMn_{1-x} have been used as well. The fluid carrier might be polar (water, in which case the FF are highly biocompatible) or nonpolar (hydrocarbons, fluorocarbons). Obtaining magnetic nanoparticles itself constitutes an important research field, and different methods have been attempted, with precipitation being the most common one (see Volume I, Chapter 7 and Chapter 3, this volume). Current applications of FF include positioning systems for multiaxle movement (see Figure 8.49), compliant positioning systems, dynamic sealing (e.g., on computer hard disks), bearings, adaptive gripping devices and dampers, and heat dissipation. Suspended superparamagnetic particles are also used to tag biomaterials of interest, the targeted species can be removed from a matrix using a magnetic field, but they do not agglomerate (i.e., they stay suspended).

Although perhaps not immediately or obviously useful, FF present a striking behavior when subjected to a normal magnetic field of strength above a critical value. Such fields induce a normal-field instability, causing the FF to instantly protrude from the surface in spikes aligned to the field lines, as

* On the other hand, it is also because of superparamagnetism that computer hard drive technologies are expected to stop growing once they reach a density of 150 gigabits per square inch.

Technical Data:

Load capacity: F < 10 N (in normal direction)
Accuracy of Positioning: ~ 1 μm
Max. velocity: 0.1 m/sec (uncontrolled); 0.02 m/sec (precise control)
Max. Range: s = 10 mm (limited by the volume of fluid)
Coil current: I < 5 A (for optimized coils and field formers)

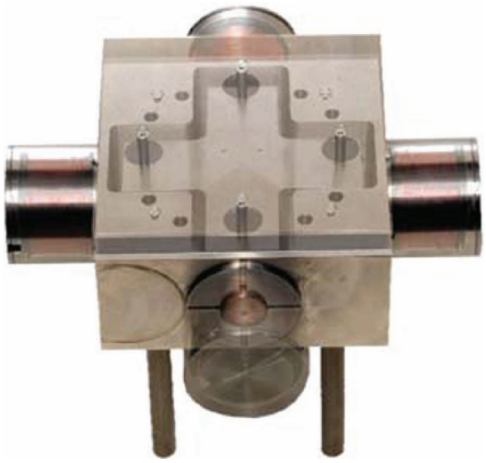


FIGURE 8.49 View of a two-axis ferrofluid positioning system with two paired coils for each direction (Institute for Machine Tools and Factory Management, IWF, Germany).

shown in Figure 8.50. The spikes are characteristic of equilibrium of gravitational force, magnetic force, and surface tension.

Many excellent videos have been created demonstrating this interesting behavior. A special mention goes to those obtained by Sachiko Kodama (http://www.kodama.hc.uec.ac.jp/protrudeflow/index.html), where art and technology are cleverly mixed to obtain beautiful FF patterns and behaviors.

Another interesting application along the same artistic lines is the creation of fractal droplets on a surface by confining a drop of FF between two glass slides in a Hele-Shaw cell* with a 0.9–1.4 mm gap (see Rinaldi and Zahn[145] and http://www.aps.org/units/dfd). When an FF drop is placed in such a cell and is subjected to simultaneous in-plane rotating and DC axial uniform magnetic fields and the magnitude of the fields increases, different, striking patterns are obtained (see Figure 8.51). The fluorocarbon-based FF used for the results shown here has a saturation magnetization of ≅400 Gauss and a low field magnetic susceptibility of $\chi \approx 3$. The FF is surrounded by a 50/50 mixture of isopropyl alcohol and deionized water that prevents the FF from wetting the glass plates. The rotational field strength is up to 100 Gauss rms at frequencies 20–40 Hz while the dc axial field is varied from 0 to ~250 Gauss.[146]

FIGURE 8.50. A normal-field instability caused by a magnet placed underneath the ferrofluid. (From Sachiko Kodama, http://www.kodama.hc.uec.ac.jp/protrudeflow/index.html.)

FIGURE 8.51 Fractals on a surface obtained by magnetic interaction of a field with a ferrofluid drop enclosed in a Hele-Shaw cell.[146] (From Prof. Markus Zahn, Massachusetts Institute of Technology.)

Next we detail an important microfluidic application of FF (i.e., valving and pumping) and highlight the prospect of ferrowaxes in fluidic platforms.

Hartshorne et al. demonstrated a FF-based valve.[147] In their work, a 600-cP FF was used to implement plugs to block fluidic channels. In air, this FF valve could hold pressures of up to 12 kPa (1.74 psi). When tested in water, water leakage at the FF/channel wall interface was observed. By making the channel walls hydrophobic, this water leak could be reduced. In the latter case, the valve could withstand pressures up to 10 kPa (1.45 psi). Besides their low pressure rating, FF plugs require the application of a magnetic

* A Hele-Shaw cell consists of two flat plates that are parallel to each other and are separated by a small distance, with at least one of the two plates being transparent.

field at all times in order to maintain the plug at the desired position.

An elegant approach for pumping with FF was introduced by Hatch and coworkers.[148] This team manipulated two FF plugs in a circular channel using a magnet that moves in a circular motion and maintains the plugs in place using stationary magnets. Using this scheme, one is able to pump defined volumes of liquid, given by the separation between the ferroplugs and the radius of the circular channel containing both pumping liquid and plugs.

Our team at University of California, Irvine, is attempting to use modified FF valves to create an inexpensive, normally closed, reusable valve that does not consume any power to stay closed. This type of valve is essential for all types of microfluidic platforms. For example, diagnostic reagents often must be stored on a chip or platform for months before they actually are used (the shelf-life of a diagnostics lab-on-a-chip device must be at least six months). Unless one has good, inexpensive liquid and vapor barriers/valves, liquids or vapors from the different reagent compartments will contaminate the system and even mix with each other, rendering the platform useless in a matter of hours. In principle, a regular pure wax valve fits the description: inexpensive, normally closed, and a good barrier for both liquids and vapors. Pal and colleagues[149] implemented pure wax valves, using paraffin wax and other waxes with higher melting points. Moving the resistance-heated molten wax by applying air pressure toggled the valve. Similarly, Liu and coworkers[150] used paraffin wax in their single-use valve. In this case, resistive heaters or Peltier elements melted the wax, while air pressure was again used for moving the molten wax. The authors claimed zero leakage when checking sealing against air. They measured a maximum holding pressure of 40 psi, while Pal et al. achieved 250 psi. Both groups succeeded in using a wax valve for sealing a polymerase chain reaction (PCR) chamber (remember that the maximum temperature needed for PCR thermocycling is ~92°C; see Figure 2.53). Liu et al. reported the use of paraffin valves (melting point ~50°C) 10 mm away from the PCR chamber, while Pal et al., using waxes with higher melting points (~85°C), could operate as close as ~5 mm from the PCR chamber. Use of microcrystalline waxes, with melting points up to 95°C, would be better yet for use in PCR applications.

An important drawback of the aforementioned wax valves is the need for an air pump to control the movement of the molten wax. This is where modified FF valves come in. A normally closed, reusable FF valve that provides effective sealing of a channel, for both vapors and liquids, and can be moved magnetically would thus be a very desirable feature. In the case of a regular FF valve, though, that objective is not met: to stay closed, a magnetic field needs to be applied, which is impractical during storage. Ferrowaxes, a modified type of FF valve, might be able to solve this problem.

In a ferrowax, magnetic nanoparticles are dispersed in a molten wax, and phase changes between solid and liquid are temperature controlled. A magnetic wax offers the opportunity of magnetic manipulation at will while the wax is molten and solidification at desired location (e.g., a valve seat) upon removal of the heat source. Additionally the wax, in its liquid phase, conforms to its holding container, and through the use of different kinds of waxes, the melting point can be tailored to different applications. Paraffin waxes appear to be the best choice for melting points of around 70–80°C, whereas microcrystalline waxes can take temperatures of around 95°C. Thanks to its nonpolar nature, the wax is immiscible in most biological samples. Advantages and disadvantages of the use of ferrowax are summarized in Table 8.15.

Kwang W. Oh and colleagues[151] demonstrated a ferrowax valve in a Y-shaped microfluidic junction. The valve is opened and closed by heating (using resistance heating) and magnetically displaced at will. The valve presented no leakage up to pressures of 50 psi.

Park and coworkers[152] implemented a single-use ferrowax valve in a compact disc (CD)-like centrifugal microfluidic device. This research group at Samsung took advantage of the absorption of laser energy by the embedded magnetic nanoparticles. The laser source operates at relatively low intensity to induce heat in the ferrowax causing a quick phase change to liquid. The pressure generated by the phase change causes the plug to move in a direction established a priori by the device's geometry, thus either

Table 8.15 Pros and Cons of Using Ferrowaxes

Pros	Cons
Relatively inexpensive	Heat source is required adding to the cost and complexity of the system
Energy efficient	Finite time of ferrowax phase change limits the time available for manipulation
Absorption of specific electromagnetic energy	Not yet commercially available (as of 2008)
Precision control	Industry scalability, since they require deposition of hot, viscous slurry
Nonpolar carrier used	
Melting point can be tailored	Melting point restricts the temperatures the at which the device might be used and/or stored
Do not interact with sample	
Conform to container geometry and provide an excellent seal after solidifying	

blocking or unblocking a channel or chamber. In this case, the magnetic nanoparticles are used solely to enhance heat absorption of the ferrowax and to induce a quicker phase change than that of paraffin wax alone. Our team at University of California, Irvine, tested the quality of the sealing of a ferrovalve for vapors and liquids qualitatively by immersing ferrowax sealed channels, containing calcium chloride, in a liquid environment. Even after more than twenty months, the calcium chloride did not show any sign of water uptake. A solidified ferrowax plug of around 2 mm in length with a cross-section of 1.75 mm^2 was also tested in a circular channel on the CD platform; the plug was located at a distance of 80 mm from the center of the rotating disc and did not show leakage up to 6000 rpm.

Regarding a real pumping application with ferrowax on the CD, the author's research group is currently exploring the use of these pumps to bring liquid samples back toward the center of the CD.[153] Indeed, a serious drawback in current CD microfluidics, besides the need of the type of valve described above, is the inability, after spinning fluids from the center to the rim, to move them back toward the center of the CD. This prevents the implementation of more complex diagnostic systems on a single standard-size CD, since the whole process of sample preparation and detection cannot be fitted easily onto a single CD platform. An FF plug could pump the sample back to the center of the CD for additional process steps (see also Volume I, Chapter 6, on fluidics, for more details on this CD platform).

The concept that we are implementing is the use of a ferrowax or FF plug as a pulling or pushing plug for the magnetic field to move the liquid inside a microchannel. To minimize water leakage at the interface of the pull/push plug and the CD plastic the walls of the channel are made hydrophobic. The concept is sketched in Figure 8.52. Theoretically, engineering a very soft ferrowax could eliminate the need for a heat source in a ferrowax system. In this case, the torque generated by the magnetic particles under a magnetic field could perhaps be made greater than the cohesion forces within the material, making the material move when a magnetic field is applied. The soft material might be a soft paraffin, gel-like polymer, or gelatinized nonpolar carrier that exhibits a "solid-like" behavior.

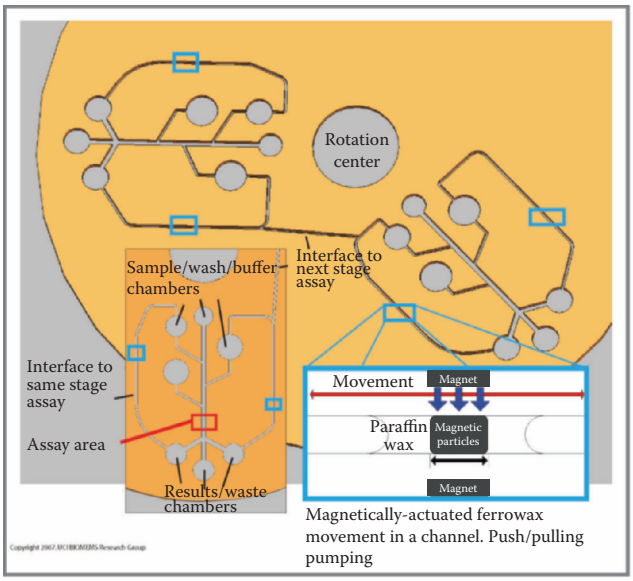


FIGURE 8.52 Schematic showing an example of a ferrowax pump. Ferrowax plugs, magnetically actuated by an external magnetic field, pull/push a sample toward the center of a compact disc (CD)-like centrifugal microfluidics platform. This way a sample might be reprocessed in one single CD platform, enabling a full diagnostic test (i.e., separation, filtration, amplification, and quantification) to be contained in a single platform. (From Zoval, J., and M.J. Madou. 2004. Centrifuge-based fluidic platforms. *Proc IEEE* 92(1):140–53.[153])

Superconductivity

Introduction

Given that actuators and power sources, in general, exhibit poor scaling in the microdomain, superconductors could play a dominant role in the more efficient powering and actuating of micromechanisms. The superconducting state is a new phase state in which, below a certain critical temperature (T_c), zero resistance and perfect diamagnetism is reached. It is well known that superconductors can levitate magnets because of their diamagnetism (Meissner effect). In this section we elaborate on superconductivity as it makes for the ideal Meissner levitators.

The levitating permanent magnets described above are ferromagnets with a positive and very large magnetic susceptibility, χ_m (χ_m usually is only constant in a small range of magnetization). Levitation can also be based on diamagnetic materials for which χ_m is small and negative. In diamagnetism, as a magnet approaches a diamagnetic material, magnetic dipole moments are induced in the diamagnetic material that oppose the applied field. These dipole moments lead to magnetic forces that tend to push away the magnet. Some ordinary diamagnetic materials, such as carbon and bismuth, with a χ_m of about -1×10^{-6} (compared with -0.20×10^{-6} for Au or -0.11×10^{-6} for Cu), can levitate magnets at room temperature. Although only a small force per unit mass, this kind of levitation does not require power, operates at room temperature, and can be used with a variety of materials. Kim et al., for example, levitated a permanent-magnet mover by applying the Meissner effect.[110] The ideal would of course be to have superconductivity at room temperature, but even a material superconducting at liquid nitrogen temperature (77 K) is amazing [liquid nitrogen ($1.04/L) is much cheaper than liquid helium ($3.75/L) and far easier to handle].

The Discoveries

Superconductivity represents a macroscopic quantum effect: one of the few cases where quantum effects become observable on a macroscopic scale. The electrons in a superconductor behave as a single particle. The Dutch physicist, Heike Kammerlingh-Onnes (1853–1926) (Figure 8.53), discovered the phenomenon in mercury in 1911, and in 1913, he received a Nobel Prize for his work.* The superconducting state may be defined as a new phase state in which, below a certain critical temperature (T_c), zero resistance and perfect diamagnetism is reached.

Heike Kamerlingh Onnes
(1853-1926)

FIGURE 8.53 The Dutch physicist Heike Kammerlingh-Onnes (1853–1926).

Most high-purity metals, when cooled to temperatures nearing 0 K, exhibit a gradually decreasing electrical resistivity, approaching some small yet finite value characteristic of the particular metal. For a few superconducting materials, such as mercury and indium, resistivity abruptly plunges from a finite value to one that is virtually zero at very low temperatures and remains there upon further cooling. Loss of resistance to electrical current flow occurs below a critical temperature, T_c. Below that temperature superconductors offer an energy transfer medium with virtually no power loss. In the case of pure elements such as mercury (see Figure 8.54) and indium, the T_c is between a few and 10 K. The change from the normal state to the superconducting state occurs over a temperature range of about a thousandth of a degree. One might wonder then if the resistance, R, really is zero and how one measures a zero resistance. To measure R = 0, a loop of a superconductor is made, a current is injected in it, and the magnetic field due to this current is measured as a function of time. Carrying out these experiments

* Kammerlingh-Onnes was also the first to liquefy helium at 4.2 K, using a "fancy" refrigerator, in 1908.

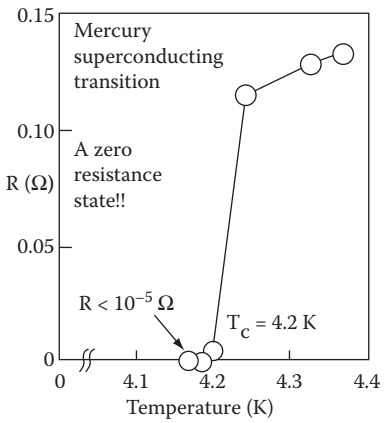


FIGURE 8.54 Mercury becomes a superconductor below 4.2 K.

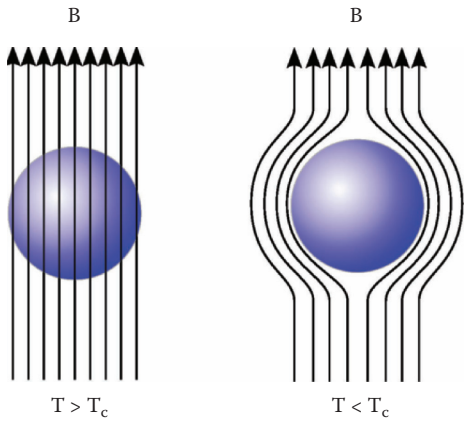


FIGURE 8.55 The Meissner effect for type I superconductors at $T > T_c$ and $T < T_c$.

turns out to be very boring, after monitoring the current via the magnetic field it generates for a number of *years* and observing no change, scientists typically shut the experiment down. We cannot say from this experiment that the resistance is really zero, but we can at least say it is far smaller than any capability or patience of ours to measure (the theory says it really is zero). In other words, once set in motion, electrical current will flow forever in a closed loop of a superconductor (R = 0), and all magnetic fields will be expelled from the same chunk of material (perfect diamagnetism). The Holy Grail for superconductivity researchers is, of course, a room-temperature superconductor, which would exhibit perpetual motion. But even a material superconducting at liquid nitrogen temperature (77 K) is amazing. [Liquid nitrogen ($1.04/L) is much cheaper than liquid helium ($3.75/L) and far easier to handle.]

The curious magnetic property of superconductors in which the material expels all traces of magnetism from its interior is called the *Meissner effect*; it was discovered in Berlin, in 1933, by Walther Meissner and Robert Ochsenfeld. Superconductivity baffled classical physicists. Below T_c there can be no resistance, so there can be no voltage inside the superconductor, and from Faraday's law it follows that the magnetic field inside cannot change. However, we actually observe that the field lines from the external field are "expelled" from the superconductor. How is that possible? The explanation is that a supercurrent (screening current) is induced on the surface in a direction such that it cancels the external field (see Figure 8.55). The magnetic field generated is strong enough to overcome gravity, so that a magnet can be made to levitate over the superconductor (see below, Figure 8.60). A superconductor such as this, which exhibits a complete Meissner effect, is called a *type I superconductor*, a soft superconductor, or a low-critical-temperature superconductor (LTS). Such a superconductor is limited in its current-carrying capability because it can tolerate only very small magnetic fields. As we shall see below, type II semiconductors can carry much larger currents but are not perfectly diamagnetic.

Among pure elements, niobium shows the highest temperature for superconductivity, and there are about a dozen niobium compounds (examples include niobium carbide, niobium titanium, niobium gallium, and niobium germanium) that all show super conductivity up to about 30 K (liquid hydrogen range). Nb_3Ge was discovered in 1973 and was long the record-holder for highest T_c (see Figure 8.56). Some of the niobium compounds have been drawn into wires to produce very compact magnetic fields for use in, for example, magnetic resonance imaging (MRI). In Table 8.16 we present a review of typical metal superconductors.

Until 1986, the T_c was 23 K for the best low-temperature superconductors, requiring liquid helium for maintaining the superconductivity state. In 1986, Bednorz and Muller (Figure 8.57), working at the IBM Zurich Research Laboratory, discovered higher-temperature (35 K) superconductivity in a lanthanium-barium-copper oxide. Their compound was a ceramic material and, ironically, a very poor conductor at room temperature.

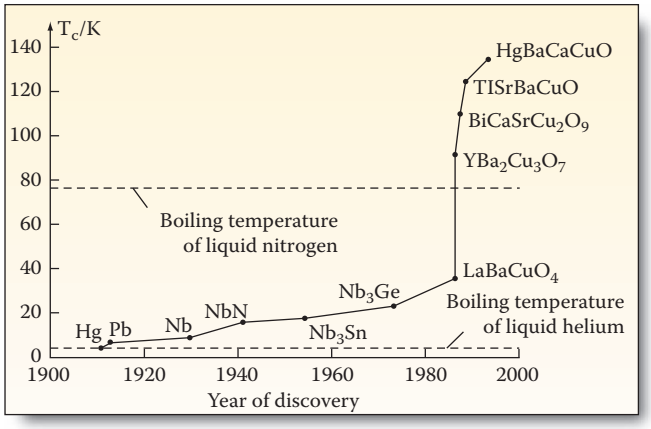


FIGURE 8.56 The critical temperature, T_c, of various superconductors plotted against their discovery dates.

FIGURE 8.57 George Bednorz and Alex Muller, 1987 Noble Prize winners.

Bednorz and Muller's ceramic was the first nonmetal that begins to superconduct above the boiling point of liquid hydrogen. Their invention triggered a race to find superconducting materials with transition temperatures above 77 K, the boiling point of liquid nitrogen. Overnight, anybody who knew about superconductors or ceramics started cooking up new compounds in his or her furnace, and a record number of high-temperature superconductors was announced over the next couple of years (see Figure 8.56). Paul Chu, of the University of Houston, made a stunning breakthrough when he announced the discovery of a material that became superconducting at 93 K, well above liquid nitrogen temperature. Early in 1987, Chu circulated a preprint of a paper describing superconductivity at 93 K in a ceramic containing Yb, Ba, Cu, and O. When the paper appeared in print, every Yb (ytterbium) had been replaced by Y (yttrium)! It was just a typo, Chu* explained. By January 1988, another copper-oxide material was shown to exhibit superconductivity at a high temperature, 110 K. The world record T_c for a superconductor is 138 K, for a compound containing mercury, thallium, barium, calcium, copper, and oxygen. Some superconductor ceramic materials are listed in Table 8.17. The new breed of ceramic superconductors is referred to as *hard superconductors*, *type II superconductors*, or *high-critical-temperature superconductors* (HTS). The two types of superconductors are compared in Figure 8.58.

In the case of a small applied magnetic field, type II superconductors act like type I superconductors. In the presence of large magnetic fields, a type II "sacrifices" part of itself so that the rest can remain superconducting. These superconductors contain vortices that are filaments of the material in the normal

TABLE 8.16 Type I Superconductors, Soft Superconductors, and Low-Critical-Temperature Superconductors

Compound	T_c (K)
Al	1.17
α-Hg	4.15
In	3.41
Nb	9.25
Ru	0.49
Sn	3.72
Ti	0.40
Zn	0.85
Al_2Y	0.35
$AuPb_2$	3.15
InPb	6.65
Ir_2Th	6.50
Nb_2SnV	9.80
CuS	1.62
Nb_3Sn	18.00
TiO	0.58
SnO	3.81
$(SN)_x$	0.26

* This would not be the last intrigue haunting superconductor research. In May 2002, outside researchers presented evidence to Bell Labs management of possible manipulation of data involving five separate papers published by its scientists in *Science, Nature,* and *Applied Physics Letters* over a two-year period by Jan Hendrik Schön as lead author and his colleagues at Murray Hill and elsewhere as coauthors. Schön is the only researcher who coauthored all five papers in question.

TABLE 8.17 Examples of Type II Superconductors with Their Critical Temperatures (T_c)

Compound	T_c (K)
$YBa_2Cu_3O_7$	93
$YBa_2Cu_4O_8$	80
$Y_2Ba_4Cu_7O_{15}$	93
$Bi_2CaSr_2Cu_2O_8$	92
$Bi_2Ca_2Sr_2Cu_3O_{10}$	110
$Tl_2CaBa_2Cu_2O_8$	119
$Tl_2Ca_2Ba_2Cu_3O_{10}$	128
$TlCaBa_2Cu_2O_7$	103
$TlCa_2Ba_2Cu_3O_8$	110
$Tl_{0.5}Pb_{0.5}Ca_2Sr_2Cu_3O_9$	120

state as shown in Figure 8.59. The magnetic field lines pass through those vortices. High-temperature superconductors thus exhibit only a partial Meissner effect. Compared with type I superconductors, type II superconductors can carry enormous currents (in some cases, superconducting wires can carry currents that are thousands of times larger than those that can be handled by ordinary wires of similar dimensions), but these brittle ceramics are very difficult to fabricate into wires.

In Figure 8.60, we show a magnetic cube, placed on a disc made of a superconductor material. As the disc is cooled down with liquid nitrogen, superconductivity sets in below its critical temperature and the magnetic repulsion levitates the cube.

The Model of John Bardeen, Leon Cooper, and Robert Schrieffer (BCS Model)

The mechanism of superconductivity remained a mystery from 1911 until 1957. The abrupt change from normal state to superconductive state indicated

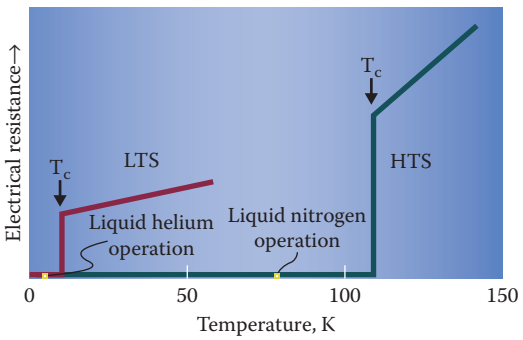


FIGURE 8.58 Low- and high-critical-temperature (T_c) superconductors compared.

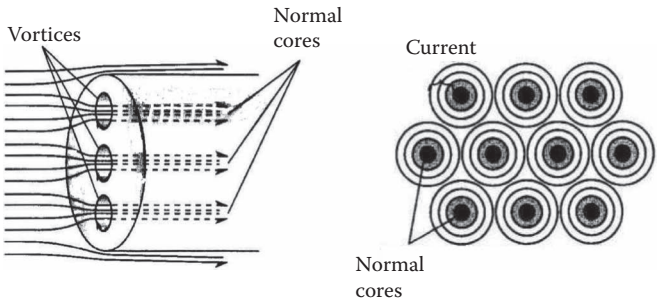


FIGURE 8.59 Type II superconductors expel magnetic fields just as type I superconductors do but contain vortices of the metal in the normal state. This is called a *partial Meissner effect*.

clearly that superconductivity is not merely an extension of ordinary conductivity. Three important observations having to do with superconductors that gave hints about its origins were the following:

1. The Meissner effect exists.
2. The "best" metals do not seem to be superconductors. Among superconducting metals, better metals make poorer superconductors, and vice versa.
3. A lighter isotope of mercury becomes superconducting at a higher temperature (i.e., more readily) than a heavier isotope.

From quantum theory, we know that electrical conductivity arises because of acceleration of electrons near the Fermi surface and that electrons move in conductor-like waves rather than as particles. Because they act as waves, electrons may travel

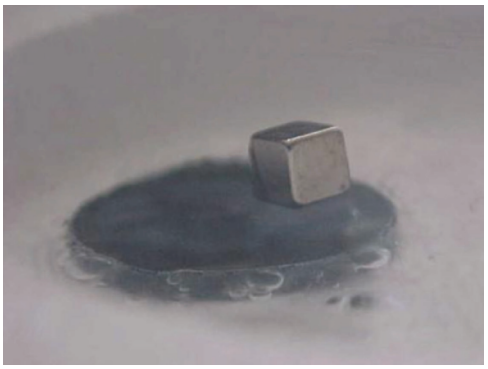


FIGURE 8.60 A magnetic cube is placed on a disc made of a superconductor material. As the disc is cooled down with liquid nitrogen, superconductivity sets in below its critical temperature, and the magnetic repulsion levitates the cube.

FIGURE 8.61 The theory of superconductivity by Bardeen, Cooper, and Schrieffer (left to right) earned them a joint Nobel Prize in 1972.

through a metal without colliding with the metal ions, and consequently, in principle at least, one could get zero resistance. If electrons do not collide with metal ions, what do they collide with? After all, metals do have finite resistivities? As we discovered in Volume I, Chapter 3, even in a pure material, thermal oscillations of the atoms, or phonons, will scatter moving electrons, causing finite resistivity; electrons do not bump into ions at the regular crystal sites but scatter from ions that happen to be vibrating and in the "wrong" place when the electron wave passes by. Impurity atoms and defects such as vacancies, dislocations, and twin and grain boundaries will scatter the electrons even more, further increasing the resistivity. When the temperature is lowered, the phonon contribution to resistivity is expected to fall away, but the contribution caused by impurities should remain finite. In the case of superconductivity, all these resistance-causing factors are somehow overcome.

John Bardeen, Leon Cooper, and Robert Schrieffer (Figure 8.61), all at the University of Illinois, first explained superconductivity in 1957 and won the 1971 Nobel Prize for their BCS model (this was Bardeen's second Nobel Prize). The model works for low-temperature superconductivity only; there is no explanation yet for the ceramic HTS. The key observation that led to the BCS model was the isotope effect, which puts one on the trail of phonons; since isotopes are involved, atoms, and perhaps their movement, must somehow play a role in the phenomenon. To start the argument, remember that in a normal metal conductivity (σ) is proportional to scattering time (τ). In superconductors, there are no final states for the scattered electrons to go to, so scattering is simply quantum mechanically not allowed. What this means is that the difference between a normal metal and a superconductor is that in the case of the superconductor there is a bandgap, as shown in Figure 8.62.

Because of this bandgap in momentum space, scattering is forbidden, i.e., there are no final states for scattered electrons to go to. The energy gap is caused by attractive interaction between electrons. Somehow pairing of electrons reduces their energy. This is rather curious; how can electrons attract each other and form pairs with a lower energy content? The answer is that there is a *phonon glue* that causes their interaction to be attractive.

This needs some further detailing. Phonons, as we know, are periodic vibrations of atoms in solids; think of plucking an atom with nanotweezers. The forces between the different atoms pull the plucked

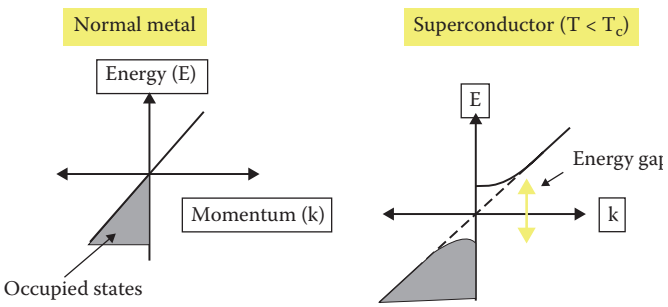


FIGURE 8.62 An energy bandgap in momentum space in a superconductor.

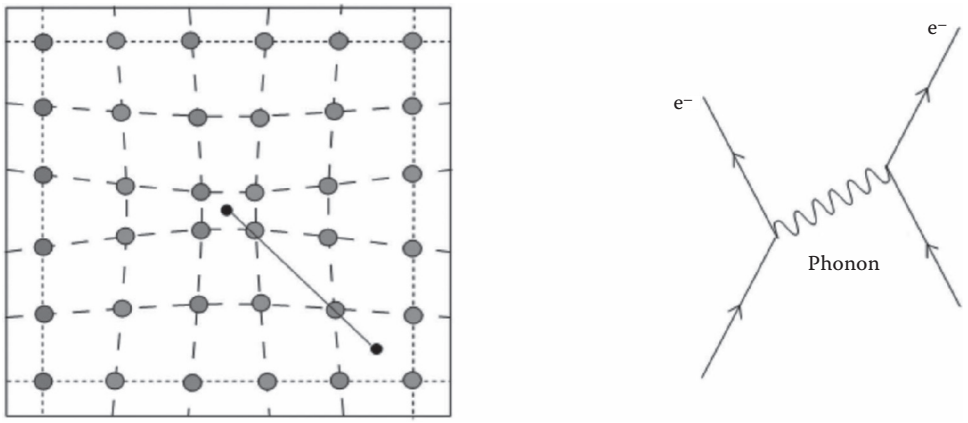


FIGURE 8.63 Pairing is mediated by crystal vibrations (phonons), and a distortion of the lattice as shown causes Cooper pair formation.

atom back into position, setting up a wave of atomic vibrations that travels through the material. Pairing is mediated by these crystal vibrations (phonons), and it is a distortion of the lattice, as shown in Figure 8.63, that allows the formation of the Cooper pairs, as shown in Figure 8.64, to take place.

A close-up of the crystal deformations leading to the formation of Cooper pairs is shown in Figure 8.65. As electrons travel between the ion cores of the metal they exert an attractive force on the positive nuclei, and these move very slightly away from their idealized positions. As a result of this disturbance, the distribution of positive charges is no longer uniform and features regions of small excess of positive charges. This extra charge is sufficient to attract a second electron, and so two electrons are linked, albeit indirectly. If the attraction exceeds the electron-electron repulsion, a "bound pair" is formed. The bound pair is called a *Cooper pair*. This mechanism explains why good conductors such as Cu and Ag are poor superconductors; these materials are good conductors at room temperature exactly because their conduction electrons do not strongly interact with the ion cores, and as a consequence we would not expect such materials to form good superconductors since the mechanism for forming pairs is very weak. This pair formation also explains the abrupt change from normal to superconducting state. Such a sudden transition is known as a *phase change*. Just as ice melts suddenly at 0°C, the thermal motion of the molecules at a critical temperature abruptly destroys the Cooper pairs.

The model also explains why a lighter isotope of mercury becomes superconducting at a higher temperature (i.e., more readily) than a heavier isotope; the smaller the mass of the positive ion core, the greater the distortion an electron can induce (to be more exact, $T_c M^{1/2}$ = constant). But how do those Cooper pairs explain superconductivity? Electrons in a superconductor move in correlated pairs in such a way that the properties of each pair are the same

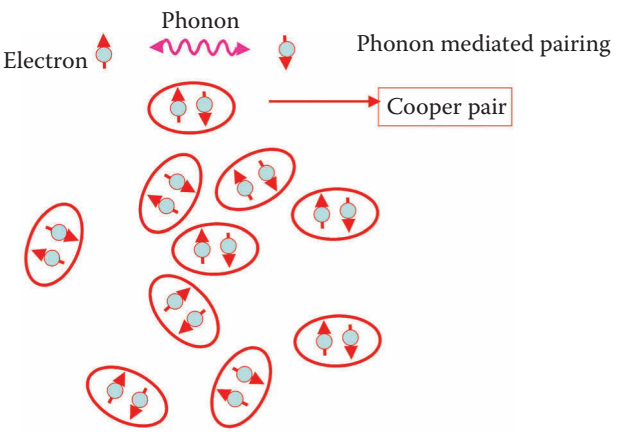


FIGURE 8.64 Cooper pair formation.

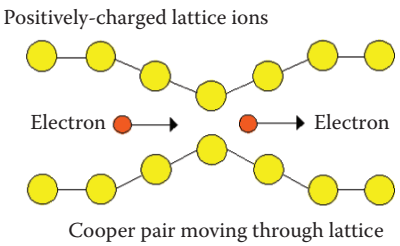


FIGURE 8.65 A lattice deformation forming a Cooper pair.

as those of all the others. These paired electrons are not necessarily close; on average they might be separated by hundreds or even thousands of atom spacings. The maximum distance up to which the states of paired electrons are correlated to produce superconductivity is called the *coherence length*. Coherence length of Cooper pairs is about 10^{-4} cm.

The electrons in Cooper pairs have opposite spins and equal and opposite linear momenta, or the pairs have zero spin and zero momentum, which makes them bosons instead of fermions (see Figure 8.66 and Volume I, Chapter 3).

In the absence of a current there is no net movement of electrons, and all the electron pairs have an average velocity of zero. Zero-spin particles such as Cooper pairs are bosons (like photons and phonons), and they "like" each other so much that they all crowd together in the same state (energy and momentum). Cooper pairs condense into one macroscopic quantum state with ~10^{23} particles that are all coherent. If a current flows, all the pairs have the same velocity and move in the same direction, as one collective body. In the superconducting state, the Cooper pairs are constantly scattering each other, but the total momentum remains constant. Remember—the paired electrons have opposite momentum. If one gains momentum, the other loses momentum. Because all Cooper pairs act together, a single lattice ion cannot scatter a single Cooper pair or they do not lose energy by interacting with the lattice. With all the electrons acting as a team, the only resistance possible is for all electron pairs to simultaneously scatter into a different state, an event that is so implausible that all attempts to measure any resistance in a superconductor have failed. This is the same as saying that there are no energy levels for the scattered pairs to go to, or that there is an energy gap. Because all the electrons act together to form a single quantum state, electrons flow around a wire as electrons in an atom do. Zeeman splitting takes place when a material is placed in a high magnetic field. When Zeeman splitting occurs in superconductors, the field is great enough to force a Cooper pair to occupy the same energy level and thus spin in the same direction. This breaks the Cooper pairs and superconductivity.

In summary, the BCS model is based on a phonon-mediated attractive interaction between electrons, leading to the formation of Cooper pairs; these pairs act like bosons, which fall into a coherent superconducting state. BCS theory beautifully explained all aspects of superconductors, but it seemed to predict that 30 K (give or take a couple of degrees) is the highest possible temperature at which a material could be superconducting, as above that temperature lattice vibrations destroy the Cooper pairing: the BCS model explains the LTS but not the HTS. BCS theory is clearly a good starting point, and just as clearly it is not the final answer. If we understood HTS, it might be possible to propose a room-temperature superconductor.

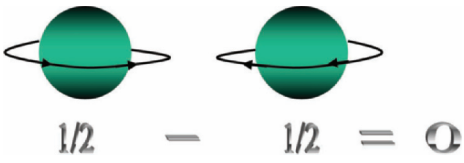


FIGURE 8.66 How to make bosons out of electrons.

Thermal Actuators[68]

Introduction

In thermomechanical actuators (TMAs), thermal energy is transformed into mechanical movement. A μ-TMA generates controlled motions via thermal strains. The simplicity of the system is the major reason that many MEMS devices are actuated by thermal means.

In Figure 8.67, we illustrate an example of a TMA, i.e., a parallel beam μ-TMA. By applying a current, the narrow beam shown in this structure is heated

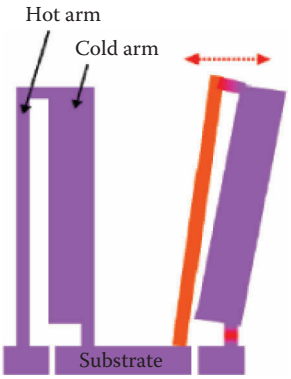


FIGURE 8.67 Parallel beam thermomechanical actuator.

up more and expands also more because of its higher resistance and higher current density than the wider beam, causing a movement as indicated. Removing the current restores the structure to its original position.

Scaling of Thermal Actuators

Scaling laws in heating and cooling, such as the total amount of heat required to start a thermal actuator, its heating and cooling rate, the thermal stresses or distortions expected, and possible damage to delicate MEMS and NEMS components, were detailed in Volume I, Chapter 6. In the same chapter we also analyzed the breakdown of the heat transfer continuum theory in the microdomain, and in Volume I, Chapter 3, we discussed quantized thermal transport in nanoscale devices.

Here we reiterate the most salient points of those scaling insights as they apply to thermal actuators:

1. Small devices, having a smaller mass, heat and cool faster than larger structures (see Volume I, Equation 6.68).
2. The Biot number (Bi) relates the heat transfer resistance inside and at the surface of a body (Volume I, Equation 6.70). This dimensionless number gives an indication of the potential for thermal stresses. For Bi $\ll$ 1, internal temperature gradients become small and the microstructure can be treated as having a uniform temperature. In the microscale, the Bi is small, so that conduction accounts for 99% of the energy transport. Consequently, with small devices there is less worry about thermal stresses induced by thermal gradients, and systems with larger internal heat generation capacity by volume can be built.
3. The increased phonon-boundary scattering in nanostructures lowers the thermal conductivity κ (see Volume I, Chapter 3 and Chapter 3, this volume). This contributes, for example, to making better heat insulators and more efficient thermoelectrics.

From the above, the micro- and nanodomain seems especially appealing for thermal actuator applications.

Types of Thermomechanical Actuators

Introduction

In this section we compare pneumatic, bimetallic, and shape memory alloy thermal actuators by considering their use in microvalves. These three types of thermal actuator mechanisms employed in microvalves are illustrated in Figure 8.68. In the section that follows we detail the choice of actuator types in commercially available valves.

Thermopneumatic Actuators

The thermopneumatic valve (see Figure 8.68a) traps a liquid in a sealed cavity containing a thin film heating resistor along one side of the cavity, with a flexible diaphragm wall forming the opposite side.[154] Upon heating, the fluid in the chamber expands and evaporates, raising the pressure in the sealed cavity and causing the flexible wall to bulge outward. In a normally open valve, this bulging diaphragm closes a nearby orifice. In a closed valve, as illustrated in Figure 8.68a, actuation levers a silicon body away from an orifice. Compared with other microvalve technologies, thermopneumatic actuation exerts tremendous force through a long stroke. In a normally open valve, over 20 N through a stroke of 50 μm was demonstrated.[154] This long stroke allows a thermopneumatically actuated valve to control higher flow rates and a wider range of pressures. This type of valve was made at Redwood Microsystems, Inc., a company that is now defunct.

Bimetallic Actuators

The second thermal actuator, illustrated in Figure 8.68b, is a bimetal. A bimetal involves the combination of two materials with distinct thermal expansion coefficients. A bimetal strip will bend when heated or cooled from the initial reference temperature because of incompatible thermal expansions of the materials that are bonded together. The bimetal returns to its initial reference shape once the applied heat is removed.

In the case of a bimetal structure based valve, as illustrated in Figure 8.68b, the nickel and silicon relax and the valve is closed when cold; upon heating, the nickel expands more than the silicon, lifting

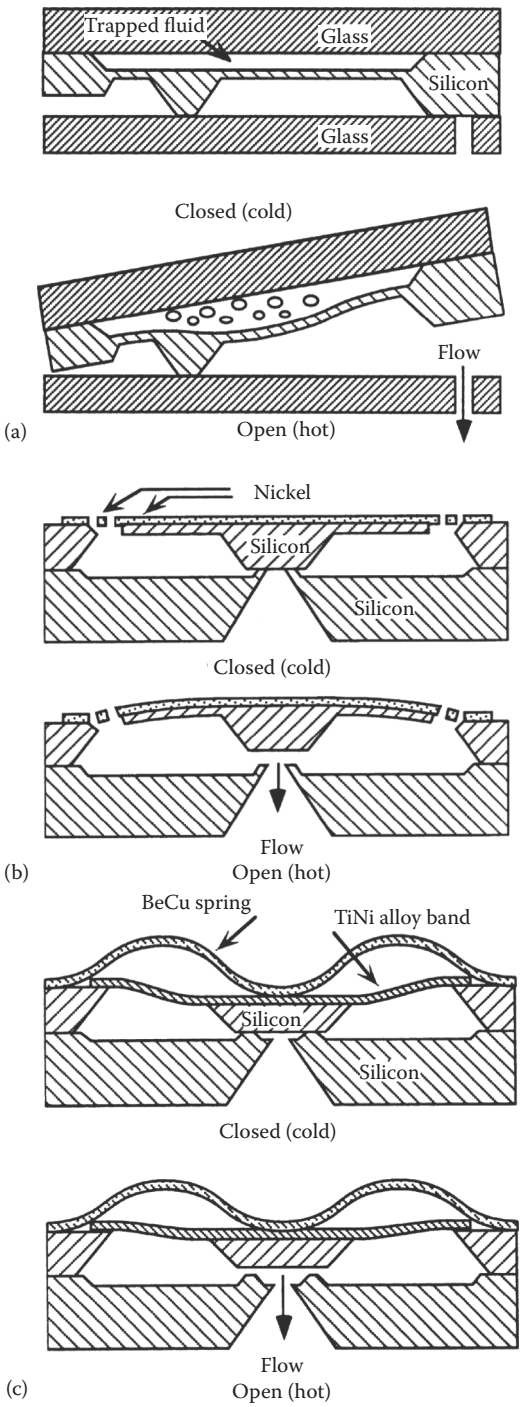


FIGURE 8.68 Schematics of thermally actuated valves. (a) Schematic of a normally closed thermopneumatic valve. (After Zdeblick, M.J., R. Anderson, J. Jankowski, B. Kline-Schoder, L. Christel, R. Miles, and W. Weber. 1994. In *Technical Digest: 1994 Solid State Sensor and Actuator Workshop,* 251. Hilton Head Island, SC.[154]) (b) Schematic of a normally closed nickel/silicon bimetal valve. (c) Schematic of a normally closed shape memory alloy valve. [After Barth, P.W. 1995. In *8th International Conference on Solid-State Sensors and Actuators (Transducers '95),* 276. Stockholm, Sweden.]

the silicon body from the valve seat. In terms of size, bimetal actuators are smaller than thermopneumatic ones and consume less power, but they lack the longer stroke of the latter.[155]

Shape Memory Alloy (SMA) Actuators

A third thermal actuator is based on shape memory alloys (see Figure 8.68c). Shape memory alloys undergo reversible thermal-mechanical transformations of their atomic structure at certain temperatures. Specifically, SMAs such as NiTi undergo crystalline phase transformations from a weak and easily deformable state at low temperatures to a hard and difficult-to-deform state at higher temperatures. The material is first held in the desired shape and heated to well above the transition temperature (T_{tr}). At these temperatures, the crystal structure is in the austenite or parent phase. Cooling transforms the material into martensite, the low-temperature phase, which plastically deforms (solder-like) by as much as 10% at relatively low stresses (yield strength is only 100 Mpa). When the SMA is heated again above its transition temperature, T_{tr}, it transforms back to the high-temperature phase (austenite) and reverses to its originally high-temperature shape, exerting a substantial force (>100 MN/m^2). The austenite phase has a much higher yield point and can sustain stresses up to 560 MPa without permanent deformation.[156] The large forces make SMAs ideal for actuation purposes; piezoelectric and electrostatic actuators exert only a fraction of the SMA force, but they act much faster. The temperature interval over which the shape change takes place typically is 10–20°C. The lowest energy path to the austenite exactly retraces the atomic movements responsible for the deformation, causing the shape memory. The discovery of the effect in 1951 involved a gold-cadmium alloy, but this was soon extended to a broad range of other alloys, including titanium-nickel, copper-aluminum-nickel, iron-nickel, and iron-platinum alloys. For practical applications, Ni-Ti, Cu-Zn-Al, and Cu-Al-Ni all have been tried. Depending on the type of alloy and the alloy composition, critical temperatures range between –150°C and +150°C.

Besides their use in μ-valves, shape memory alloys have been used in numerous applications (see Figure 8.69) including medical (e.g., microsurgical

(a) (b)

FIGURE 8.69 Examples of shape memory alloy (SMA) applications. SMA valve opens and releases hot water at the proper temperature (a). NiTi surgical tools: precise bending of the tool to the proper shape in microsurgery (b).

instruments, stents, artificial muscle), aerospace (e.g., connectors, lock rings), automotive (e.g., Ni-Ti thermostat), industrial (e.g., pipe connectors), consumer (e.g., eyeglass frames), and safety (e.g., fire safety valves).

The most extensive work with shape memory alloys involves bulk materials and bulk titanium-nickel alloys in the form of wires and rods, which are commercially available under the name Nitinol™. Nitinol, a composition transforming near room temperature, is traditionally prepared with 50% Ti and 50% Ni. Increasing the Ni content decreases the transition temperature by about 25°C per 0.2 at % Ni. The material is a good electrical conductor, with a resistivity of 80 μΩ·cm, but a relatively poor thermal conductor, with a thermal conductivity about one-tenth that of silicon.

The more IC-compatible sputtered films were made in the late 1980s by Walker et al.[157] at Bell Labs and by Busch and Johnson[156] at TiNi Alloy Company (http://www.sma-mems.com). Thin-film, nickel-titanium shape memory alloys up to 50 μm in thickness have been sputter deposited. These films exhibit memory behavior comparable to bulk material, and the phase transition to and from martensite lies entirely above ambient temperature (see Figure 8.70).[158] When deposited at room temperature, NiTi films are amorphous. Heating to 500°C causes the film to crystallize and acquire its shape memory property.[158]

Some materials and processes are incompatible with such high temperatures, requiring either rapid thermal annealing (RTA) or a special choice of materials and careful selection of the order in which processes are carried out. Another limitation of thin-film NiTi is the upper limit of 70–80°C for transition temperature T_{tr}, which is too low for applications such as valves for fuel injection systems and many military applications. Experimentation with higher-temperature tertiary alloys such as TiNiHf is under way to alleviate this problem. The bandwidth of the NiTi films is slightly improved, to about 5 Hz, compared with the NiTi wire (about 1 Hz) because of the more effective cooling in the microdomain. Another advantage of TiNi thin-film actuators is that they are biocompatible and transistor-transistor logic (TTL) voltage compatible.

We compared the energy density and drive conditions for NiTi with magnetic, electrostatic, and piezoactuators in Table 8.8. NiTi outperforms magnetic and electrostatic actuation in terms of energy density and may cause large motions in the range from 10 μm to 1 mm, operational requirements not well matched by electrostatic or piezoelectric

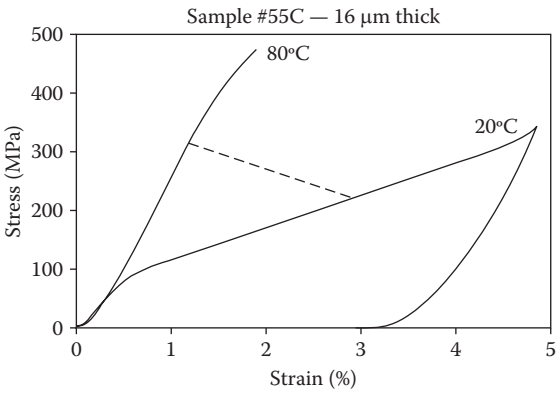


FIGURE 8.70 Stress-strain curve for a TiNi shape memory alloy thin film. The sloping, broken line indicates a load line spring force resisting the actuator, showing a 2% repeatable strain recovery. (After Johnson, A.D. 1991. Vacuum-deposited TiNi shape memory film: Characterization and applications in microdevices. *J Micromech Microeng* 1:34.[158])

technologies. As the required voltage for actuation is low (a couple of volts), it becomes compatible with IC technology. The phase change in SMAs is temperature driven, and as heat must be removed before the next cycle starts, this makes the system inefficient—the cycle rate may be slowed down by the achievable rate of heat transfer. The ultimate efficiency cannot exceed that of the Carnot cycle operating between the same temperature limits.[159] Practical SMA actuators have an efficiency typically less than 10%; for example, 50 J of heat input might be required to obtain 1 J of mechanical energy output.[158] However, in very small devices, heat transfer is rapid, which makes this type of actuator more attractive. Another potential disadvantage is the need, in most applications, of mechanical biasing devices to provide a return force such as the BeCu spring, enabling cyclical behavior in the normally closed valve sketched in Figure 8.68c. The beryllium-copper spring puts bands of SMA in tension over the orifice at low temperature. The SMA bands constitute electrical resistors and are heated by current passage. At high temperature, the SMA element works against the BeCu spring to reach its high-temperature preset form and moves away from the valve seat, opening the valve. At low temperatures, the spring easily pushes the silicon body back onto the valve seat, closing the valve. The spring adds somewhat to the size of the SMA actuator and also further slows down the response time (0.5–2 Hz). If the high-temperature shape does not change upon cooling the specimen and transforming it to the martensic phase, the phenomenon is referred to as a *one-way* shape memory effect, and springs or other mechanical biasing mechanisms are required. Solutions to avoid mechanical biasing might be within reach: SMAs can, under certain circumstances, remember their low-temperature shape as well, enabling cyclic device operation without the need of a bias.[160] In the latter case, one refers to a "reversible" shape memory effect. The low bandwidth of SMA usually is assumed to be determined by the relatively long cooling thermal time constant. Hunter et al. report an interesting unexplained effect in which very large brief current pulses (>10^9 A/m^2), imposed during externally shortening and lengthening cycles, alter the subsequent NiTi switching properties.[161] The altered NiTi shortens and lengthens very rapidly (within 40 ms) and generates a maximum extrapolated stress of 230 MN/m^2 and a peak power/mass approaching 50 kW/kg.[161]

Shape memory plastics also are being developed. Norsorex, for example, is the trade name for a polynorbornene polymer with excellent shape recovery properties at shape memory temperatures of over 35°C. Another polymer is Zeon Shable, a polyester-based polymer blend.[162] There are many potential applications for shape memory alloys and patents abound in this area (upwards of 15,000).

Summarizing, the advantages of SMA actuation include high force/weight and force/volume ratios, large deformation, simplicity of implementation via a heating current fed through the SMA material, simple cooling by the ambient environment (especially effective for small structures with a small Biot number), and inexpensive raw materials. Disadvantages include the one-way operation necessitating a bias force, heating/cooling cycles that reduce the bandwidth, the number of cycles reducing the maximal deformation (after >10^5 cycles, the maximal deformation is less than 4%), the fact that cycling changes the properties of the alloy, hysteresis (10 … 30°C), and nonlinearity.

Comparison of Actuator Principles in Commercial Valves

Miniature, highly accurate valves are of importance in many industrial operation processes. In Table 8.18, listing four commercially available thermal valve technologies, we compare the merits of the various underlying thermal actuator mechanism reviewed in this section.[155] Parameters of interest for valve performance include power consumption, actuation speed, force, stroke, cost, and package size.

Hewlett-Packard (HP) has come up with a thermally actuated valve intended for a very small, possibly handheld, gas chromatography (GC) system.[163,164] A schematic of HP's normally closed nickel/silicon bimetal valve was shown earlier in Figure 8.68c (a more detailed view is presented in Figure 8.71b). Resistive heating of the nickel/silicon bimetal causes the membrane deformation and opening of the valve. The competing IC Sensor's

TABLE 8.18 Comparison of Micromachined Thermal Valves

Parameters	Shape Memory Alloy (Microflow Analytical, Inc.)	Al (IC Sensors, Inc.)	Ni/Si (HP)	Thermopneumatic (Redwood Microsystems)
Pressure (PSIG)	80	25	150	100
Flow (sccm, air or N_2) at T = 25 °C	6000	100	1000	2000
Power (W)	0.29	0.5	1.03	2.0
Response time (ms)	100	100	200	400

Source: After Barth, P.W. 1995. In *8th International Conference on Solid-State Sensors and Actuators (Transducers '95),* 276. Stockholm, Sweden.[155]

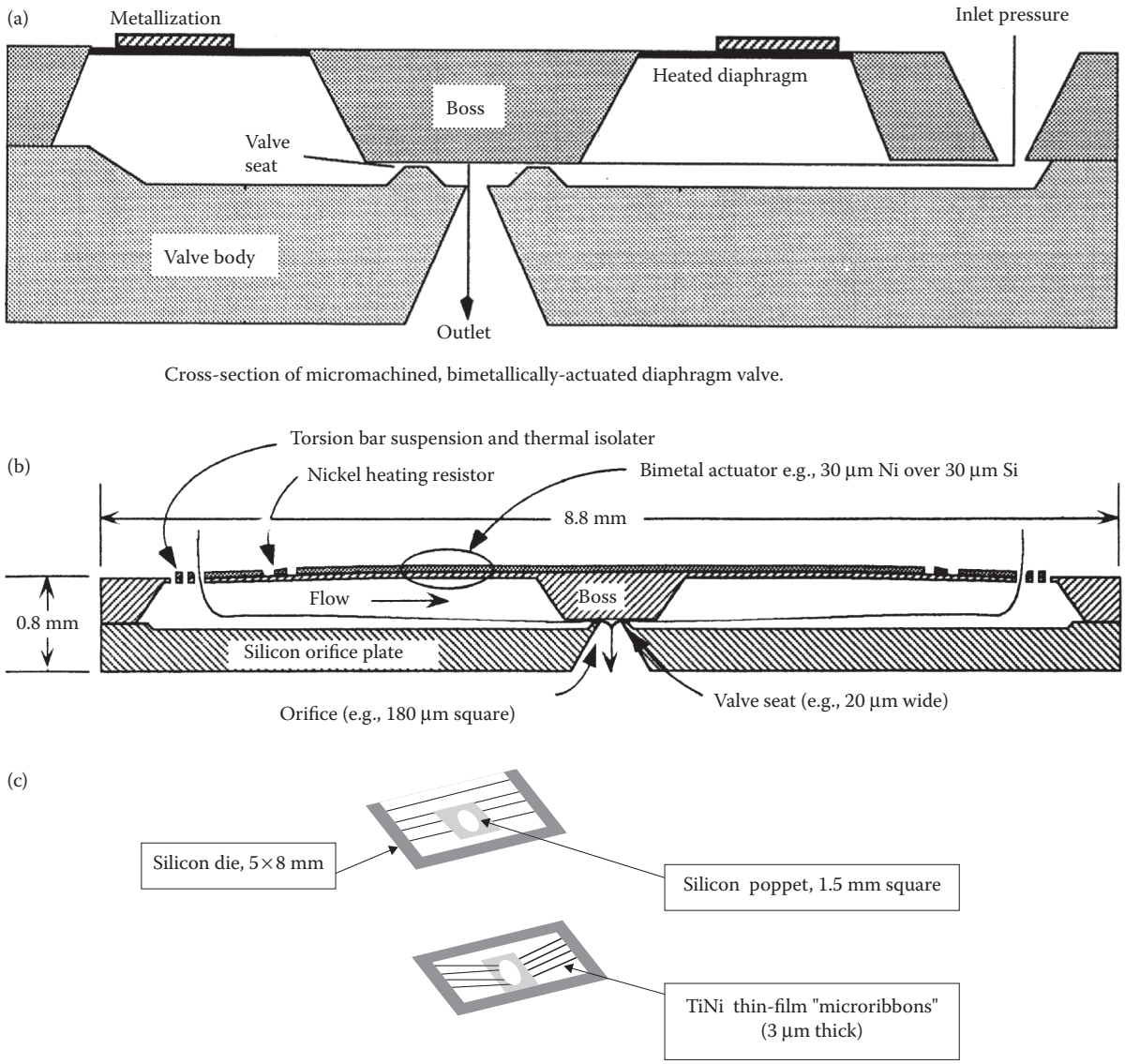


FIGURE 8.71 Commercially available, micromachined, thermally actuated valves. (a) IC Sensor's bimetallic Al/Si valve. (b) Hewlett-Packard's bimetallic Ni/Si valve. (c) The TiNi actuator consists of a rectangle of silicon, 5 mm wide and 8 mm long, that has been partially removed to leave a poppet supported by 8 TiNi thin-film "microribbons." The lower drawing shows these microribbons stretched. When electrical current runs through these microribbons, the shape memory property causes them to contract. The upper view shows the microribbons in their contracted state. The poppet displacement is more than 100 µm, and it can exert a force of as much as 0.5 N (http://www.tinialloy.com/devices.html).

valve (Figure 8.71a) is also a normally closed bimetallic valve but uses an annular aluminum region on a Si diaphragm with integral diffused resistors instead. By varying the electrical power dissipated in the resistors, and thus the temperature of the Si diaphragm, the displacement of a central boss can be controlled.[165] The HP valve exhibits an increase of an order of magnitude in both pressure and flow control capability compared with the IC Sensor's valve. This improvement is made possible by using a torsion bar suspension (see Figure 8.71b) and a nickel-and-silicon bimetal combination, which allows larger valve dimensions and higher stress levels for consequent longer stroke and higher force. For very small structures, this thermal type of actuation becomes an attractive option because of geometric scaling (see comments above and in Volume I, Chapter 6 on the Biot number). The amount of thermal mass decreases as the volume of heated material decreases, and the thermal loss decreases as the thickness of the support structure for the bimetal decreases, while the force per unit area remains high. The normally closed thermopneumatic valve (the Fluistor™) from Redwood Microsystems (now defunct) is shown in Figure 8.68a; in this valve, a liquid-filled cavity, upon heating, forces a silicon diaphragm to lift off the valve seat (a normally open valve is also available from the same company). Although providing a tremendous force (>20 N) through a long stroke (between 50 and 150 μm), this valve is thermally inefficient because of the large volume of heated fluid and is difficult to assemble because of the working fluid in the device. Moreover, it is quite slow (~ 400 ms). A fourth alternative is the normally closed SMA valve from Microflow, sold by TiNi Corporation, which is shown in Figure 8.71c.[166] Sputtered thin-film NiTi is the SMA element and is deformed in tension through an integral BeCu bias spring, pushing it against a valve seat. Passing a small current through the actuator element heats it up, and it regains its original shape, lifting it against the spring from the valve seat and allowing gas to flow. From the point of view of performance, the HP valve stands out; in terms of manufacturability and IC process compatibility, the IC valve has the edge. Redwood Microsystems has gone the furthest in incorporating its valves into higher-level subsystems, including pressure and feedback control, to create a completely electronically programmable pressure regulator. This company has also paid more attention to applying its valves in biomedical instruments carrying fluids.[154,155]

Table 8.19 Salient Points of Thermal Actuators

1. Maximum stroke: 100 μm
2. Maximum speed: 10 s to 1 kHz (low bandwidth; see Figure 8.19)
3. Maximum force: 1–10 mN (high force; see Figure 8.19)
4. Easy to fabricate
6. Footprint: 10 μm to mm
7. Special drive conditions: 1–5 V
8. High power and temperatures
9. Multiaxis operation

Summary: Thermal Actuators

In Table 8.19, we list the salient points associated with thermal actuators. See also Figure 8.14.

Other Actuators and Some Final Actuator Comparisons: Final Comparison

Introduction

In closing this chapter on actuators, we review two more interesting actuator principles, i.e., those based on chemical/electrochemical forces and those based on the quantum mechanical Casimir force, and present some final actuator comparison figures.

Chemical/Electrochemical Actuators

A variety of electrochemical and chemical actuators have been demonstrated. In Volume II, Example 4.3, we reviewed an electrochemical disruptable valve invented by the author (see Volume II, Figure 4.98). Hamberg et al.[167] used electrolysis of water to create pressure on a micromachined membrane. Osmotic pumps use a semipermeable membrane over a pump body to realize a mechanical stroke; the mechanism is clarified in Figure 8.72.[168] A solution in a closed vessel is covered on one side with a semipermeable membrane. A hypertonic solution outside the pump draws solvent from the hypotonic solution inside the pump body, reducing the volume inside and bending the membrane inward. Such an

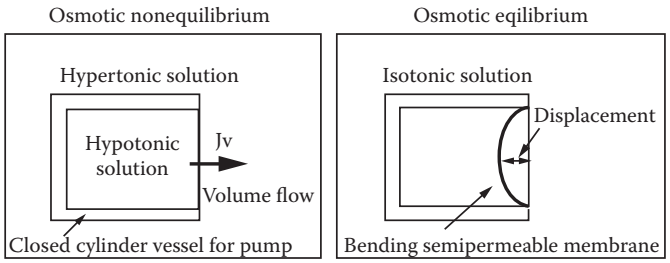


FIGURE 8.72 Mechanism of the osmotic pump. The decrease in volume bends the semipermeable membrane. JV corresponds to the volume flow out of the pump chamber. [After Nagakura, T., K. Ishihara, T. Furukawa, K. Masuda, and T. Tsuda. 1995. In *8th International Conference on Solid-State Sensors and Actuators (Transducers '95),* 287. Stockholm, Sweden.[168]]

osmotic pump was proposed for delivering insulin to a patient, depending on the patient's changing glucose concentration. In this application, the membrane displacement caused by a high glucose concentration in the patient pushes insulin out of the pump into the patient's body.[168]

Also, hydrogels and redox polymers and blends of the two have been used as chemical or electrochemical actuators. Examples include the actuator hydrogel valve in Volume II, Figure 8.15. One of the most useful features of a polymer actuator is that swelling and shrinking of the material can be triggered by a wide variety of external stimuli. This may involve pH, ionic strength, magnetic field, electrical field, etc. A number of these environmentally sensitive or "intelligent" materials are listed in Table 8.20.[169] Most of these polymers exhibit reversible structural modifications upon repeated changes in the external environment. These polymer systems are of great interest as drug delivery vehicles. A Web site describing progress in the general area of electroactive polymers is the WorldWide Electroactive Polymers Newsletter, or Artificial Muscles Newsletter (http://ndeaa.jpl.nasa.gov/nasa-nde/lommas/eap/EAP-web.htm).

In Volume I, Chapter 6, we described a gold/polypyrrole actuator μ-flap (see Volume I, Figure 6.64) as a means of active mixing in the microdomain. A thin Au film (1 μm), covered with a thin layer of polypyrrole, forms an electrochemical actuator. If submerged in an ionic solution, upon application of a bias small solvated cations in the solution move in and out of the redox polymer matrix, swelling and shrinking, respectively, its apparent volume. The volume changes in the polymer generate stresses on the Au layer surface, bending it upward or downward in a "flapping" motion. Movement frequencies in the 1–20 Hz range can be achieved with this type of actuator.

Casimir Forces: A Quantum Actuator

We normally think of a vacuum as space with nothing in it, and we would then also tend to conclude that its energy is zero. But this is inconsistent with the time-energy uncertainty principle that predicts that the energy of a vacuum fluctuates (Volume I, Equation 3.107). It is therefore possible that empty space locally

TABLE 8.20 Environmentally Sensitive Polymers for Drug Delivery

Stimulus	Hydrogel	Mechanism
pH	Acidic or basic hydrogel	Change in pH—swelling—release of drug
Ionic strength	Ionic hydrogel	Change in ionic strength—change in concentration of ions inside gel—change in swelling—release of drug
Chemical species	Hydrogel containing electron-accepting groups	Electron-donating compounds—formation of charge/transfer complex—change in swelling—release of drug
Enzyme substrate	Hydrogel containing immobilized enzymes	Substrate present—enzymatic conversion—product changes swelling of gel—release of drug
Magnetic	Magnetic particles dispersed in aliginate microspheres	Applied magnetic field—change in pores in gel—change in swelling—release of drug
Thermal	Thermoresponsive hydrogel poly(N-isopropylacrylamide)	Change in temperature—change in polymer-polymer and water-polymer interactions—change in swelling—release of drug
Electrical	Polyelectrolyte hydrogel	Applied electric field—membrane charging—electrophoresis of charged drug—change in swelling—release of drug
Ultrasound irradiation	Ethylene-vinyl alcohol hydrogel	Ultrasound irradiation—temperature increase—release of drug

Source: After Brannon-Peppas, L. 1997. Polymers in controlled drug delivery. *Med Plastics Biomater* 4:34.[169]

does not have zero energy but may have sufficient energy ΔE for a very short time, Δt, to create particles and their antiparticles. This is called the *zero point energy*. Empty space is a seething mass of quantum fluctuations creating virtual particles that recombine and annihilate, just below the threshold of reality (Volume I, Figure 7.29). These fluctuations have been measured as the Casimir effect. The Casimir effect is a small attractive force that acts between two close parallel uncharged conducting plates in a vacuum as illustrated in Volume I, Figures 3.69 and 7.30. Casimir realized that only virtual particles of certain wavelengths can appear between the plates, while those of any wavelength can appear on the outside. Particles other than the photon also contribute a small effect, but only the photon force is measurable. All bosons, such as photons, produce an attractive Casimir force, while fermions make a repulsive contribution.

The attractive Casimir force, F_c, between two plates of area A separated by a distance r can be calculated as follows:

$$F_c = -\frac{\pi^2}{240}\frac{\hbar c}{r^4}A \qquad (8.92)$$

where c is the speed of light. Steve K. Lamoreaux[170] measured this tiny force in 1996 with a torsion pendulum (Figure 8.73). One can quickly establish from Equation 8.92 that two mirrors of 1 cm^2 separated by a distance of 1 μm have an attractive Casimir force of 130 nN. Although this force appears small,

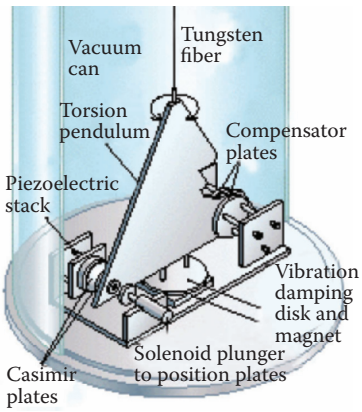


FIGURE 8.73 This experiment by Steve K. Lamoreaux[170] relies on a torsion pendulum. A current applied to the piezoelectric stack tries to move the Casimir plate on the pendulum. The compensator plates hold the pendulum still. The voltage needed to prevent any twisting serves as a measure of the Casimir effect.

at distances below a micrometer, the Casimir force becomes the dominant force between two neutral objects. At separations of 10 nm, the Casimir effect produces the equivalent of 1 atmosphere pressure! Obviously this is a force to reckon with in MEMS and NEMS devices.

As summarized in Table 8.21, Lamoreux's measurement of the Casimir force was followed by many subsequent attempts using MEMS and NEMS actuators to measure this force.

The zero point energy is also connected to Einstein's famous cosmological constant:

$$\Lambda = \frac{8\pi g}{3c^2}\rho \qquad (8.93)$$

where g = gravitational constant
c = speed of light
ρ = energy density of vacuum

Einstein included this term in the equations for general relativity because he was dissatisfied that his equations would not allow for a stationary universe, which was the model of the universe at the time. Soon after Einstein developed his theory, observations by Edwin Hubble (1929) proofed that the universe was expanding. Einstein then abandoned the cosmological constant and called it the "biggest blunder" of his life. The cosmological constant remains of interest though, as observations made in the late 1990s of distance-redshift relations suggest that the universe is accelerating. These observations can be explained by assuming a very small positive cosmological constant in Einstein's equations. The current standard model of cosmology, the Lambda-Cold Dark Matter model (CDM), thus includes the cosmological constant, which is measured to be on the order of 10^{-35} s^{-2}, or 10^{-47} GeV^4, or 10^{-29} g/cm^3, or about 10^{-120} in Planck units.

Final Comparison of Actuator Mechanisms

Finally, in Figures 8.74 and 8.75 we show two very complete figures summarizing the state of the art in actuators by Bell.[171] The first is a figure comparing all types of actuators in terms of maximum displacement versus maximum frequency; the second compares their maximum displacement versus maximum force exerted.

Table 8.21 Attempts Using MEMS and NEMS Actuators to Measure Casimir Force

Investigators	Year	Geometry	Method	Distance Scale (nm)	Materials	Pressure (mbar)	Temp. (K)	Accuracy (%)
S.K. Lamoreaux	1997		Torsion pendulum	600–6000	Au (500 nm)	10^{-4}	300	5
U. Mohideen and A. Roy	1998		AFM	100–900	Al (300 nm) + AuPd (20 nm)	5×10^{-2}	300	2
A. Roy and U. Mohideen	1999		AFM	100–900	Al (250 nm) + AuPd (8 nm)	5×10^{-2}	300	2
G.L. Klimthitskaya, A. Roy, U. Mohideen, and V.M. Mostepanenko	1999		AFM	100–900	Al (300 nm) + AuPd (20 nm)	5×10^{-2}	300	1
T. Ederth	2000		Piezo-tube manipulator	20–100	50 μm Au wires coated in thiol SAM	1000	300	1
H.B. Chan, V.A. Aksyuk, R.N. Kleiman, D.J. Bishop, and F. Capasso	2001		MEMS torsion bar capacitance	90–1000	Au (200 nm) + Cr underlayer	1000	300	1
G. Bressi, G. Carugno, R. Onofrio, and G. Ruoso	2002		Interferometry	500–3000	Cr (50 nm) on Si	10^{-5}	300	15
R.S. Decca, D. Lopez, E. Fischbach, and D. E. Krause	2003		MEMS torsion bar capacitance	200–2000	Cu/Au	10^{-4}	300	1
NANOCASE	2005		AFM, MEMS	10–1000	Si, Au	10^{-11}	20–1000	<1

Source: Antezza Mauro, University of Trento, INFN and CNR-INFM BEC Center on Bose-Einstein Condensation, Trento, Italy.

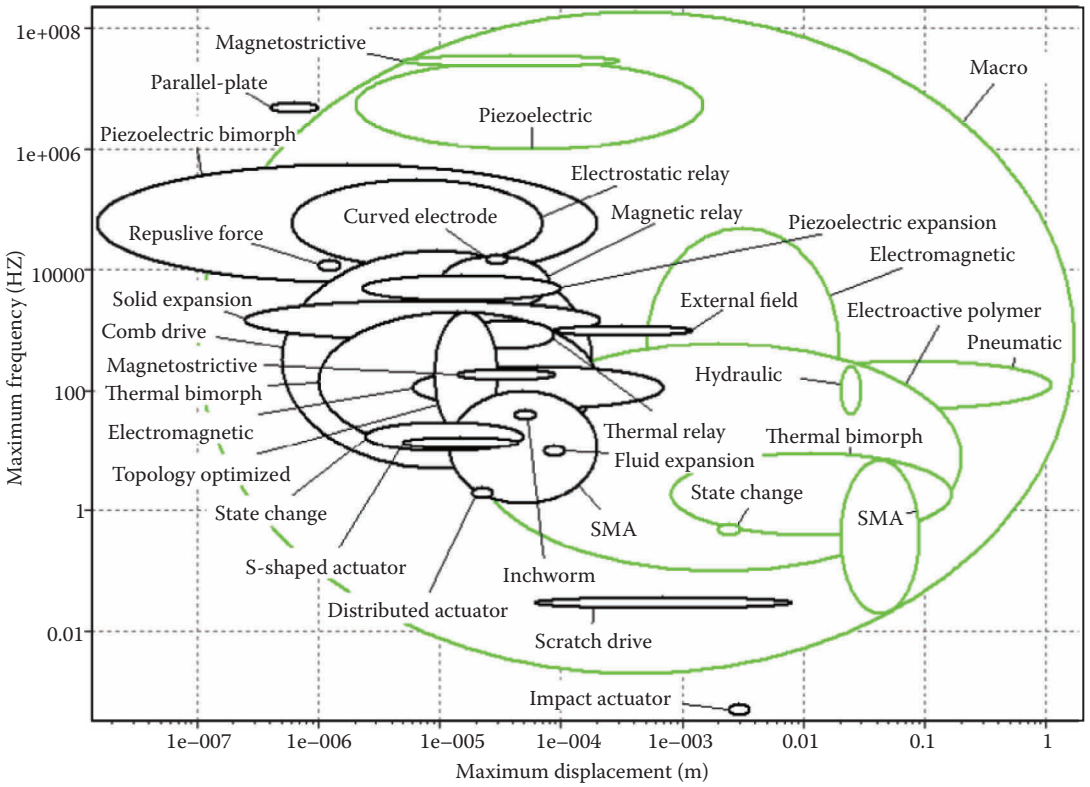


FIGURE 8.74 Comparison of all types of actuators in terms of maximum displacement versus maximum frequency.[171]

Example 8.1: Mass Sensing Actuators

In Volume I, Example 4.1, we analyzed the design criteria for a static mass sensitive piezoresistive cantilever. In that example, we were interested only in the static bending of the cantilever upon mass loading. Here we consider the mass sensitivity of different types of resonators, including cantilever beams. A mass sensitive resonator transforms an additional mass loading into

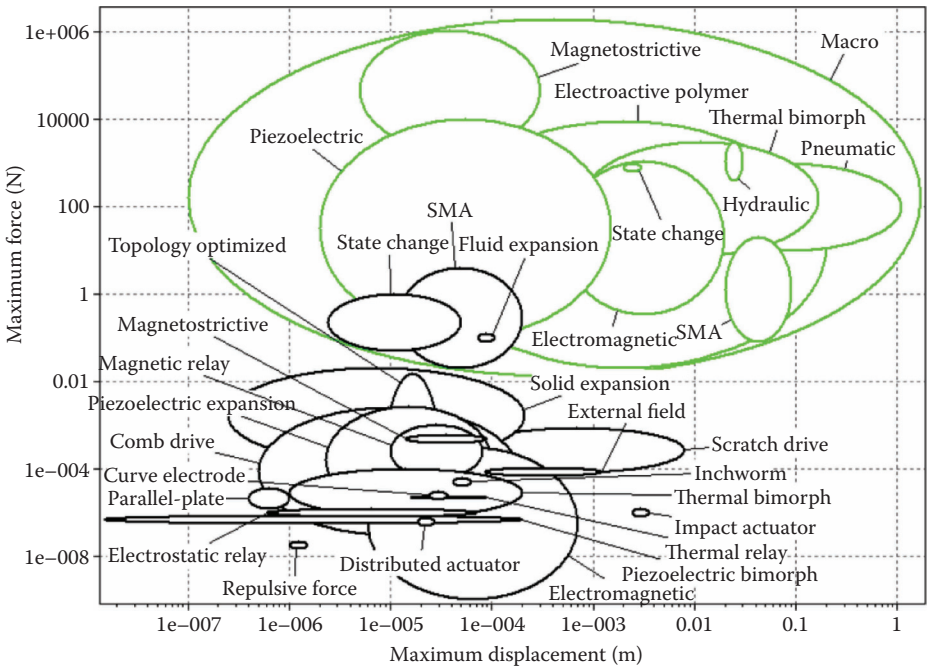


FIGURE 8.75 Comparison of all types of actuators in terms of maximum displacement versus maximum force exerted.[171]

a resonance frequency shift. Resonance frequency shifts of cantilevers can be measured with high precision using optical reflection with a diode laser and a linear position sensitive detector (PSD). The cantilever measuring principle may also be based on the detection of changes in phase, deflection, amplitude, and *Q*.[172] The excitation of MEMS resonators, say a comb drive or a cantilever, is most often realized by electrostatic actuation. The drawbacks of this approach include the relatively high supply voltage and the nonlinearity, linked to the varying distance between the electrodes. An alternative actuation scheme uses the combination of a high-quality piezoelectric thin film with a mechanical resonator. This has several advantages: it avoids the nanometer-size gaps typically required for electrostatic transduction and has a better power handling capability, better impedance handling with electronics, and better potential to reach 10 GHz and above (especially important for very sensitive mass-sensing applications).

Consider as the resonator a uniform beam with a rectangular cross-section, which is fixed at one end and deflected at the other (Figure 8.76). Such a beam experiences lengthwise stress that is compressive below the centerline and tensile above it. The effective spring constant, *K*, of this beam is given by the following:

$$K = EWt^3/4L^3 \qquad (8.94)$$

with *E* the Young's modulus and *W*, *t*, *L* the width, thickness, and length of the beam, respectively. Representative MEMS values for *W*, *t*, and *L* are 20, 0.6, and 100 μm, with a resulting value for *K* of 0.1 N/m.

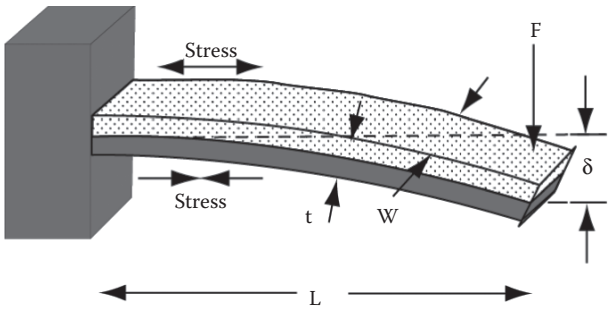


FIGURE 8.76 Stresses in a uniform cantilever beam when deflected.

For this simple cantilever beam the maximum deflection, δ, induced by a weight, W_e, at its free end is given by the following:

$$\delta = \frac{W_e L^3}{3EI} \qquad (8.95)$$

in which *E* and *I* are, respectively, the Young's modulus of the beam material and the moment of inertia of the beam, which is given as $Wt^3/12$. The weight $W_e = Mg$, that is, the mass (*M*) multiplied by the gravitational acceleration (*g*). The stiffness constant, *K*, from Equation 8.94, corresponds to the load/deflection, or $W_e/\delta = 3EI/L^3$, resulting in a natural frequency, f_0, of:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{3EI}{ML^3}} = \frac{1}{2\pi}\sqrt{\frac{EWt^3}{4ML^3}} = \frac{t}{4\pi L^2}\sqrt{\frac{E}{\rho}} \qquad (8.96)$$

where ρ is the density of the material.* For a Si-based stress-free MEMS cantilever with dimensions of 100 × 3 × 0.1 (*L* × *W* × *t*, in μm), the calculated resonant frequency is 12 kHz (Young's modulus E = 160 GPa); for a NEMS beam of 0.1 × 0.01 × 0.01 (*L* × *W* × *t*, in μm), the frequency is 1.2 GHz. Using a stiffer material (higher Young's modulus *E*) such as SiC (E = 400 GPa), the corresponding numbers are even higher, that is, 19 kHz and 1.9 GHz, respectively— reflecting the higher $(E/\rho)^{1/2}$ in Equation 8.96. In the NEMS range, mechanical devices are almost as fast as today's electronic devices. Making things yet smaller, the ultimate resonant frequency is reached in the THz range, that is, for molecular vibrations.[174] An intermediate frequency of 1.25 MHz was observed for an 8.3-μm-long, 95-nm-thick SiO_2 beam.[175]

In Figure 8.77 we show an example of a set of micromachined single-crystal silicon cantilevers. These beams can be used for real-time, in situ measurements of physical parameters such as viscosity,[176] pressure, density, flow rate and temperature. The latter, for instance, is achieved simply by coating a cantilever with metal on one side to form a bimetal.[177] Mercury vapor,[177] moisture, volatile mercaptans, DNA hybrization,[178] discrimination of single-nucleotide mismatches in DNA,[179] protein conformational changes,[180] and antibody-antigen binding all have been

* For a rigorous derivation, including the case of a thin cantilever that has been metalized with a layer that contributes to the functioning of the device, see Den Hartog.[173]

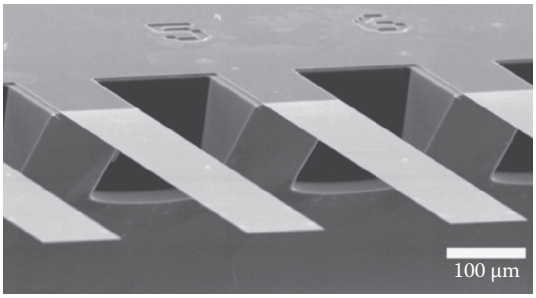


FIGURE 8.77 Scanning electron micrograph of a section of a microfabricated silicon cantilever array (eight cantilevers, each 1 μm thick, 500 μm long, and 100 μm wide, with a pitch of 250 μm, spring constant 0.02 Nm^{-1}. (Reprinted with permission from Fritz, J., M.K. Baller, H.P. Lang, H. Rothuizen, P. Vettinger, E. Meyer, H.-J. Güntherodt, C. Gerber, and J.K. Gimzewski. 2000. Translating biomolecular recognition into nanomechanics. *Science* 288:316. Copyright 2000, American Association for the Advancement of Science. Courtesy of Dr. J. Fritz, Massachusetts Institute of Technology Media Lab.[182])

monitored using cantilevers.[181] For chemical and biochemical sensing, the microcantilever surface must be derivatized with chemically selective coatings, perhaps through self-assembled alkanethiols or organosilane films, direct covalent attachment of molecular receptors, dip coating (e.g., polymeric resins), or adsorption/evaporation (e.g., sol-gel matrices).

Under the assumption of negligible variation in spring constant and uniformly distributed mass loading of the beam, Equation 8.96 describes the resonance frequency dependence on mass of the cantilever. Adding a discrete mass, M_d, to the cantilever defines a total effective mass, M_T, of the cantilever-adsorbate system as $M_T = M + M_d$. Equation 8.96 describes the mass dependence of a mass sensing cantilever, in which case the effective mass is that of the cantilever plus an adsorbate. The effective mass of the beam, M, is related to the mass of the beam, M_b, through $M = nM_b$, where n is a geometric parameter, which for a rectangular bar is 0.24.[183] This results in an expression for the resonant frequency given by the following:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{EWt^3}{4M_TL^3}} = \frac{1}{2\pi}\sqrt{\frac{EWt^3}{4L^3(M_d + 0.24WtL\rho)}} \tag{8.97}$$

where ρ is the density of the cantilever. To compare the mass sensitivity, S_m (given by Equation 8.64), of a cantilever sensor to that of other gravimetric devices, see Table 8.9.

In a case where mass loading is distributed evenly over the cantilever surface, S_m, as noted in Table 8.9, is given by the following (in Table 8.9, d was used for thickness instead of t).

$$S_m = \frac{1}{\rho t} \tag{8.98}$$

The positive sign indicates that, as mass is added, the resonant frequency increases, corresponding to the increase of the cross-sectional thickness of the beam. For an end-loaded cantilever (mass only allowed to accumulate in a small area at the free end of the beam) this results in:

$$S_m = \frac{-\xi}{2\rho(\xi t_d + 0.24t)} \tag{8.99}$$

where ξ and t_d are the fractional area coverage and thickness of the deposited mass, respectively. The negative sign denotes that the frequency is decreasing with increasing mass piling up.

From the above we conclude that for mass sensing with a microcantilever, a uniform mass loading is preferred. From Equations 8.98 and 8.99, it is further advantageous to make the cantilever as low density (ρ) and as thin as possible (t). Dramatic further improvements in analytical figures of merit can be achieved by scaling the cantilever size down to the nanoscale.

In Figure 8.78 we compare the size of a typical quartz crystal microbalance (QCM) transducer, with a minimal volume of 80 mm^3, with an array of ten micromachined cantilevers (e.g., based on silicon nitride) that requires as little as 0.001 mm^3 of volume.

From Table 8.9 cantilevers are much more sensitive than quartz microbalances and are

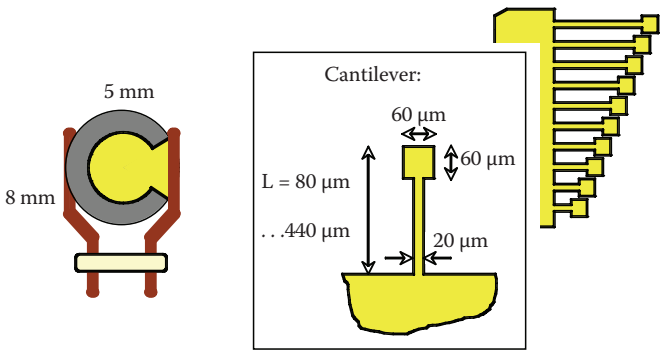


FIGURE 8.78 Size comparison of a quartz crystal microbalance and a micromachined array of cantilevers.

easier to downscale. Nanomechanical resonators enable the measurement of mass with extraordinary sensitivity. Samples as light as 7 zeptograms have been weighed in vacuum with cantilevers.[184] Resolving such small mass changes requires the resonator to be very light with a high *Q*. In solution, viscosity severely degrades both of these characteristics, preventing many sensing applications where fluid is required. One needs to excite with a positive feedback loop of a piezoelectric or measure in air or vacuum. FBARs, we saw earlier, can be downscaled more easily than BAWs, and for similar size they are even more sensitive than cantilevers (see Table 8.9).

An alternative approach, avoiding the viscous damping problem, is to use a static formulation, pioneered by Thundat et al. from Oak Ridge National Laboratory,[177] and involves the measurement of bending induced by the adsorption of molecules onto a thin microcantilever that has two chemically different surfaces. The Thundat et al. approach was covered in Volume I, Chapter 4 (Example 4.1).

Burg et al. pioneered yet a different approach to eliminate viscous damping by placing the solution inside a hollow resonator that is surrounded by vacuum.[185] This group demonstrated that suspended microchannel resonators can weigh single nanoparticles, single bacterial cells, and submonolayers of adsorbed proteins in water with subfemtogram resolution (1 Hz bandwidth). The combination of the low resonator mass (100 ng) and high-quality factor (Q = 15,000) enables an improvement in mass resolution of six orders of magnitude over a high-end commercial quartz crystal microbalance. This development gives access to intriguing applications, such as mass-based flow cytometry, the direct detection of pathogens, or the nonoptical sizing and mass density measurement of colloidal particles.

There are several (near) commercial biosensors based on microcantilever structures, as well as companies aiming at commercializing such products (e.g., Protiveris, Graviton, Concentris and Cantisens® Research). The current trend in SAW and QCM devices is to work at higher frequencies for better sensitivity, and in silicon MEMS, micro- and nanocantilvers are replacing quartz devices, as they are easier to integrate and easier to make into many device arrays. The key strength of all these devices is that they hold the promise of label-free measurements. In Chapter 10, this volume, we will also see that the RF-MEMS community is hard at work trying to replace quartz and AlN resonators with silicon MEMS-based oscillators.

Acknowledgments

With thanks to Ramses Madou, Marleen Madou, and George O'Quinn.

Questions

Questions and answers by Dr. Marc Madou, assisted by Mr. George O'Quinn, UC Berkeley

8.1: Fill in the blank with the correct word.
- (a) Piezoelectric materials produce _____ when they are deformed by a force.
- (b) Silicon is _____ piezoelectric.
- (c) Viscosity of a fluid is the tendency of the fluid to _____ .
- (d) Large field strength can be used to produce large _____ with smaller _____ .
- (e) _____ is the major loss mechanism in microactuators.
- (f) Shape memory alloys recover the shape they are given when they are heated to a temperature that is higher than the _____.

8.2: What is the difference between a sensor and an actuator?

8.3: What is the underlying principle of ferrofluids (FF) and how do they differ from magnetorheological fluids (MR)?

8.4: Describe, using equations, why one might prefer a parallel-driven parallel plate capacitor in an electrostatic actuator over a vertical-driven one?

8.5: Prove that the snap voltage V_{snap} in a parallel plate capacitor is given by 1/3 g_0, where g_0 is the initial gap spacing (with V = 0).

8.6: Make a table listing corresponding mechanical and electrical equivalent circuit components.

8.7: Suggest a normally closed valve technology based on phase-change of a material that

works for both vapors and liquids and does not require power to remain closed but only to open.

8.8: What transduction principles are used in biosensors?

8.9: What is the difference between a self-exciting and a modulating sensor? Give at least one example of each used to sense the same quantity.

8.10: Describe different methods to levitate and move a biological cell to a desired spot on a glass slide.

8.11: Why are wobble motors effective when small, yet ineffective when large, and what role does friction play in a wobble motor?

8.12: Explain the Seebeck effect, Peltier effect, Thomson effect, and Joule heating. How are they important to thermoelectric devices?

8.13: What are the characteristics of shape memory alloy (SMA) actuators that cause them to have a slower response time, but higher force than piezoelectric and electrostatic actuators?

8.14: What is the main difference between piezoelectricity and ferroelectricity?

8.15: How are Cooper pairs formed?

8.16: Why do good conductors make poor superconductors?

8.17: Why is the Casimir force an important consideration in MEMS and NEMS?

8.18: Explain magnetostriction and compare it to the piezoelectric effect.

8.19: Name five mechanical measurands.

References

1. White, R. M. 1987. A sensor classification scheme. *IEEE Trans Ultrason Ferroelectr Freq Control* UFFC-34:124–26.
2. Middlehoek, S., and S. Audet. 1989. *Silicon sensors.* London: Academic Press.
3. Lion, K. S. 1969. Transducers: Problems and prospects. *IEEE Trans Ind Electron* 16:2–5.
4. Gopel, W., J. Hesse, and J. N. Zemel, eds. 1989. *Sensors: A comprehensive survey.* New York: VCH.
5. Habekotte, E. 1993. In *Sensoren and Actuatoren in de Werktuigbouw/Machinebouw 15–96.* The Hague, the Netherlands: Centrum voor Micro-Electronica.
6. National Research Council. 1995. NRC.
7. Tang, W. C.-K. 1990. Electrostatic comb drive for resonant sensor and actuator applications. PhD thesis, University of California, Berkeley.
8. Tang, W. C., T.-C. H. Nguyen, and R. T. Howe. 1989. Laterally driven polysilicon resonant microstructures. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '89).* Salt Lake City, UT, February 20–22, 1989, pp. 53–59.
9. Denn, M. M. 1980. *Process fluid mechanics.* Englewood Cliffs, NJ: Prentice Hall.
10. Schmidt, M. A. 1988. Microsensors for the measurement of shear forces in turbulent boundary layers. Ph.D. thesis, Massachusetts Institute of Technology.
11. Trimmer, W. S. N., and K. J. Gabriel. 1987. Design considerations for a practical electrostatic micro-motor. *Sensors Actuators* 11:189–206.
12. Mohr, J., P. Bley, M. Strohrmann, and U. Wallrabe. 1992. Microactuators fabricated by the LIGA process. *J Micromech Microeng* 2:234–41.
13. Roy, S., A. K. McIlwain, R. G. DeAnna, A. J. Fleischman, R. K. Burla, C. A. Zorman, and M. Mehregany. 2000. SiC resonant devices for high Q and high temperature applications. *Solid-State Sensor and Actuator Workshop*, pp. 22–25. Hilton Head Island, SC: Transducers Research Foundation.
14. Roukes, M. L. 2000. Nanoelectromechanical systems. *Solid-State Sensor and Actuator Workshop*, pp. 367–76. Hilton Head Island, SC: Transducers Research Foundation.
15. Busch-Vishniac, I. J. 1992. The case for magnetically driven microactuators. *Sensors Actuators A* A33:207–20.
16. Mahadevan, R. 1987. In *Proceedings: IEEE Micro Robots and Teleoperators Workshop*, 15/1–8. Hyannis, MA.
17. Mahadevan, R. 1990. In *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '90).* Napa Valley, CA, February 11–14, 1990, pp. 120–27.
18. Kumar, S., D. Cho, and W. N. Carr. 1992. Experimental study of electric suspension for microbearings. *J Microelectromech Syst* 1:23–30.
19. Mehregany, M., P. Nagarkar, S. D. Senturia, and J. H. Lang. 1990. Operation of microfabricated harmonic and ordinary side-drive motors. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '90).* Napa Valley, CA, February 11–14, 1990, pp. 82–88.
20. Mehregany, M., S. D. Senturia, and J. H. Lang. 1990. Friction and wear in microfabricated harmonic side-drive motor. *Technical Digest: IEEE Solid-State Sensor and Actuator Workshop.* Hilton Head Island, SC, June 4–7, 1990, pp. 17–22.
21. Fujita, H., and K. J. Gabriel. 1991. New opportunities for micro actuators. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91).* San Francisco, CA, June 1991, pp. 14–20.
22. Trimmer, W. S., ed. 1997. *Micromechanics and MEMS classic and seminal papers to 1990.* New York: IEEE.
23. Price, R. H., S. J. Cunningham, and S. C. Jacobsen. 1992. Field analysis for the electrostatic eccentric drive micromotor (wobble motor). *J Electrost* 28:7–38.
24. Furuhata, T., T. Hirano, L. H. Lane, R. E. Fontana, L. S. Fan, and H. Fujita. 1993. Outer rotor surface-micromachined wobble motor. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '93).* Fort Lauderdale, FL, February 7–10, 1993, pp. 161–66.
25. Tai, Y., L. Fan, and R. Muller. 1989. IC processed micromotors: design, technology, and testing. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '89).* Salt Lake City, UT, February 20–22, 1989, pp. 1–6.
26. Fujita, H. 1990. In *Proceedings: Micro System Technologies '90.* Berlin, Germany, p. 818.

27. Ohnstein, T., T. Fukiura, J. Ridley, and V. Bonne. 1990. Micromachined silicon microvalve. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '90)*. Napa Valley, CA, February 11–14, 1990, pp. 95–98.
28. Marxer, C., O. Manzardo, H.-P. Herzig, R. Dändliker, and N. F. de Rooij. 1999. An electrostatic actuator with large dynamic range and linear displacement-voltage behaviour for a miniature spectrometer. *Transducers '99* 1:786–89.
29. Petersen, K. 1979. Micromechanical membrane switches on silicon. *IBM J Res Dev* Report no. 23:376.
30. Cady, W. G. 1922. The piezo-electric resonator. *Proceedings of the Institute of Radio Engineers,* April 1922, pp. 83–114.
31. Fukuda, E., and I. Yasuda. 1957. On the piezoelectric effect in bone. *J Phys Soc Jpn* 12:1158–62.
32. Gandhi, M. V., and B. S. Thompson. 1992. *Smart materials and structures*. London: Chapman and Hall.
33. Hueter, T. F., and R. H. Bolt. 1955. *Sonics.* New York: Wiley.
34. Pallas-Areny, R., and J. G. Webster. 1991. *Sensors and signal conditioning.* New York: Wiley.
35. Herbert, J. M. 1982. *Ferroelectric transducers and sensors.* New York: Gordon and Breach Science Publishers.
36. Khazan, A. D. 1994. *Transducers and their elements.* Englewood Cliffs, NJ: PTR Prentice Hall.
37. Rogacheva, N. N. 1994. The theory of piezoelectric shells and plates. Boca Raton, FL: CRC Press.
38. Cady, W. G. 1964. *Piezoelectricity.* New York: Dover.
39. Zelenka, J. 1986. *Piezoelectric resonators and their applications.* Amsterdam: Elsevier.
40. White, R. M., and F. W. Voltmer. 1965. Direct piezoelectric coupling to surface elastic waves. *Appl Phys Lett* 7:314–316.
41. Flynn, A. M., L. S. Tavrow, S. F. Bart, R. A. Brooks, D. J. Ehrlich, K. R. Udayakumar, and L. E. Cross. 1992. Piezoelectric micromotors. *J Microelectromech Syst* 1:44–52.
42. Abe, T., and M. L. Reed. 1994. In *IEEE International Workshop on Micro Electro Mechanical Systems (MEMS '94)*, 164–69. Oiso, Japan.
43. Hunter, I. W., and S. Lafontaine. 1992. A comparison of muscle with artificial actuators. *Technical Digest: 1992 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC, June 22–25, 1992, pp. 178–85.
44. AMP. 1993. *Piezo film sensors.* Valley Forge, PA: AMP.
45. Nalwa, H. S., ed. 1995. *Ferroelectric polymers: Chemistry, physics, and applications.* New York: Marcel Dekker.
46. Polla, D., and R. S. Muller. 1986. Zinc-oxide thin films for integrated-sensor applications. *Technical Digest: IEEE Solid-State Sensors Workshop.* Hilton Head Island, SC, June 2–5, 1986.
47. Muller, R. S. 1987. From ICs to microstructures materials and technologies. *Proceedings: IEEE Micro Robots and Teleoperators Workshop*. Hyannis, MA, November 9–11, 1987, IEEE Catalog No. 87TH0204-8.
48. Tjhen, W., T. Tamagawa, C.-P. Ye, C.-C. Hsueh, P. Schiller, and D. L. Polla. 1991. In *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '91)*. Nara, Japan, pp. 114–19.
49. Hirata, Y., H. Okuyama, S. Ogino, T. Numazawa, and H. Takada. 1995. In *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '95)*. Amsterdam, the Netherlands, pp. 191–96.
50. Preu, G., A. Wolff, D. Cramer, and U. Bast. 1991. In *Proceedings: Second European Ceramic Society Conference (2nd ECerS '91)*, 2005–9. Augsburg, Germany.
51. Lubitz, K. 1989. *Mikrostrukturierung von piezokeramik.* Berlin: VDI-Tagungsbericht, 796.
52. Son, I.-S., A. Lal, B. Hubbard, and T. Olson. 2000. In *Solid-State Sensor and Actuator Workshop,* 206–9. Hilton Head Island, SC: Transducers Research Foundation.
53. Christensen, D. A. 1988. *Ultrasonic bioinstrumentation.* New York: Wiley.
54. Udayakumar, K. R., S. F. Bart, A. M. Flynn, J. Chen, L. S. Tavrow, L. E. Cross, R. A. Brooks, and D. J. Ehrlich. 1991. Ferroelectric thin films for ultrasonic micromotors. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '91)*. Nara, Japan, January 30–February 2, 1991, pp. 109–13.
55. van Lintel, H. T. G., F. C. M. van der Pol, and S. Bouwstra. 1988. A piezoelectric micropump based on micromachining of silicon. *Sensors Actuators* 15:153–67.
56. Smits, J. G. 1990. Piezoelectric micropump with three valves working peristaltically. *Sensors Actuators A* A21:2030–6.
57. Judy, J. W., D. L. Polla, and W. P. Robbins. 1990. Experimental model and IC-process design of a nanometer linear piezoelectric stepper motor. *Microstructures, Sensors, and Actuators*. American Society of Mechanical Engineers, November 1990, pp. 11–17.
58. Shoji, S., S. Nakagawa, and M. Esashi. 1990. Micropump and sample-injector for integrated chemical analyzing systems. *Sensors Actuators A* A21:189–92.
59. Nyborg, W. L. M. 1965. Acoustic streaming. In *Physical acoustics,* ed. W. P. Mason, 265–331. New York: Academic Press.
60. Moroney, R. M., R. M. White, and R. T. Howe. 1990. Ultrasonic micromotors: physics and applications. *IEEE Micro Electro Mechanical Systems.* New York, February 1990, pp. 182–87.
61. Moroney, R. M., R. M. White, and R. T. Howe. 1991. Microtransport induced by ultrasonic waves. *Appl Phys Lett* 59:774–76.
62. Miyazaki, S., T. Kawai, and M. Araragi. 1991. A piezo-electric pump driven by a flexural progressive wave. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '91)*. Nara, Japan, January 30–February 2, 1991, pp. 283–88.
63. Ballantine, D. S., R. M. White, S. L. Martin, A. J. Ricco, E. T. Zellers, G. C. Frye, and H. Wohltjen. 1997. *Acoustic wave sensors: Theory, design, and physico-chemical applications*. San Diego, CA: Academic Press.
64. Sauerbrey, G. Z. 1959. The use of an oscillator for weighing thin layers and for microweighing. *Z Phys* 155:206–12.
65. Black, J. P., A. Elium, R. M. White, M. G. Apte, L. A. Gundel, and R. Cambie. 2007. 6D-2 MEMS-enabled miniaturized particulate matter monitor employing 1.6 GHz aluminum nitride thin-film bulk acoustic wave resonator (FBAR) and thermophoretic precipitator. *Ultrasonics symposium, 2007,* New York, October 28–31, 2007, pp. 476–79.
66. Kobrin, P., C. Seabury, A. Harker, and R. O'Toole. 1999. *Thin-film resonant chemical sensor with resonant acoustic isolator*. US Patent 5936150.
67. Wohltjen, H., and R. E. Dessy. 1979. Surface acoustic wave probe for chemical analysis, parts I–III. *Anal Chem* 5:1458–75.
68. Hoenk, M. E., T. R. Van Zandt, D. A. McWatters, R. K. Watson, C. Kukkonen III, W. Kaiser, and D. Cheng. 1995. *Surface acoustic wave hygrometer flight tests on the NASA DC-8 airborne laboratory.* Pasadena, CA: Jet Propulsion Laboratory.
69. Bernstein, J. J., J. Bottari, K. Houston, G. Kirkos, and R. Miller. 2000. In *Solid-State Sensor and Actuator Workshop,* 281–84. Hilton Head Island, SC: Transducers Research Foundation.

70. Allen, R. 1984. Sensors in silicon. *High Technology,* September: 43–81.
71. Kolesar, E. S., and C. S. Dyson. 1995. Object imaging with a piezoelectric robotic tactile sensor. *J Microelectromech Syst* 4:87–96.
72. Glaser, K., and G. Fuhr. 1987. The spin of cells in rotating high frequency electric fields. In *Mechanistic approaches to interactions of electric and electromagnetic fields with living systems,* eds. M. Blank and E. Findl. New York: Plenum Press, 271–89.
73. Fuhr, G., R. Hagedorn, T. Muller, B. Wagner, and W. Benecke. 1991. Linear motion of dielectric particles and living cells in microfabricated structures induced by travelling electric fields. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '91)*. Nara, Japan, January 30–February 2, 1991, pp. 259–64.
74. Moesner, F. M., and T. Higuchi. 1995. Devices for particle handling by an ac electric field. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '95)*. Amsterdam, the Netherlands, January 29–February 2, 1995, pp. 66–71.
75. Fuhr, G., R. Hagedorn, T. Muller, W. Benecke, and B. Wagner. 1992. Microfabricated electrohydrodynamic (EHD) pumps for liquids of higher conductivity. *J Microelectromech Syst* 1:95–98.
76. Bart, S. F., L. S. Tavrow, M. Mehregany, and J. H. Lang. 1990. Microfabricated electrohydrodynamic pumps. *Sensors Actuators A* A21:193–97.
77. Shoji, S. 1998. In *Microsystem technology in chemistry and life sciences,* eds. A. Manz and H. Becker, 161–88. New York: Springer.
78. Richter, A., A. Plettner, K. Hoffmann, and H. Sandmaier. 1991. Electrohydrodynamic pumping and flow measurement. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '91)*. Nara, Japan, January 30–February 2, 1991, pp. 271–76.
79. Choi, J.-W., and C. H. Ahn. 2000. An active micro mixer using electrohydrodynamic (EHD) convection. *Solid-State Sensor and Actuator Workshop*. Hilton Head Island, SC, June 4–8, 2000, pp. 52–55.
80. Bart, S. F., T. A. Lober, R. T. Howe, J. H. Lang, and M. F. Schlecht. 1988. Design considerations for micromachined electric actuators. *Sensors Actuators* 14:269–92.
81. Melcher, J. R. 1966. Traveling-wave induced electroconvection. *Phys Fluids* 9:1548–55.
82. Bart, S., and J. H. Lang. 1989. An analysis of electroquasistatic induction micromotors. *Sensors Actuators* 20:97–106.
83. Fripp, M. L. R., and N. W. Hagood. 1995. Comparison of electrostrictive and piezoceramic actuation for vibration suppression (Proceedings paper). In *Smart structures and materials 1995: Smart structures and integrated systems,* ed. I. Chopra. Bellingham, WA: SPIE, May 8, 1995, pp. 334–48.
84. Uchino, K. 1993. Ceramic actuators: Principles and applications. *MRS Bull* 18:42.
85. Lang, S. B. 1974. Sourcebook of pyroelectricity. New York: Gordon and Breach.
86. Ye, C., T. Tamagawa, and D. L. Polla. 1991. Pyroelectric PbTiO3 thin films for microsensor applications *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, CA, June 24–27, 1991, pp. 904–07.
87. Hsieh, H. Y., A. Spetz, and J. N. Zemel. 1991. Wide range pyroelectric anemometers for gas flow measurements. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, CA, June 24–27, 1991, pp. 38–40.
88. Zavracky, P. M., K. Warner, G. Jenkins, S. Etienne, C. Logan, and R. Grace. 1998. In *Solid-State Sensor and Actuator Workshop*. Hilton Head Island, SC, 209–11.
89. Hicks, L. D., and M. S. Dresselhaus. 1993. Effect of quantum-well structures on the thermoelectric figure of merit. *Phys Rev B Condens Matter* 47:727–31.
90. Yang, P., A. I. Hochbaum, R. Chen, R. D. Delgado, W. Liang, E. C. Garnett, M. Najarian, and A. Majumdar. 2008. Rough silicon nanowires as high performance thermoelectric materials. *Nature* 451:163–67.
91. Heath, J. R., A. Boukai, Y. Bunimovich, J. Tahir-Kheli, J. Yu, and W. A. Goddard. 2008. Silicon nanowires as efficient thermoelectric materials. *Nature* 451:168–71.
92. Shakouri, A. 2006. Nanoscale thermal transport and microrefrigerators on a chip. *Proceedings of the IEEE* 94:1613–38.
93. Voorthuyzen, J. A., and P. Bergveld. 1988. The PRESSFET: An integrated electret-MOSFET based pressure sensor. *Sensors Actuators* 14:349–60.
94. Voorthuyzen, J. A., P. Bergveld, and A. J. Sprenkels. 1989. Semiconductor-based electret sensors for sound and pressure. *IEEE Trans Electr Insul* 24:267–76.
95. Sprenkels, A. J., R. A. Groothengel, A. J. Verloop, and P. Bergveld. 1989. Development of an electret microphone in silicon. *Sensors Actuators* 17:509–12.
96. Wolffenbuttel, R. F., J. F. L. Goosen, and P. M. Sarro. 1991. Design considerations for a permanent-rotor-charge excited micromotor with an electrostatic bearing. *Sensors Actuators A* A25–27:583–90.
97. Mascarenhas, S. 1973. The electret state: A new property of bone. *Electrets,* ed. M. M. Perlman Princeton, NJ: The Electrochemical Society, p. 650.
98. Mascarenhas, S. 1974. The electret effect in bone and polymers and the bound-water problem. *Ann NY Acad Sci* 238:36.
99. Fukada, E., T. Takamaster, and I. Yasuda. 1975. Callus formation by electret. *Jpn J Appl Phys* 14:2079.
100. Duff, A. W. 1896. The viscosity of polarized dielectrics. *Phys Rev* 41:23–38.
101. Winslow, W. M. 1947. *Method and means for translating electrical impulses into mechanical force*. US Patent 2,417,850.
102. Winslow, W. M. 1947. Induced vibration of suspensions. *J Appl Phys* 20:1137–40.
103. Rabinow, J. 1948. The magnetic fluid clutch. *AIEE Trans* 67(II):1308–15.
104. Gandhi, M. V., and B. S. Thompson. 1992. *Smart materials and structures.* London: Chapman & Hall.
105. Ball, P. 1997. *Made to measure: New materials for the 21st century.* Princeton, NJ: Princeton University Press.
106. Wen, W., X. Huang, S. Yang, K. Lu, and P. Sheng. 2003. The giant electrorheolgical effect in suspension of nanoparticles. *Nature Materials* 2:727–30.
107. Zipernowsky. 1889. Zipernowsky electrostatic motor. *Electr World* 14:260.
108. Trimmer, T. 1990. Micromechanical systems. *Proceedings, Integrated Micro-Motion Systems: Micromachining, Control, and Applications (3rd Toyota Conference)*. Aichi, Japan, October 1990, pp. 1–15.
109. Trimmer, W. S. N. 1989. Microrobots and micromechanical systems. *Sensors Actuators* 19:267–87.
110. Kim, Y.-K., M. Katsurai, and H. Fujita. 1990. Fabrication and testing of a micro superconducting actuator using the Meissner effect. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '90)*. Napa Valley, CA, February 11–14, 1990, pp. 61–66.

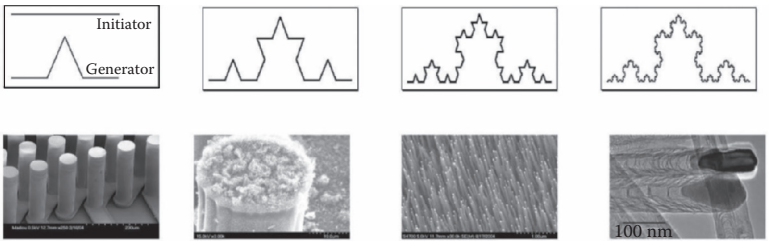
111. Pelrine, R., and I. Busch-Vishniac. 1987. Magnetically levitated micro-machines. *Proceedings: IEEE Micro Robots and Teleoperators Workshop*. Hyannis, MA, November 1987, 19/1–5.
112. Rousselet, J., G. H. Markx, and R. Pethig. 1998. Separation of erythrocytes and latex beads by dielectrophoretic levitation and hyperlayer field-flow fractionation. *Colloids Surfaces A* 140:209–16.
113. Kumar, S., D. Cho, and W. N. Carr. 1992. Experimental study of electric suspension for microbearings. *J Microelectromech Syst* 1:23–30.
114. Pister, K. S. J., R. S. Fearing, and R. T. Howe. 1990. A planar air levitated electrostatic actuator system. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '90)*. Napa Valley, CA, February 11–14, 1990, pp. 67–71.
115. Wagner, B., M. Kreutzer, and W. Benecke. 1993. Permanent magnet micromotors on silicon substrates. *J Microelectromech Syst* 2:23–29.
116. Benecke, W. 1991. Silicon-microactuators: Activation mechanisms and scaling problems. In *1991 International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, CA, June 24–27, 1991, pp. 46–50.
117. Wagner, B., M. Kreutzer, and W. Benecke. 1991. Electromagnetic microactuators with multiple degrees of freedom. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, CA, June 24–27, 1991, pp. 614–17.
118. Gongora-Rubio, M., L. Sola-Laguna, M. Smith, and J. J. Santiago-Aviles. 1999. In *Microfluidic devices and systems II (proceedings of the SPIE)*. Santa Clara, CA: SPIE, 230–39.
119. Price, R. H., S. J. Cunningham, and S. C. Jacobsen. 1992. Field analysis for the electrostatic eccentric drive micromotor (wobble motor). *J Electrost* 28:7–38.
120. Goemans, P. A. F. M. 1994. Microsystems and energy: the role of energy. *Microsystem technology: Exploring opportunities*, ed. G. K. Lebbink. Alphen aan de Rijn/Zaventem: Samsom BedrijfsInformatie bv, pp. 50–64.
121. Christenson, T. R., H. Gückel, K. J. Skrobis, and J. Klein. 1992. In *Solid-State Sensors and Actuators Workshop*. Hilton Head Island, SC, June 22–25, 1992, pp. 51–54.
122. Guckel, H., K. J. Skrobis, T. R. Christenson, J. Klein, S. Han, B. Choi, E. G. Lovell, and T. W. Chapma. 1991. Fabrication and testing of the planar magnetic micromotor. *J Micromech Microeng* 1:135–38.
123. Ahn, C. H., Y. J. Kim, and M. G. Allen. 1993. A planar variable reluctance micromotor with fully integrated stator and wrapped coils. *IEEE microelectromechanical systems*. Fort Lauderdale, FL, February 1993, pp. 1–6. New York: IEEE.
124. Wallrabe, U., P. Bley, B. Krevet, W. Menz, and J. Mohr. 1994. Design rules and test of electrostatic micromotors made by the LIGA process. *J Micromech Microeng* 4:40–45.
125. Busch-Vishniac, I. J. 1992. The case for magnetically driven microactuators. *Sensors Actuators A* A33:207–20.
126. Edelstein, R. L., C. R. Tamanaha, P. E. Sheenan, M. M. Miller, D. R. Baselt, L. J. Whitman, and R. J. Colton. 2000. The BARC biosensor applied to the detection of biological warfare agents. *Biosens Bioelectron* 14:805–13.
127. Busch, J. D., and A. D. Johnson. 1990. Shape-memory properties in Ni-Ti sputter-deposited film. *J Appl Phys* 68:6224–28.
128. Fujita, H., and A. Omodaka. 1987. Electrostatic actuators for micromechatronics. *Proceedings: IEEE Micro Robots and Teleoperators Workshop*. Hyannis, MA , November 1987, 14/1–10.
129. Fujita, H. 1990. In *Proceedings: Micro System Technologies '90*, 818. Berlin, Germany.
130. Takeshima, N., K. J. Gabriel, M. Ozaki, J. Takahashi, H. Horiguchi, and H. Fujita. 1991. Electrostatic parallelogram actuators. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, CA, June 24–27, 1991, pp. 63–66.
131. Romankiw, L. T. 1989. In *Proceedings: Symposium on Magnetic Materials, Processes, and Devices*, 39–53. Hollywood, FL.
132. Cardot, F., J. Gobet, M. Bogdanski, and F. Rudolf. 1994. Fabrication of a magnetic transducer composed of a high-density array of microelectromagnets with on-chip electronics. *Sensors Actuators A* A43:11–16.
133. Ahn, C. H., Y. J. Kim, and M. G. Allen. 1993. A fully integrated micromachined toroidal inductor with a nickel-iron magnetic core (the switched dc/dc converter application). *7th International Conference on Solid-State Sensors and Actuators (Transducers '93)*. Yokohama, Japan, June 7–10, 1993, pp. 70–73.
134. Honda, T., K. I. Arai, and M. Yamaguchi. 1994. Fabrication of actuators using magnetostrictive thin films. In *IEEE International Workshop on Micro Electro Mechanical Systems (MEMS '94)*. Oiso, Japan, January 25–28, 1994, pp. 51–56.
135. Quandt, E., and K. Seemann. 1995. Fabrication of giant magnetostrictive thin film actuators. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '95)*. Amsterdam, the Netherlands, January 29–February 2, 1995, pp. 273–77.
136. Rombach, P., and W. Langheinrich. 1994. An integrated sensor head in silicon for contactless detection of torque and force. *Sensors Actuators A* A41–42:410–16.
137. Hathaway, K. B., and A. E. Clark. 1993. Magnetostrictive materials. *MRS Bull* April: 34.
138. Fukuda, T., H. Hosokai, H. Ohyama, H. Hashimoto, and F. Arai. 1991. Giant magnetostrictive alloy (GMA) applications to micro-mobile robot as a micro actuator without power supply cables. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '91)*. Nara, Japan, January 30–February 2, 1991, pp. 210–15.
139. Panholzer, R. 1963. Electromagnetic pumps. *Elec Eng* 2:128–35.
140. Heng, K.-H., L. Huang, W. Wang, and M. C. Murphy. 1999. Development of a diffuser/nozzle type micropump based on magnetohydrodynamic (MHD) principle. *Microfluidic devices and systems II*, eds. C. H. Ahn and A. B. Frazier. San Jose, CA, September 20–21, 1999, pp. 66–73.
141. Lemoff, A. V., A. P. Lee, R. R. Miles, and C. F. McConaghy. 1999. An ac magnetohydrodynamic micropump: towards a true integrated microfluidic system. *Technical Digest: 10th International Conference on Solid-State Sensors and Actuators*. Sendai, Japan, pp. 1126–29.
142. Lemoff, A. V., and A. P. Lee. 2000. In *Micro total analysis systems*, eds. A. van den Berg, W. Olthuis, and P. Bergveld. Enschede, the Netherlands: Kluwer Academic Publishers.
143. Phule, P., and J. Ginder. 1998. The materials science of field-responsive fluids. *MRS Bull* 23:19–21.

144. Herzer, G. 1997. Nanocrystalline soft magnetic alloys. In *Handbook of magnetic materials,* ed. K. H. J. Buschow, 417–61. Amsterdam: Elsevier Science B.V.
145. Rinaldi, C., and M. Zahn. 2002. Effects of spin viscosity on ferrofluid duct flow profiles in alternating and rotating magnetic fields. *J Magnetism Magnetic Materials* 252: 172–75.
146. Lorenz, C., and M. Zahn. 2003. Hele-Shaw ferrohydrodynamics for rotating and dc axial magnetic fields. 2003 Gallery of Fluid Motion. *Phys Fluids* 15:S4.
147. Hartshorne, H., C. J. Backhouse, and W. E. Lee. 2004. Ferrofluid-based microchip pump and valve. *Sensors Actuators B* 99:592–600.
148. Hatch, A., A. Kamholz, G. Holman, P. Yager, and K. Bohringer. 2001. A ferrofluidic magnetic micropump. *J Microelectromech Syst* 10:215–21.
149. Pal, R., M. Yang, B. N. Johnson, D. T. Burke, and M. A. Burns. 2004. Phase change microvalve for integrated devices. *Anal Chem* 76:3740–48.
150. Liu, R. H., J. Bonanno, J. Yang, R. Lenigk, and P. Grodzinski. 2004. Single-use, thermally actuated paraffin valves for microfluidic applications. *Sens Actuators* B 98:328–36.
151. Oh, K. W., K. Namkoong, and C. Park. 2005. A phase change microvalve using meltable magnetic material: Ferro-wax. In *9th International Conference on Miniaturized Systems for Chemistry and Life Sciences: uTAS*. Boston, MA, October 9–13, 2005, pp. 554–56.
152. Park, J. M., Y. Cho, B.-S. Lee, J.-G. Lee, and C. Ko. 2007. Multifunctional microvalves controlled by optical illumination on nanoheaters and its application in centrifugal microfluidic devices. *Lab Chip* 7:557–64.
153. Zoval, J., and M. J. Madou. 2004. Centrifuge-based fluidic platforms. *Proc IEEE* 92(1):140–53.
154. Zdeblick, M. J., R. Anderson, J. Jankowski, B. Kline-Schoder, L. Christel, R. Miles, and W. Weber. 1994. Thermopneumatically actuated microvalves and integrated electro-fluidic circuits. *Technical Digest: 1994 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC, June 13–16, 1994, pp. 251–55.
155. Barth, P. W. 1995. Silicon microvalves for gas flow control. *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, pp. 276–80.
156. Busch, J. D., and A. D. Johnson. 1990. Shape-memory properties in Ni-Ti sputter-deposited film. *J Appl Phys* 68:6224–28.
157. Walker, J. A., K. J. Gabriel, and M. Mehregany. 1990. Thin-film processing of TiNi shape memory alloy. *Sensors Actuators A* A21:243–46.
158. Johnson, A. D. 1991. Vacuum-deposited TiNi shape memory film: Characterization and applications in microdevices. *J Micromech Microeng* 1:34–41.
159. Dario, P., M. Bergamasco, L. Bernardi, and A. Bicchi. 1987. A shape memory alloy actuating module for fine manipulation. *Proceedings: IEEE Micro Robots and Teleoperators Workshop*. Hyannis, MA, November 1987, 16/1–5.
160. Quandt, E., C. Halene, H. Holleck, K. Feit, M. Kohl, and P. Schlossmacher. 1995. Sputter deposition of TiNi and TiNiPd films displaying the two way shape memory effect. *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, pp. 202–05.
161. Hunter, I. W., and S. Lafontaine. 1992. A comparison of muscle with artificial actuators. *Technical Digest: 1992 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC, June 22–25, 1992, pp. 178–85.
162. Gandhi, M. V., and B. S. Thompson. 1992. *Smart materials and structures.* London: Chapman & Hall.
163. Gordon, G. B., and P. W. Barth. 1991. Thermally-actuated microminiature valve, Hewlett-Packard, Palo Alto, CA. US Patent 5058856.
164. Barth, P. W., C. C. Beatty, and L. A. Field. 1981. Characterization of the silicon electrode. *Surf Sci* 108:135–52.
165. Jerman, H. 1991. Electrically-activated, normally-closed diaphragm valves. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, CA, June 24–27, 1991, pp. 1045–48.
166. Editorial. 1995. *Microflow Analytical*.
167. Hamberg, M. W., C. Neagu, J. G. E. Gardeniers, D. J. Ijntema, and M. Elwenspoek. 1995. Electrochemical micro actuator. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '95)*. Amsterdam, the Netherlands, January 29–February 2, 1995, pp. 106–10.
168. Nagakura, T., K. Ishihara, T. Furukawa, K. Masuda, and T. Tsuda. 1995. Auto-regulated medical pump without energy supply. *8th International Conference on Solid-State Sensors and Actuators (Transducers '95)*. Stockholm, Sweden, June 25–29, 1995, pp. 287–90.
169. Brannon-Peppas, L. 1997. Polymers in controlled drug delivery. *Med Plastics Biomater* 4:34–44.
170. Lamoreaux, S. K. 1997. Demonstration of the Casimir force in the 0.6 to 6 μm range. *Phys Rev Lett* 78:5.
171. Bell, D. J., T. J. Lu, N. A. Fleck, and S. M. Spearing. 2005. MEMS actuators and sensors: Observations on their performance and selection. *J Microelectromech Syst* 15:S153–64.
172. Finot, E., A. Passian, and T. Thundat. 2008. Measurement of mechanical properties of cantilever shaped materials. *Sensors* 8:3497–541.
173. Den Hartog, J. P. 1956. *Mechanical vibrations.* New York: McGraw-Hill.
174. Roukes, M. L. 2000. In *Solid-State Sensor and Actuator Workshop*. Hilton Head Island, SC: Transducers Research Foundation, pp. 367–76.
175. Petersen, K. E. 1982. Silicon as a mechanical material. *Proc IEEE* 70:420–57.
176. Oden, P. I., G. Y. Chen, R. A. Steele, R. J. Warmack, and T. Thundat. 1996. Viscous drag measurements utilizing microfabricated cantilevers. *Appl Phys Lett* 68:1465–69.
177. Thundat, T., P. I. Oden, and R. J. Warmack. 1997. Microcantilever sensors. *Nanoscale Microscale Thermophys Eng* 1:185–99.
178. Thundat, T., G. Y. Chen, R. J. Warmack, D. P. Allison, and E. A. Wachter. 1995. Vapor detection using resonating microcantilevers. *Anal Chem* 67:519–21.
179. Hansen, K. M., H.-F. Ji, G. Wu, R. Datar, R. Cote, A. Majumdar, and T. Thundat. 2001. Cantilever-based optical deflection assay for discrimination of DNA single-nucleotide mismatches. *Anal Chem* 73:1567–71.
180. Moulin, A. M. 2000. Microcantilever-based biosensors. *Ultramicroscopy* 82:23–31.
181. Thundat, T., L. A. Bottomley, S. Meller, W. H. Velander, and R. Van Tassell. 2001. In *Immunoassays: Methods and protocols,* eds. A. L. Ghindilis, A. R. Pavlov, and P. B. Atanajov. Totowa, NJ: Humana Press.

182. Fritz, J., M. K. Baller, H. P. Lang, H. Rothuizen, P. Vettinger, E. Meyer, H.-J. Güntherodt, C. Gerber, and J. K. Gimzewski. 2000. Translating biomolecular recognition into nanomechanics. *Science* 288:316–18.
183. Chen, G. Y., T. Thundat, E. A. Wachter, and R. J. Warmack. 1995. Adsorption-induced surface stress and its effects on resonance frequency of cantilevers. *J Appl Phys* 77:3618–22.
184. Yang, Y. T., C. Callegari, X. L. Feng, K. L. Ekinci, and M. L. Roukes. 2006. Zeptogram-scale nanomechanical mass sensing. *Nano Lett* 6:583–86.
185. Burg, T. P., M. Godin, S. M. Knudsen, W. Shen, G. Carlson, J. S. Foster, K. Babcock, and S. R. Manalis. 2007. Weighing of biomolecules, single cells and single nanoparticles in fluid. *Nature* 446:1066–69.

9

Power and Brains in Miniature Devices



Different orders of a Koch curve seen in parallel with carbon fractal electrodes at different magnifications. Fractal electrodes can minimize internal resistance while maximizing surface-to-volume ratios. Carbon-MEMS shows great promise for use in the fabrication of fractal electrodes for electrochemical applications.

Outline

Introduction
Power in MEMS and NEMS
MEMS and NEMS Batteries, Capacitors, Fuel Cells, Combustion Engines, and Solar Cells
Fuel Cells
Microfuel Cells
Fuels and Fuel Storage in Microfuel Cells
Heat Engines in MEMS
Solar Cells
MEMS and NEMS in Solar Cells
Energy Scavenging (Harvesting)
Fractal Power
Brains in Miniature Devices
Artificial Intelligence
Questions
References

Introduction

In this chapter we consider power and brains in miniature devices. We analyze the challenges and promises of powering miniaturized devices and of the miniaturization of power sources themselves. The power section starts with a comparison of onboard energy sources, such as batteries, capacitors, fuel cells, internal combustion engines, and solar cells. Next, we consider energy scavenging methods, such as taking advantage of environmental vibrations or thermal gradients.

Miniaturization of power sources scales disadvantageously, as the specific energy (energy per volume unit: l^3) of the power source determines the volume for a given application. That volume, besides the active materials (say, Li in a Li-ion battery), must also include packaging, and packaging efficiency degrades as the power source gets smaller.

As in the case of cell phones, personal digital assistants (PDAs), distributed sensor networks, portable computers, etc., small, lightweight, and long-lasting energy sources are one of the most urgently needed MEMS and NEMS breakthrough technologies. A pertinent quote from *Wired* magazine in April 2004 reads, "If we don't do something about increasing

battery life, we're toast. The biggest impediment to our technological future isn't extending Moore's law. No, the biggest challenge to progress is much more ordinary: It's battery life."[1]

Micro- and nanosystems require even smaller power sources, as weight and volume of onboard energy sources are disproportionately large compared with the systems they power. The roles of energy storage and energy dissipation in microsystems differ considerably from the world of practical daily experience. Designing micro- and nanosystems demands close examination of energy budgets and taking advantage of the merits of smallness as well as minimizing its adverse effects. In Figure 9.1 we compare progress made in improving processors, hard disk drive (HDD) capacity, and battery energy stored. Compared with the other technologies, progress in battery improvement is obviously much more sluggish, with a growth of, at most, 2–3% per year: there is obviously no Moore's law for batteries.

The section on power in MEMS and NEMS should help the reader in choosing a preferred power source for his or her MEMS/NEMS device.

We also consider the implementation of computing power or "brains" in miniaturized systems. After a short summary of current and projected computer technology and its comparison with nature's computers, we consider the latest developments in a field closely related and important to microrobotics: artificial intelligence (AI). We will discuss how both bottom-up and top-down approaches to AI might impact miniaturization technology. Complexity theory and artificial life are examples of bottom-up approaches to building AI that feature the self-organizing or emergent properties touched upon in Chapter 3. Artificial neural networks represent a bottom-up approach to developing brains heavily dependent on learning cycles. Electronic noses and tongues have already been developed on the basis of such neural nets (see Chapter 10). As in the case of machining, we will learn that bottom-up approaches for embedding intelligence in miniature devices are gaining more and more interest.

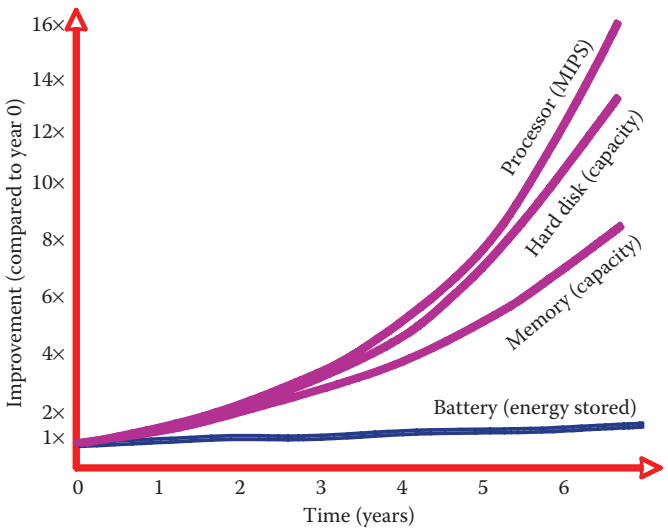


FIGURE 9.1 Progress in improvements made in processors, hard disk capacity, and battery energy stored. No Moore's law in batteries: 2–3% growth per year. MIPS, million instructions per second.

Power in MEMS and NEMS

Onboard Energy Sources

Introduction

Onboard energy sources for MEMS devices may include radioactive sources, batteries, capacitors, fuel cells, internal combustion engines, solar cells, etc. Combustion processes for electrical power generation provide advantages over onboard batteries, even when the low conversion efficiency in the combustion process from thermal energy to electrical energy is taken into account. For example, hydrocarbon fuels provide an energy storage density between 40 and 50 MJ/kg, whereas modern nickel metal hydride batteries, commonly used in laptop computers, provide only 0.4 MJ/kg. Thus, even at only 4% conversion efficiency from thermal to electrical energy, hydrocarbon fuels provide about five times higher energy storage density than batteries.

There has been a lot of attention given recently to MEMS and even NEMS power sources. The goal is to put as much energy as possible in the smallest amount of space. To form an idea of what is achievable with known power systems, we list the energy one can store in 1 mm^3 of various power systems in Table 9.1. From this table, the highest energy density available is from a radioactive source, but such sources may never be practical for consumer applications because of environmental and safety concerns (although americium is used as the current source

Table 9.1 Energy Stored in 1 mm^3

Power capacitor	4 μJ/mm^3	1 μW for 4 s
Thick-film battery	1 J/mm^3	270 μW for 1 h
Thin-film battery	2.5 J/mm^3	0.7 mW for l h
Solar cell (1 × 1 × 0.1 mm^3)		0.1 mW
Gasoline	300 J/mm^3	3 mW for 1 day
180 Ta ($T_{1/2}$ = 8 h)	≈1 MJ/mm^3	34 W
178 Hf ($T_{1/2}$ = 31 years)	>10 MJ/mm^3	160 mW

in some fire alarms). Chemical energy in the form of batteries provides the next highest density. Solar light, at noon on a sunny day, gives us 100 mW/cm^2 (office lights are 7.2 mW/cm^2). Collectors of solar light work with varying efficiency; single-crystal Si provides a 14–30% efficient conversion of light to power. Polysilicon provides about 10–15% efficiency, and organic dyes are 5–10% efficient.

Future stand-alone microsystems will almost surely employ hybrid strategies, using low levels of power for low-level activities and quick bursts of energy, perhaps relying on a capacitor, for computing, communication, or maneuvering.

MEMS and NEMS Batteries, Capacitors, Fuel Cells, Combustion Engines, and Solar Cells

Batteries

To judge the benefits and problems associated with different battery technologies, some general properties of batteries must be introduced. It is usually desirable that the amount of energy stored in a given mass or volume of active battery material is as high as possible. To compare the energy content of batteries, the terms *specific energy* (expressed in Wh/kg) and *energy density* (in Wh/L) are used, whereas the rate capability is expressed as *specific power* (in W/kg) and *power density* (in W/L). To reach the goal of a high specific energy and energy density, two fundamental requirements must be met by the battery electrode materials: 1) a high specific charge (in Ah/kg) and charge density (in Ah/L), i.e., a high number of available charge carriers per mass and volume unit of the material; and 2) a high (positive electrode) and low (negative electrode) standard redox potential of the respective electrode redox reactions, leading to a high cell voltage (see also Volume I, Chapter 7). Moreover, in rechargeable batteries (also called *reversible* or *secondary batteries*), reactions at both negative and positive electrode have to be highly reversible to maintain the specific charge for hundreds of charge/discharge cycles. With an assumed cycling efficiency of even 99% in each cycle, the available specific charge after 100 cycles would be only ~37% of the initial value (after 500 cycles, less than 1%), and electrode materials should thus combine high specific charge with good rechargeability.

In Table 9.2 we list primary and secondary batteries for onboard power systems, and for comparison we also included data on microfuel cells and

Table 9.2 Examples of Onboard Power Systems

On Chip			
Power Sources	**Energy Density (Wh/kg)**	**Wh/L**	**Power Density (W/kg)**
Primary batteries			
Zinc air	200	1050–1560	150
Zn-Cl_2	150	100–180	90
Alkaline	125–225	150–440	
Silver Oxide	155–285	250–500	
Lithium	100–200	340–500	>200
Secondary batteries			
Lead-acid	30–40	40	30–40
Ni-Cd	200	170	150–200
NiMH	150–200	250	150
Li–ion	100	300	200
Microfuel cell		4000	
Ultracapacitors		13–28	200

ultracapacitors. The specific energy (energy per volume unit) of the onboard power sources listed in Table 9.2, determines the proper active volume for a given application. If the volume of the packaging is taken into account, energy volume densities of primary lithium batteries may reach 340–500 Wh/L; these batteries generate the highest specific energy of any commercially available battery. If we assume that the average power consumption of a MEMS device is 100 mW, we need slightly more than 1 cm^3 of lithium battery to last for 1 year, assuming we can use 100% of the charge in the battery. The energy density of rechargeable batteries is typically less than half of that of primary batteries, so in that case we need to be recharging every 3–4 months. Today, commercially available batteries are obviously too big for most MEMS devices, and they are even too big for many mobile consumer applications.

An important technical challenge that arises in considering downscaling batteries, besides the obvious fact that their capacity scales as l^3, is that battery packaging efficiency scales poorly in the microdomain: there is relatively more packaging in smaller batteries than in large ones. What is needed, then, in the microbattery arena is a set of new packaging strategies, new materials, and new manufacturing techniques. In order to be able to make an onboard micropower source using MEMS technology, one also expects that the materials involved are compatible with microfabrication processes (e.g., being stable under vacuum conditions; no liquid electrolytes), that the battery is rechargeable (because its life cycle will be short), and that the reaction products are compatible with the MEMS device.

In terms of commercially available batteries for MEMS, Li primary and Li-ion reversible batteries are the two most important choices today. In lithium primary batteries, the anode, consisting of high-purity lithium, may be combined with many different cathode materials, resulting in different voltages ranging from 1.5 to 3.9 V. For electrolyte, batteries contain organic solvents in which lithium salts are dissolved or they feature solid polymers that also may function as electrolytes. Beyond the button Li batteries 4.8 mm (diameter) by 1.4 mm (high) used in watches and cameras, progress in miniaturizing high-energy-density power sources has been limited. Button Li cells are now the best available primary energy sources for microsystems. The basic principle of rechargeable Li-ion batteries is electrochemical intercalation and de-intercalation of lithium in both electrodes. As we noted above, an ideal secondary or reversible battery has a high-energy capacity, fast charging time, and long cycle life. The capacity is determined by the lithium saturation concentration of the electrode materials. Many materials combination have been researched, and an often touted approach is the use of carbon nanotubes (CNTs). The highest Li concentration can be stored in CNTs if all the interstitial sites (inter-shell van der Waals spaces, inter-tube channels and inner cores) are accessible for Li intercalation. Nalimova et al. proved that two Li atoms per C atom can be stored in multiwall carbon nanotubes (MWCNTs) under a high pressure of 6 GPa.[2,3] This is a very high Li loading, but it involves a chemical reaction, not an intercalation, and its usefulness for a battery is not established. Gaucher et al. reported high degrees of irreversible (500 mAh/g) and reversible (450 mAh/g) insertion of lithium ions in MWCNTs in an aprotic medium.[4] They noticed that the structure of the MWCNTs plays an important role in specific capacity and cycle stability. The capacity of less-graphitized MWCNT was higher than that of the well-graphitized MWCNT, although the latter exhibited better cycle stability and charge/discharge rates.[4] In spite of such interesting behaviors of CNTs as the host of lithium intercalation, their large irreversibility capacity and their poor discharge curves relative to those of lithium metal leave the material not satisfactory yet for commercial applications.

Batteries and fuel cell materials deposited with integrated circuit (IC) technologies on the device substrate itself remain in the research stage.[5] Often, the thin-film materials deposited in constructing those batteries, such as Li, TiS_2, V_2O_5, etc., prove incompatible with the IC process, and the prospect of integrating them with ICs remains remote. Ultrathin, solid-state Li cells—"energy paper"—are beginning to emerge. Kanebo introduced the polymeric polyacenic semiconductor (PAS)-based battery as far back as 1993. The polymer PAS film in the battery is only 200 μm thick and has an active surface area of 2200 m^2/g. It serves as the anode, and the cathode is

lithium based. The voltage is 3.3 V, corresponding to 3 Ni-Cd elements in series. Unfortunately, the energy density, taking the complete, packaged battery into account, is only 5.5 Wh/L.[6]

If a MEMS device uses 10 μW of power continuously, in a day it will consume ~1 J of energy. Since the energy density for Li-ion thin-film systems is in the range of ~2 J mm^{-3}, matching or exceeding standard commercial lithium-ion systems, it would appear that this battery fits the task. In reality, this is not so: thin-film batteries, as shown in Figure 9.2, are two-dimensional (2D) devices and are necessarily thin in order to prevent power losses typically associated with the slow transport of ions, and this form factor leads to a low capacity per mm^2. The energy per unit area as reported for several lithium thin-film batteries ranges from 0.25 to ~2 × 10^{-2} Jmm^{-2};[7] consequently, with a typical device thickness on the order of 15 μm, a thin-film battery would require a significant amount of area in order to supply 1 Jmm^{-3}.

The real question then becomes whether a thin-film battery can supply the energy that the MEMS device requires within the constraints of the areal "footprint" available for the battery. Unfortunately, a thin-film battery of 1 mm^2, despite having an excellent energy per unit volume, falls far short of being able to power a typical MEMS for more than one day. If the areal footprint is made 100 times larger (at 1 cm^2), the thin-film approach becomes more acceptable,[7,8] These important consequences of the 2D nature of thin-film batteries are frequently overlooked. The calculation by Koeneman et al.,[9] for example, ignored the 2D character of thin-film batteries when they concluded that these batteries could carry out some 60,000 actuations of a "smart bearing."[7] When the actual area available for the power source on the device is considered, only about 1200 actuations are possible. In order to power devices with limited real estate and maintain a small areal footprint, batteries must somehow make better use of their thickness dimension.[7]

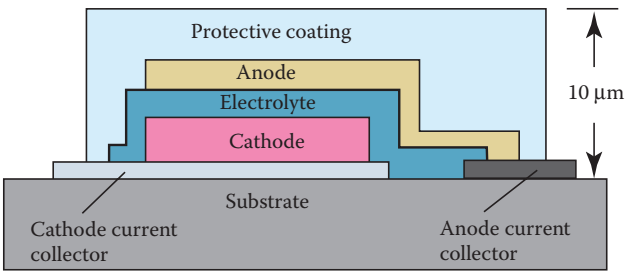


FIGURE 9.2 Example thin-film battery.

In recent years there has been a growing realization that improved microbattery performance can be attained by configuring the electrode materials currently employed in 2D batteries into micro-three-dimensional (μ-3D) architectures (high-rise batteries). In this approach, cell structures are designed to maximize power and energy density while maintaining short ion transport distances.[7] There are several possible architectures that can achieve this same goal, but an underlying principle in all μ-3D batteries is that the transport between the electrodes remains one-dimensional at the microscopic level, while the electrodes are configured in complex nonplanar geometries in order to increase the energy density of the cell within the same footprint area.[7,10]

Detailed calculations by Long et al.[7] show that microbatteries configured as μ-3D structures can readily achieve 1 Jmm^{-2}, the amount required to power a typical MEMS device for a day. Perhaps the most obvious μ-3D design is one consisting of interdigitated electrode posts, shown in Figure 9.3a. The anode and cathode consist of arrays of rods separated by a continuous electrolyte phase. The spatial arrangement of the anode and cathode arrays determines the current-potential distribution. The short transport distances lead to a much lower ohmic resistance as compared with traditional planar battery configurations. The molecular diffusion equation indicates that diffusion over a 10-μm distance is 1 million-fold faster than diffusion over a 1-cm distance, since $x = (2D\tau)^{1/2}$ (Volume I, Equation 6.100), where D is the diffusion constant and τ the time required for a molecule to diffuse over distance x. Thus, μ-3D configurations with narrow electrode posts offer a means of maintaining "short" diffusion distances, thereby increasing the power capacity of the battery. For example, Hart et al. first consider a thin-film 2D battery that comprises a 1-cm^2 area for both anode and cathode, each 22.5 μm thick and separated by a 5-μm-thick electrolyte.[10] The total volume of electrodes and electrolyte in this case is 5 × 10^{-3} cm^3. It is relatively straightforward

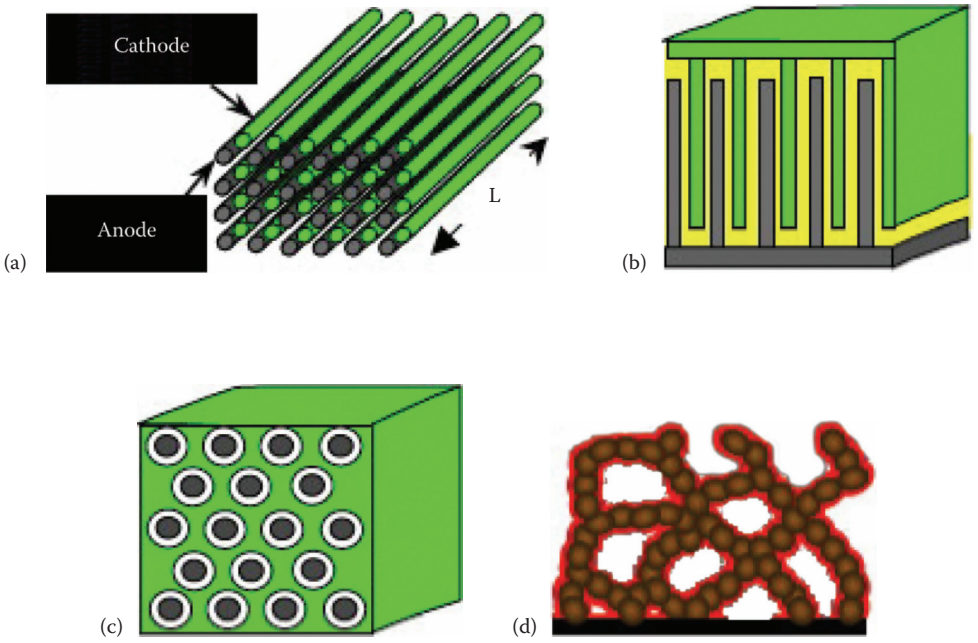


FIGURE 9.3 Examples of prospective three-dimensional architectures for charge-insertion batteries. (a) Array of interdigitated cylindrical cathodes and anodes; (b) Interdigitated plate array of cathodes and anodes; (c) Rod array of cylindrical anodes coated with a thin layer of ion-conducting dielectric (electrolyte), with the remaining free volume filled with the cathode material; (d) Aperiodic "sponge" architectures in which the solid network of the sponge serves as the charge-insertion cathode, which is coated with an ultrathin layer of ion-conducting dielectric (electrolyte), and in which the remaining free volume is filled with an interpenetrating, continuous anode.[7]

to calculate then that a corresponding μ-3D array battery (see Figure 9.3) with a series of 5-μm-radius interdigitated cathodes and anodes, a 5-μm surface-to-surface electrode spacing, and occupying the same total volume (i.e., 5×10^{-3} cm^3) contains only about 39% of the energy capacity of the thin-film 2D design. The lower-energy capacity is due to a higher percentage of the total volume being occupied by the electrolyte. On the other hand, the active cathode and anode surface areas in the three-dimensional (3D) design are 3.5 cm^2 each, significantly larger than the 2D design. There are some other intriguing advantages of the μ-3D design that are not directly reflected in the above numbers. For instance, the transport length scale in the thin-film 2D battery is 350% larger than in the 3D design. This means that the μ-3D design is significantly less susceptible to ohmic losses and other transport limitations. To achieve equal transport length scales in the 2D design would require a 350% increase in the areal footprint in order to maintain equal cell volume, a significant disadvantage in employing these devices in MEMS and microelectronic applications. While the above comparison of 2D and μ-3D designs indicates that the μ-3D cell has an inherently lower-energy capacity per total cell volume, in fact, the capacity of the 3D design can be increased without limit by increasing the battery height, without sacrificing the small area footprint or high-power density.

For an example of this type of μ-3D electrode configuration, see the carbon-MEMS (C-MEMS) battery in Figure 5.18. This type of battery is being developed by Enevate (http://www.enevate.com/).

Capacitors

Rechargeable and disposable (primary) batteries, reviewed above, use a chemical reaction to produce energy. The problem is that after many charges and discharges, even a rechargeable (secondary) battery loses capacity, to the point where the user has to discard it. However, capacitors contain energy stored in an electric field between charges on two metal electrodes separated by a dielectric. Capacitors charge faster and last longer than normal batteries, and their reversibility, absence of polarity, and extended lifetime make them an attractive alternative for power in microsystems. They are useful for on-chip power conversion, but their energy density

is too low to be a real secondary storage component. Electrochemical capacitors improve that perspective; with an energy density on the order of 75 J/cm^3, they are a good potential for secondary storage.

Electrochemical capacitors are called by a number of names: *supercapacitors*, *ultracapacitors*, and *electrochemical double-layer* (ECDL) *capacitors*. Typically, supercapacitor devices consist of two electrodes and an insulating material that separates the electrodes. The insulator prevent electrical contact but allows ions from the electrolyte to pass through. At each electrode, there exists a dipole layer, the so-called *ECDL*. Typically ~0.4 nm thick, the ECDL acts as a capacitor if the potential difference across it is less than that needed to dissociate the electrolyte (into hydrogen or oxygen in the case of water). If the electrolyte is water, this restricts the voltage to 1.23 V, but higher voltages of 2.3 V are possible by using a nonaqueous electrolyte, such as tetraethyl ammonium fluoroborate in propylene carbonate. The capacitance is given by $C = \varepsilon A/d$, where ε is the dielectric constant, A is the electrode surface area, and d is the ECDL thickness. The electrodes are made of materials with a high effective surface area, such as porous metal oxides (e.g., IrO_x), porous carbon, or carbon aerogels, in order to maximize the surface area, A, of the double layer and to achieve a small double layer thickness, d. By choosing and fabricating electrodes with extremely high surface area A (e.g., 1000 m^2/g or greater) and a small d, very large capacitances are achievable (e.g., 1000 F). Carbon nanotube filaments can further increase the surface area of the electrodes and allow capacitors to store yet more energy. Single-walled carbon nanotubes (SWNTs) have the largest surface area to volume ratio of any carbon material, 3000 m^2/g, as all their atoms are on the surface, and they make for the ultimate electrode material. This advantage is used for energy storage in nanotube supercapacitors.[7]

The performance of supercapacitors is judged in terms of their energy density (per unit volume) or $\frac{1}{2}CV^2$ and power density or V^2/R, where R is their series discharge resistance. Another advantage of CNTs in this regard is that they have a much lower resistance than the often-used activated carbon, which could greatly increase their power density. The limiting factors are contacting CNTs to the electrode backing and having the ability to do this at low temperatures and at a low cost, as the competing activated carbon is very cheap.

Since it is possible to carbonize photoresist materials, make them porous, and charge them, it is possible to integrate ultracapacitors on ICs (see C-MEMS in Chapter 5). The most challenging technical issue, just as in the case of a battery, is packaging. Supercapacitors and batteries incorporate very corrosive and reactive materials, making the challenge even more daunting.

The performance of supercapacitors can be gleaned from the *Ragone plot*, as shown in Figure 9.4. A Ragone plot is a log-log plot of the specific power (W/kg) versus the specific energy (Wh/kg).

As can be seen from this plot, supercapacitors or ultracapacitors bridge the gap between capacitors and batteries and have the ability to store greater amounts of energy than conventional capacitors, and they are able to deliver more power than batteries. Supercapacitors also possess a number of desirable qualities that make them an attractive energy storage option: they store and release charge completely reversibly, they are extremely efficient and can withstand a large number of charge/discharge cycles, they can store or release energy very quickly, and they can operate over a wide range of temperatures.

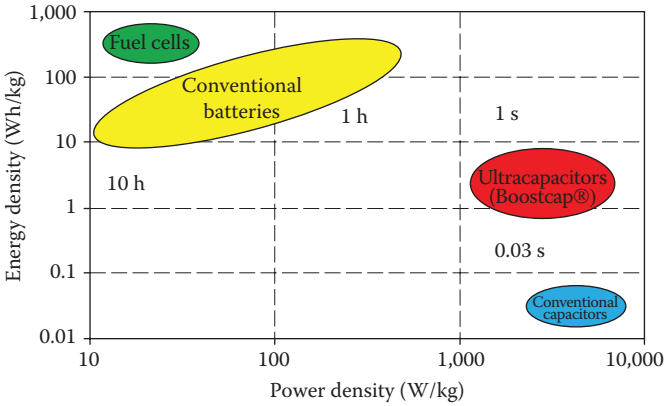


FIGURE 9.4 Ragone plot. The power density versus energy density plot distinguishes various electricity storage systems. Batteries are high-energy density but low-power density, whereas normal capacitors are low-energy density but high-power density. Li^+ batteries have the highest energy density of any battery. Supercapacitors have a high-energy density and high-power density and are intermediate between normal capacitors and batteries. See http://www.tecategroup.com/ultracapacitors/maxwellboostcaps.php.

Fuel Cells

Introduction

Two major challenges of our times are the shrinking of available fossil energy resources and climate change associated with global warming. With humans consuming carbon-based fuels (carbon, natural gas, coal, and tar shale) faster than they can be replaced, yet another "oil war" making the daily news, GM still selling 15–20,000 Hummers a year, and Tata planning to sell millions of its new Nano cars, independence from fossil fuels has finally again become what former President Jimmy Carter called the "moral equivalent of war."

One of the often advanced solutions to the energy crisis is to replace fossil hydrocarbon-based energy with the energy from carbon-free sources such as the sun, nuclear energy, or the hot interior of the earth and use hydrogen as the energy carrier. Hydrogen can be produced by hydrolysis of water using energy from carbon-free sources, and the hydrogen thus produced can serve as fuel in fuel cells to generate electricity, either stationary or mobile.

Fuel cells, discovered in the 1850s, are batteries with renewable energetic materials (fuels) (see inset in the heading of Part II and Figure 9.5). The electrodes in a fuel cell do not take part in the cell reactions they act as reaction sites only. The basic reaction in a fuel cell is reversed electrolysis, or:

$$2H_2 + O_2 = 2H_2O \qquad \text{Reaction 9.1}$$

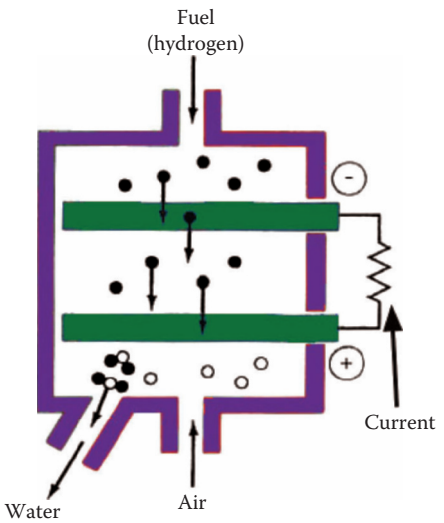


FIGURE 9.5 Schematic of a fuel cell.

Electricity is produced in an anode reaction:

$$H_2 = 2H^+ + 2e^- \qquad \text{Reaction 9.2}$$

and a cathode reaction:

$$\tfrac{1}{2}\, O_2 + 2H^+ + 2e^- = H_2O \qquad \text{Reaction 9.3}$$

Fuel cells can help us generate cleaner power from conventional sources more efficiently and can be viewed as a temporary energy solution until we can harvest more renewable resources (solar, wind) and/or or more powerful sources (fusion) directly.

Differences in electrolytes used determine the type of fuel cell:

- Phosphoric acid fuel cell (PAFC)
- Proton exchange membrane fuel cell (PEM, SPE™, and PEFC)
- Solid oxide fuel cell (SOFC)
- Molten carbonate fuel cell (MCFC)
- Alkaline fuel cell (AFC)

There are quite a few positive fuel cell attributes to enumerate. Power and energy density of a fuel cell are higher than in batteries [e.g., methanol (1550 Wh/L, 1738 Wh/kg) versus lithium-ion (450 Wh/L, 185 Wh/kg); see also the Ragone plot in Figure 9.4]. Fuels used in fuel cells typically have much higher (approximately 10×) energy densities than their battery counterparts, and to increase the power density of a fuel cell, one only needs to increase the surface-to-volume ratio within the cell. The latter is a much simpler task than engineering new material chemistries for batteries. There is no lengthy recharging, as fuel cells are rapidly refueled; instead of charging a battery for an extended amount of time, a fuel cartridge is inserted. The energy conversion process in fuel cells is simpler, and there is an unlimited access to active electrochemical material (fuel), while the amount of active materials accessible for energy conversion in batteries is limited to their volumetric energy density. Fuel cells feature negligible self-discharge, are clean and environment-friendly, and have an unlimited cycle stability. Costs of fuel cells are coming down: through sophisticated engineering the amount of expensive platinum required for catalysts is being lowered, and new manufacturing

processes are being introduced. While the cost per Joule ($/J) has gone up for energy-dense batteries (e.g., NiCd weighs 0.5 kg, lasts 1 h, and costs $20, and a comparable Li-ion battery lasts 3 h but costs more than four times more), the cost of a comparable fuel cell that lasts 30 h has gone down (to perhaps <$5).

Disadvantages of fuel cells compared with batteries are lower efficiency (up to 50–60%, compared with about 90% for Li-ion batteries) and the need for peripherals such as pumps, humidifiers, fuel processors, power conditioners, etc. These peripherals have to do with the need to obtain hydrogen from traditional fuels (oil, coal, wood, propane, natural gas, ethanol, etc.), to store fuels and hydrogen (the storage medium), and to manage heat and water (both products of the fuel cell reaction) and power (e.g., convert DC to AC). A typical process to obtain hydrogen is steam reforming, which can process methane into hydrogen (research and development is under way to also reform diesel, gasoline, ethanol, etc.). The basic steam reforming reaction is as follows:

$$CH_4 + 2H_2O = 3H_2 + CO_2 \qquad \text{Reaction 9.4}$$

Intermediate steps in this reaction produce CO, which is reacted to form CO_2. The reaction also requires catalysts to speed up reaction rates, and sulfur compounds must be removed to protect those catalysts.

In Table 9.3 we present a summary of stationary fuel cells (data are from 2003). Fuel cells are still

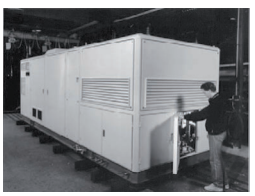
Table 9.3 Summary of Stationary Fuel Cell Products

200 KW PAFC – UTC Fuel cells

250 KW SOFC – Siemens

Type	Electrolyte	Electrode	Fuel processor	Temp (°C)	Electrical Efficiency (LHV)	Cogeneration potential	Stage of development	Manufacturer
PAFC	Phosphoric acid (liquid/gel)	Noble metal	External	200	35-40%	Hot water Steam (low psi)	Commercial sales 250 units in field	UTC fuel cells
SOFC	Zinconium oxide (solid)	Ceramic	Internal	1000	50-65%	Hot gas Steam optional	Field demonstrations Several units in field	Siemens Westinghouse
MCFC	Alkali carbonates (liquid)	Nickel	Internal	650	45-60%	Hot gas Steam optional	Field demonstrations Expect to have precommercial units in place within one year	Fuel Cell Energy, Inc.
PEMFC	Organic polymer (solid)	Noble metal	External	80	40-45%	NA	Field demonstrations Commercial sales of small residential units	Avista Ballard H Power UTC Fuel cell Nuvera Plug Power

200 KW MCFC – Fuel Cell Energy, Inc.



250 KW PEM – Ballard

too expensive, and in order to introduce them into the huge mobile market, their miniaturization and the integration/simplification of the accompanying peripheral gear is a must (see next section).

Microfuel Cells

Introduction

There are many reasons why one might want to make a miniature fuel cell. To begin with, small fuel cells, especially when they come with less or size-reduced peripheral equipment, will make many more applications possible (e.g., power for mobile phones, iPods, laptops, etc.). The comparison between 2D and μ-3D batteries carried out earlier (see Batteries) can be extended to 2D and μ-3D fuel cells. From simple theoretical considerations, for a specific μ-3D electrode designs, we calculated a 3.5-fold larger surface area than the corresponding 2D design. The surface area of a fuel cell determines its amperage, meaning that fuel cell power is directly proportional to the electrode surface area. Smaller cells have larger surface areas relative to their volume, so they automatically feature increased power density. Microfuel cells might eventually also come at a lower cost because less material is used in their construction. Because of function integration possible in microsystems, the overall complexity may also be reduced (less peripherals), further reducing total weight and volume. Moreover, microfuel cells are expected to have a higher efficiency because of their high surface-to-volume ratio and the corresponding increase in triple points (reaction points where gas, liquid, and electrode meet) and the smaller fluidic channel sizes, which allow more efficient mass transport (for serpentine flow patterns, the flow velocity is inversely proportional to the square of the length scale). Employing microfluidic strategies, gas transport in the cell can be maximized while gas stoichiometry can be maintained and "dead zones" underneath structures minimized. Microfuel cells also may benefit from increased catalyst utilization, because microfabrication and nanostructuring techniques (see e.g., carbon nanohorns, Figure 9.8) allow for more control over catalyst deposition and distribution. Finally, it is easier to maintain a homogeneous environment within a small area, and the internal resistance is lower because of shorter conductive paths.

Example 9.1:

There have been many attempts at miniaturization of fuel cells. A recent example comes from Toshiba, which unveiled a miniature direct methanol fuel cell (DMFC) developed for small electronic devices such as digital audio players and wireless headsets (Figure 9.6). The fuel cell, measures 22 × 56 × 4.5 mm, weighs 8.5 g, and can put out 100 mW of power for as long as 20 h with a 2 cc charge of 95% methanol. The air-breathing cell has only passive components.

DMFCs are a subcategory of polymer exchange membrane fuel cells (PEMFCs) where the fuel, methanol, is not reformed but is fed directly into the fuel cell, avoiding complicated catalytic reforming steps. When a DMFC is operating, the methanol solution and oxygen gas are fed to the surface of the membrane electrode assembly (MEA) through the anode and cathode plates, respectively. Methanol is consumed at the anode catalyst layer, producing protons and electrons and releasing the byproduct CO_2. The protons directly transfer through the PEM to the cathode, while the electrons, blocked by the insulating PEM, have to travel to the cathode via the external circuit, resulting in an electrical current. The protons and electrons combine with O_2 at the cathode catalyst layer to form H_2O. Storage of methanol is easier than that of hydrogen because it does not need to be done at high pressures or low temperatures, since methanol is a liquid. Liquid methanol, as

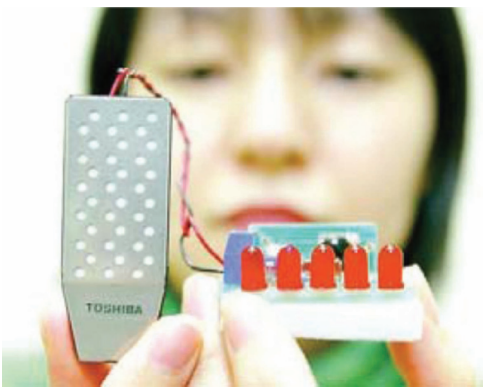


FIGURE 9.6 Toshiba's miniature direct methanol fuel cell.

the fuel of DMFC, has the advantages of easy storage and the ability to refuel instantly with no interruption of operation. The methanol cell can be directly started and operated at room temperature, which is a significant advantage for the applications of the portable electronic devices. Finally, the energy density of methanol is an order of magnitude greater than that of even highly compressed hydrogen. Methanol has an energy density of 17.6 kJ/cm^3, which is about 6 times that of a lithium battery.

Unfortunately, the efficiency of DMFCs is typically low because of the high permeation of methanol through the polymer membrane, a deleterious effect known as methanol crossover.

Fuels and Fuel Storage in Microfuel Cells

Introduction

The advent of microfuel cell-powered gadgets will require progress in the manufacture of fuel storage cartridges (see Figure 9.7).

Fuel sources for fuel cells range from highly flammable to relatively inert. Many miniaturized microfuel cells use methanol to be oxidized directly in DMFCs (see the Toshiba cell described above) or they consume hydrogen that is reformed "on demand" from methanol in reformed methanol fuel cells (RMFCs). Formic acid at a concentration of <85% by weight may be oxidized directly in a formic acid fuel cell. Liquid borohydride is either oxidized directly, in direct borohydride fuel cells (DBFCs), or reformed to hydrogen on demand, in indirect borohydride fuel cells. A butane or a butane/propane mix is oxidized directly by an SOFC system.

FIGURE 9.7 Fuel cell cartridges in consumer electronics.

Using methanol, Samsung has developed a prototype fuel cell for use in laptops that it claims lasts almost twice as long as other fuel cell systems being readied (see Figure 9.7). The Samsung fuel cell, powered by about 200 cm^3 of liquid methanol, has an energy density of 200 Wh/L. The cell measures 23 cm wide, 8.2 cm in length, and 5.3 cm deep and weighs less than 1 kg. In total, the fuel cell can supply power to a laptop for about 15 h. Such performance substantially improves over the 100–130 Wh/L energy densities of the fuel cells being developed by companies such as Toshiba (see Figure 9.6) and NEC for laptops. While hydrogen-based cells have taken off for home or automobile use, versions based on methanol for use in electronics products have yet to be commercialized. Toshiba and NEC are among the companies that promised methanol fuel cell-based laptops in previous years, but each time, technology launches have been delayed.

In what follows, we analyze in a bit more detail progress made in hydrogen storage using CNTs and carbon nanohorns and the prospect of biomass as fuels in biofuel cells.

Hydrogen Storage

The chemical energy per mass of hydrogen (142 MJ kg^{-1}) is at least three times larger than that of other chemical fuels (for example, the equivalent value for liquid hydrocarbons is 47 MJ kg^{-1}). Also, hydrogen as a fuel leads to pure water as a byproduct, and the hydrogen/oxygen reaction is a reaction for which catalysts have been researched the most. The problem of storing hydrogen safely and effectively is one of the major technological barriers currently preventing the widespread adoption of hydrogen as an energy carrier and the subsequent transition to a so-called "hydrogen economy." Practical problems with storage of hydrogen gas or liquid make solid state storage the

most attractive. To appreciate the challenge, consider that 4 kg of hydrogen gas, which is required for a practical driving distance, occupies a volume as large as 49 m^3. Storage solutions must reduce the enormous volume of the hydrogen gas. A solid medium capable of absorbing and releasing large quantities of hydrogen easily and reliably is actively being sought and various types of solids have been proposed, including microporous materials that can store physisorbed molecular hydrogen at low temperatures,[11] intermetallic hydrides that absorb atomic hydrogen as an interstitial,[12] and complex hydrides formed from light elements, such as Li, B, Na, Mg, and Al.[13,14] GE, for example, is working with the Department of Energy (DOE) on $Li_2Mg(NH)_2$, which contains 5.6% hydrogen and has reversible storage capability. As a threshold for economical hydrogen storage, the DOE has set storage requirements for the DOE Freedom CAR program at 6 weight % (wt%) to be considered for technological implementation.

Among the storage methods currently being explored, storing hydrogen by adsorption in metals appears too expensive. However, trapping it in porous materials is not only efficient (all the hydrogen adsorbed can be recovered) but cheap. What's more, the cycle of hydrogen storage and release does not require any reactivation or regeneration of the material. Accordingly, considerable research efforts are going into the evaluation of various nanostructures, such as CNTs, to find the most suitable lightweight hydrogen storage material. Because of their cylindrical and hollow geometry and nanometer-scale diameters, it was predicted that CNTs would store liquids or gases in their inner cores. Claims of storage of hydrogen in CNTs attracted worldwide interest when Dillon et al. published 5–10 wt% hydrogen storage in single-walled carbon nanotubes (SWNTs).[15] Subsequently, claims of high hydrogen storage capacities at room temperature and pressures higher than 100 bar in graphitic nanofibers (GNFs) and lithium- and potassium-doped multiwalled carbon nanotubes (MWCNTs) were made. But in the late 1990s, the identification of many erroneous measurements of gaseous hydrogen uptake by CNTs and CNFs led to a controversy that is still lingering today.*[16] In light of today's knowledge, it is not very likely that at moderate pressures and around room temperature carbon nanostructures can store the amount of hydrogen required for hydrogen-based fuel cells (at least not for fuel cells for cars): using them for storage is possible only at extremely low temperatures (below –196°C) because of the weak interaction between hydrogen and carbon. The future possibility of storing hydrogen inside carbon-based porous materials, as part of a clean energy scheme, depends on the force of the interaction between hydrogen and carbon and on how easily this force can be increased.

Another possibility for hydrogen storage is electrochemical storage in CNTs. In this case, it is not a hydrogen molecule but a hydrogen atom that is physisorbed. Currently, the storage of hydrogen in this absorbed form in carbonaceous materials is considered the most appropriate route to solve the hydrogen storage problem. While most of the studies have focused on the hydrogen storage through physisorption, recent density functional theory (DFT) calculations for single-walled CNTs indicate the potential for up to 7.5 wt% hydrogen storage capacity through chemisorption by saturating the C-C double bonds in the nanotube walls and forming C-H bonds (i.e., hydrogenation; see the articles "Generalized Chemical Reactivity of Curved Surfaces: Carbon Nanotubes"[17] and "Theoretical evaluation of hydrogen storage capacity in pure carbon nanostructures"[18]). However, direct experimental evidence of the high values of the hydrogen capacity through chemisorption has not yet been demonstrated, and so far, the only storage systems that have met the DOE targets are compressed tanks of gas and liquid. A 2004 report by the American Physical Society concluded that even the most promising solid-storage technologies are still several breakthroughs away from practical use.

* Some of the physical properties of hydrogen, including its low molar mass relative to gas phase impurities and its susceptibility to leakage and permeation, together with the sensitivity of the materials that can be used to store it, mean that hydrogen sorption measurements are particularly prone to error.

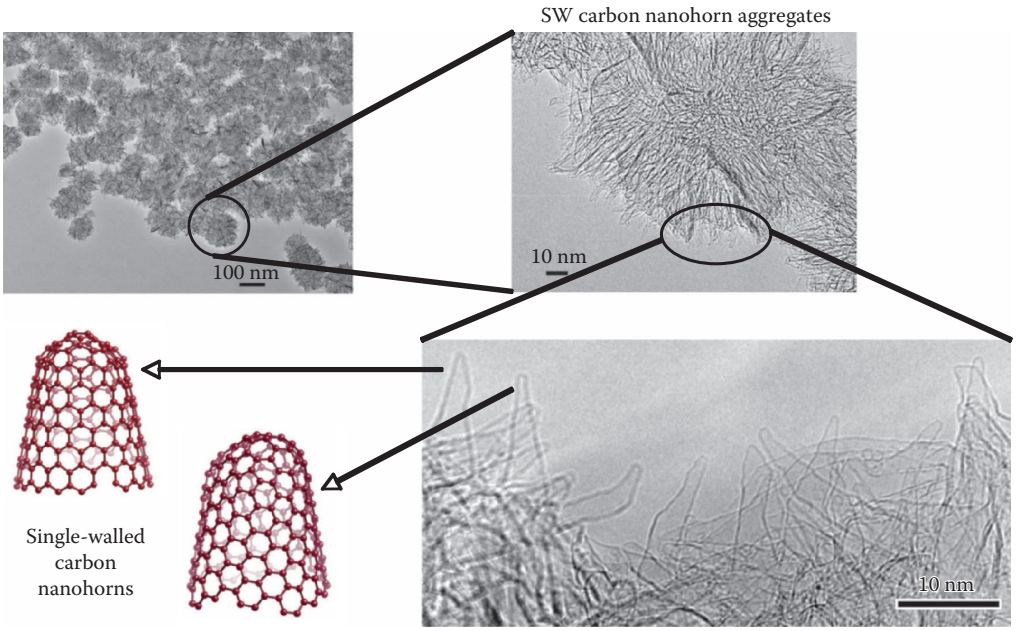


FIGURE 9.8 Carbon nanohorn aggregates and individual carbon nanohorns. (From Iijima, S., et al. 1999. *Chem Phys Lett* 309:165.[20])

Carbon nanohorns offer a better prospect for hydrogen storage in fuel cells than CNTs. Nanohorns are pure carbon structures that are 2–3 nm long, are cone-shaped, and aggregate to form dahlia-shaped structures 80–100 nm in diameter with a large number of horn-shaped, short, single-layered nanotubes that stick out in all directions (Figure 9.8). Since the tips of the cones are pointed, it was suspected that there might be a stronger hydrogen-substrate interaction than in CNTs. Recently, researchers at the Centre National de la Recherche Scientifique (CNRS) confirmed that with carbon nanohorns structures, the hydrogen-carbon interaction is indeed far more stable than with nanotubes.[19] This group demonstrated that nanohorns have remarkable adsorptive behavior because of their large surface area (about 400 m^2/g) and excellent catalytic properties because Pt catalyst is dispersed more homogeneously on the surface of carbon nanohorns. Unlike single-walled CNTs, carbon nanohorns can be made without the use of catalysts. Carbon nanohorn aggregates can be produced with a yield of more than 90% through laser vaporization of carbon at room temperature. Sumio Iijima's group showed that they could also be produced by laser ablation of graphite, and these were the researchers who came up with the name *nanohorns*.[20]

Carbon nanohorns are expected to perform excellent as gas storage material, making use of their high adsorbability.* Carbon nanohorns have actually cleared the DOE threshold for commercial adoption as methane gas storage material.

Biofuels

Not everyone agrees that improved hydrogen storage is the way to further develop microfuel cells. Some researchers point out that proponents of a hydrogen economy claim to address greenhouse gas (GHG) emissions, but in reality, these experts counterargue, the most economic methods of making hydrogen is from natural gas and these processes produce significant amounts of GHGs. Since virtually all living things in the biosphere operate by using biomass as fuel and generate power using fuel cells, these scientists propose that a biomimetic approach would appear to be the most environmentally sustainable method of producing energy. The most common energy currencies on the planet, used by virtually every living thing, are found in landfills, wastewater

* Carbon nanohorns selectively adsorb DNA fractions. Inorganic materials are mostly used today in selecting DNA fractions. However, carbon, with its excellent biocompatibility, may be a better material than inorganic substances. See http://www.nanonet.go.jp/english/mailmag/2004/023a.html.

treatment plants, wood waste, agricultural waste, grasses, etc. The mechanism of biofuel cells involves the oxidation of substrates such as carbohydrates, proteins, or fats by microorganisms or enzymes generating electrons in an external circuit. Biocatalyzed oxidation of organic substances by oxygen or other oxidizers in a two-electrode cell provides a potential means for the conversion of chemical to electrical energy, even in the body of a living subject. Like conventional fuel cells (such as methanol fuel cells), these types of cells can deliver more energy per volume weight than conventional batteries, and the theoretical efficiency of biofuel cells is estimated to be as high as 90% (compared with up to 60% for other fuel cells; see above). Abundant organic raw materials, such as amino acids or glucose, can be used as fuels in biofuel cells for the oxidation process, and molecular oxygen can act as the substrate being reduced. If the biofuel cells use concentrated sources of chemical energy, they can be made safe, small, and light. Moreover, biofuel cells differ from the traditional fuel cells by the material used to catalyze the electrochemical reaction. Instead of using precious metals as catalysts, biofuel cells rely on biological molecules, such as enzymes, to carry out the reaction.

Biofuel cells are categorized into microbial-based and enzymatic biofuel cells, which employ biocatalysts, i.e., either microorganisms or enzymes, respectively, to convert chemical energy into electrical energy. Biocatalysts are attractive compared with transition metal catalysts because 1) they operate at moderate temperatures, 2) they allow for the use of inexpensive fuel cell components, 3) they are renewable, and 4) they can be developed for a variety of cheap fuels. Enzymatic biofuel cells are preferred over microbial ones. Microbial-based biofuel cells have too many disadvantages: complex design, big size, poor reproducibility, and short lifetime based on bacterial survival.

As an example of an enzymatic biofuel cell, consider the glucose cell in Figure 9.9. Here, oxidation of glucose is coupled to the reduction of dissolved O_2. The cell incorporates an anode consisting of sugar-digesting enzymes and mediator, and a cathode comprising oxygen-reducing enzymes and mediator. The anode extracts electrons and hydrogen ions

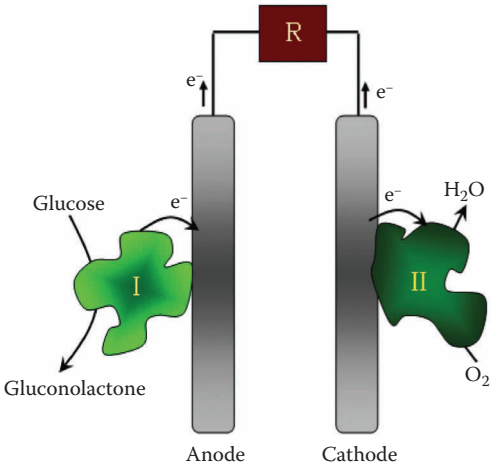


FIGURE 9.9 Biofuel cell based on glucose oxidation and oxygen reduction.

from the sugar (glucose) through enzymatic oxidation as follows.

$$\text{Glucose} \rightarrow \text{Gluconolactone} + 2\,H^+ + 2e^- \qquad \text{Reaction 9.4}$$

The hydrogen ion migrates to the cathode, and once it arrives at the cathode, the hydrogen ions and electrons absorb oxygen from the air to produce water.

$$(1/2)O_2 + 2\,H^+ + 2e^- \rightarrow H_2O \qquad \text{Reaction 9.5}$$

Through this electrochemical reaction, electrons pass through the outer circuit to generate electricity. Power output of this cell is the product of the cell voltage, V_{cell}, and the cell current, I_{cell}. The oxidation reaction may be catalyzed by glucose oxidase (I) at the anode, and oxygen reduction may be catalyzed by laccase (II) (or bilirubin oxidase) at the cathode. Since most of the redox enzymes lack direct electrical communication with electrode supports, various electron mediators have to be used to contact the biocatalyst electrically with the electrodes. A redox mediator of appropriate redox potential is required to shuttle electrons between the protein and the electrode surface, because direct electron transfer to buried redox sites within these proteins is generally not possible given the distance of the active site from the electrode surface (for more details, see Volume I, Chapter 7).

For the pictured glucose/O_2 biofuel cell, a V_{cell} higher than 0.5 V can be achieved by the utilization

of enzymes and redox mediators of carefully chosen redox potentials. The mediators and enzymes can be immobilized on the electrodes by, among others, codeposition of redox polymer/enzyme and a bifunctional linker, coelectrodeposition and enzyme reconstitution at a functionalized electrode surface to provide membrane-less biofuel cells. The immobilization concept represents an important advancement, as it precludes the need to separate the anode and cathode half-cells from each other using a membrane, provided that no solution redox reaction between fuel and oxygen occurs (as is the case for glucose). See http://www.tappi.org/content/events/10nano/papers/20.2.pdf.

Example 9.2:

Sony, one of the world's largest battery makers, has succeeded in creating a sugar-based biofuel cell. The biofuel cell does not require mixing or convection of the glucose solution or air movement; as it is a passive-type cell, it works simply by supplying sugar solution into the unit. Their test fuel cells measure 39 mm^3 and delivers 50 mW—currently the world's highest level for passive-type biofuel cells of comparable volume. The biobattery casing is made of vegetable-based plastic (polylactate) and is designed in the image of a biological cell. By connecting four cubic cells, it is possible to power a memory-type Walkman (NW-E407), together with a pair of passive-type speakers (no external power source) (see Figure 9.10).

FIGURE 9.10 Sony's biofuel cell based on glucose solutions. Four units are shown to power a Sony Walkman. (Courtesy of Sony.)

Key achievements in the development of the Sony biofuel cell include the following:

1. Enhanced immobilization of enzymes and mediator on the anode. For effective glucose digestion, the anode contains high concentration of enzymes and mediator. The Sony immobilization technology involves two polymers to attach enzyme and mediator to the anode. The polymers feature an opposite charge, so the electrostatic interaction between the two polymers effectively secures the enzymes and the mediator closely together. The ionic balance and immobilization process have been optimized for efficient electron extraction from the glucose.
2. Cathode structure for efficient oxygen absorption. The water content within the cathode environment is vital to ensure optimum conditions for the efficient enzymatic reduction of oxygen. The Sony biofuel cell employs porous carbon electrodes partitioned using a cellophane separator. The optimization of this electrode structure ensures that the appropriate water levels are maintained, enhancing the reactivity of the cathode.
3. Optimization of electrolyte. A phosphate buffer of approximately 0.1 M is generally used within enzymology research; however, an unusually high 1.0 M concentration buffer is used in the Sony biofuel cell. This is based on the discovery that such high concentration levels are effective for maintaining the activity of enzymes immobilized on the electrodes.

Sony continues its development of immobilization systems, electrode composition, and other technologies in order to further enhance power output and durability, with the aim of realizing practical applications for these biobatteries in the near future.[21, 22]

Despite the many advances in enzyme and polymer membrane modified electrode techniques, current enzymatic biofuel cells cannot compete with conventional batteries because of their low cell voltages and power densities. In addition, other key challenges of enzymatic biofuel cells include an uncertain storage and operational stability, tedious pretreatments of special and expensive

electrode materials with a subsequent complex chemical, and/or electrochemical functionalization, the adsorption of many interfering species on the electrodes, etc.

Implantable biofuel cells are proposed to convert the energy stored within chemicals present in vivo into usable electrical energy. The microbiofuel cell is potentially an ideal power source for a variety of devices that need to be implanted for short periods of time (i.e., from a few days to a period of a couple of weeks), such as medical sensor/transmitter systems and injectable biosensors, because of its easy integration with the sensor-transmitter system fabricated by MEMS technologies. Current biofuel cells cannot compete with conventional electrochemical cells in powering implantable MEMS devices, such as in situ and in vivo biosensors. Normally these cells would produce continuously a few microwatts, of which less than 1 μW l is required for the operation of the sensor; the transmitter consumes most of the power. With a small ultracapacitor storing ~10 mJ, enough for 1-ms-long 1–10 GHz bursts of 10 μW every 10 s, the transmitted information is easily acquired outside the body. Even though microwave irradiation has been used to charge these devices, if one considers the potential health risks of direct microwave irradiation, the biofuel cell could be more convenient and safe for the patients.

Micromachining of Microfuel Cells

To appreciate the many challenges involved in micromachining a fuel cell, let us start by considering how one is typically constructed (see Figure 9.11). To focus the mind we assume that we are dealing with miniaturization of a passive PEM fuel cell. In general, passive-type fuel systems have a more simple structure and are more suitable for miniaturization, whereas active type systems have a more complicated structure and are suited to higher power devices. A PEMFC contains an electrolyte that is a layer of solid polymer that allows protons to be transmitted from one face to the other. PEMFCs require hydrogen and oxygen (or air) as inputs, and these gases must be humidified in most cases. The operating temperatures are relatively low, at around 60–80°C.

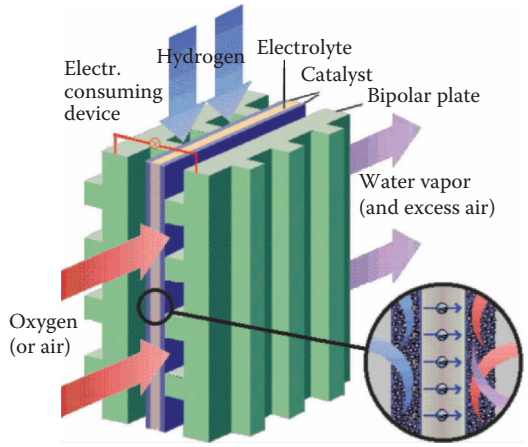


FIGURE 9.11 Construction of a typical fuel cell.

Figure 9.11 shows one fuel cell element only and these elements can be either stacked (Figure 9.12a) or put in an array. A fuel cell array can either be banded or flip-flopped as illustrated in Figure 9.12b.

Bipolar plates have a structural, an electrical, and a fluidic function. The requirements for the bipolar plates in the microfuel cells pictured in Figure 9.11 include high electrical conductivity, resistance to corrosion, chemical compatibility, high thermal conductivity, low leakage of gases, high mechanical robustness, low weight, manufacturability, and low cost. Graphite satisfies all these criteria except for mechanical robustness, manufacturability, and low cost. Bipolar plates are a significant part of the PEMFC stack and account for 80% of the total

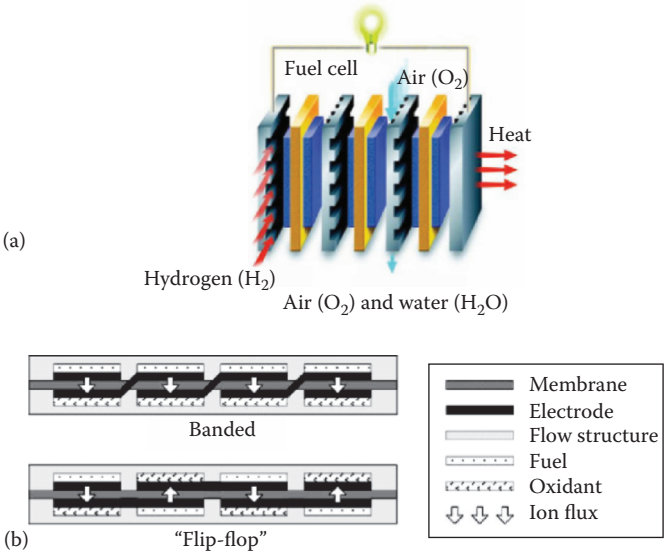


FIGURE 9.12 (a) Fuel cell stack. (b) Two methods for arraying fuel cell elements.

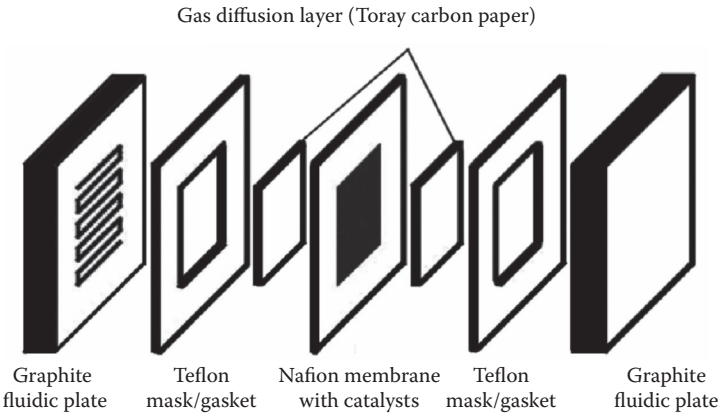


FIGURE 9.13 Bipolar plates (also called graphite fluidic plates) and typical membrane electrode assembly (MEA).

weight and 45% of stack cost. The bipolar plates will obviously be an important first target in miniaturizing a fuel cell. The membrane electrode assembly (MEA) in a PEMFC, illustrated in Figure 9.13, may consist of a Nafion® membrane (Nafion 112, Electrochem, Woburn, MA) coated with catalyst sandwiched between two carbon paper layers (e.g., Toray carbon paper). The carbon papers act as gas diffusion layers, and they are pushed up against the Nafion membrane using two Teflon gaskets.

A wide variety of micromachining techniques have been employed to construct microfuel cells, including Si MEMS-based, printed circuit (PC)-board-based, and carbon MEMS-based. An example of a Si-based MEMS-based fuel cell is shown in Figure 9.14.[23]

In this Si design, electropolished tubs in the highly conductive Si bipolar plates make for gas flow channels. Nafion is used as the membrane, and porous Si plays the role of gas diffusion layer to bring the gas to the Pt/C electrodes on the Nafion surface. Other Si MEMS-based fuel cells are reviewed in Table 9.4.

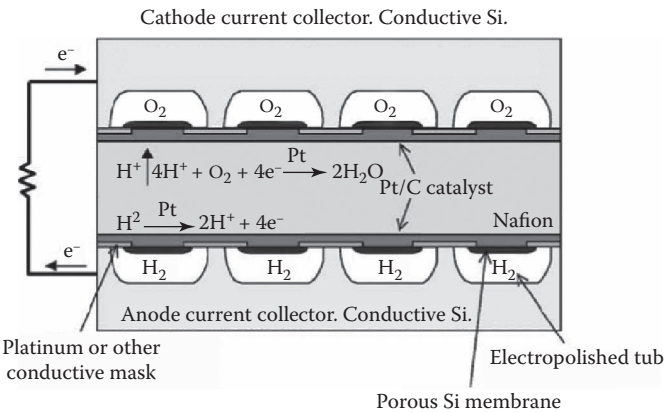


FIGURE 9.14 Silicon-MEMS-based fuel cell.[23]

Silicon is expensive compared with the printed circuit board-based fuel cell depicted in Figure 9.15.

Dr. Ben Park made the world's smallest hydrogen-air fuel cell with carbon-MEMS bipolar plates while working in the author's laboratory.[24] C-MEMS describes a manufacturing technique in which carbon devices are made by exposing a prepatterned organic structure to high temperatures in an inert or reducing environment (i.e., pyrolysis). The C-MEMS process makes the fabrication of intricate carbon shapes as easy as the patterning/machining/molding of a selected polymer material (see Chapter 5 for details).

For the manufacture of C-MEMS bipolar plates, 20-mil (~500-μm)-thick Cirlex® (Fralock, Canoga Park, CA) plastic sheets were machined using a circuit board milling tool (T-Tech, Norcross, GA) with 500-μm diameter end mills to create the fluidics. When starting with a size of the polymer bipolar plates of 1 × 1 cm, shrinkage during the pyrolysis process reduces the plates to 0.8 × 0.8 cm (see Figure 9.16).

For the membrane electrode assembly (MEA), commercial fuel cell electrodes [1 mg/cm^2 loading, 20 wt% Pt/Vulcan XC-72 (EC-20-10-7; Electrochem-Woburn, MA)] are pressed into an activated Nafion membrane* (Nafion 115) (Figure 9.17).

The fluidic channel walls and separators machined from high-temperature polymer sheets are bonded together (using PI-5878G polyimide) to create

* The Nafion membrane is activated as follows: deionized (DI) water for 1 h, 30% H_2O_2 for 1 h, 10 M H_2SO_4 for 1 h, and finally a short rinse in DI water (all at 80°C). The Nafion is stored in DI water until fabrication of the MEA.

Table 9.4 Silicon-MEMS-Based Fuel Cells

Source	Design	Fuels	Power Density
S.C. Kelley University of Minnesota	Front and back KOH-etched Si wafer, Nafion Membrane with custom Pt, Pt-Ru catalysts coatings	0.2 L/min hydrated neat **hydrogen** 0.2 L/min ambient air	130 mW/cm^2
J.P. Meyers Bell Labs	Porus Si etching, electropolished channels, Nafion membrane, Pt/C catalyst	**Hydrogen** and oxygen at 1 atm	60 mW/cm^2
S.J. Lee Stanford University	Front and back RIE Si wafer, Nafion membrane, carbon cloth and platinum, **series-connected**	Compressed **hydrogen** and oxygen	40 mW/cm^2
R. Hahn Fraunhofer, IZM	Wafer level and foil processes, screen printing	0.5 sccm **hydrogen (DMFC) NO AIR PUMPING**	80 mW/cm^2
S.C. Kelley University of Minnesota	Front and back KOH-etched Si wafer, Nafion Membrane with custom Pt, Pt-Ru catalysts coatings	0.2 L/min 0.5 M **methanol**, H_2O 0.2 L/min ambient air	15 mW/cm^2
Y.H. Seo KAIST, Korea	Front and back KOH and RIE Si Wafer, Nafion membrane, Pt catalyst	**Methanol** and ambient air **NO FUEL OR AIR PUMPING**	0.1 mW/cm^2

Source: Presentation by Hahn, R., S. Wagner, H. Reichl, M. Krumm, and K. Marquardt. 2003. Fraunhofer IZM and TU-Berlin. Development of micro fuel cells with help of MEMS technologies. MINATEC 2003.

fluidic plates as shown in Figure 9.18. The polymide physical binder also acts as the electrical binder when converted to carbon during pyrolysis. The MEA, fluidic plates, and gas inlet/outlets (cut from conventional syringes; see Figure 9.19b) are brought together, and epoxy is used to seal the entire fuel cell structure.

Making bipolar plates by C-MEMS technology has several benefits. Because one starts from an easy-to-machine polymer, new types of bipolar plates become feasible, and because the bipolar plate fluidics, gas diffusion layer, and catalyst support layer are all made of carbon, they can be integrated and fabricated into a single monolithic structure. This reduces weight, size, complexity, and internal resistance while increasing mechanical robustness. Control over the carbon precursor material allows materials engineering of the carbon itself.

The open-circuit voltage of an individual unoptimized C-MEMS cell has been measured at 871 mV, and the closed-circuit current draw of the fuel cell stabilized at 3.11 mA (with a voltage of 110 mV). The maximum power output was 0.773 mW (1.21 mW/cm^2), with an internal resistance of about 210 Ω.[24]

Our team also successfully fabricated a hydrogen-air C-MEMS fuel cell stack, as depicted in Figure 9.19.[24] The gaskets between cell elements are made of silicon rubber, which provides elastic support of the repeating parts and prevents leakage of gases. The end plates are machined in a T-Tech circuit board milling tool to create the screw holes and openings for gas ports. The latter are cut from conventional syringes. Repeating the bipolar plates, the gaskets, and the MEAs and then compressing them with four screws in the four corners of the whole structure results in the completed fuel cell stack, as shown in Figure 9.19.

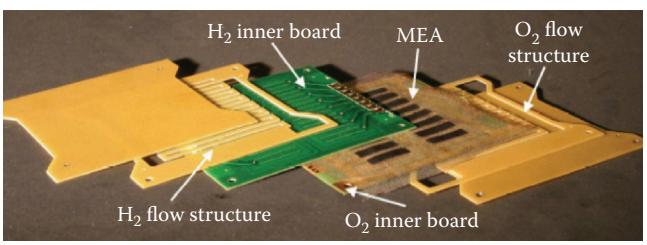


FIGURE 9.15 Printed circuit board fuel cell. (From O'Hayre, R. 2004. Microscale electrochemistry: Application to fuel cells. Thesis, Stanford University.)

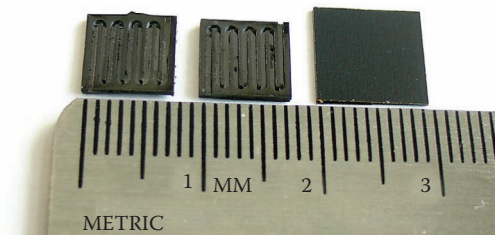


FIGURE 9.16 Cirlex bipolar plates of 1 × 1 cm (one structure on the right) shrink during the pyrolysis process to 0.8 × 0.8 cm (two structures on the left).

FIGURE 9.17 The membrane electrode assembly. A Nafion sheet (transparent) is sandwiched between two carbon electrodes.

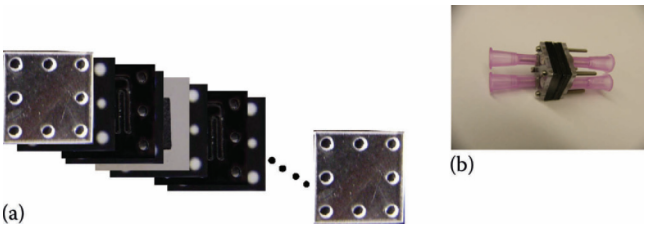


FIGURE 9.19 Hydrogen-air carbon-MEMS fuel cell stack. (a) The MEA, fluidic plates, and gas inlet/outlets (cut from conventional syringes) are brought together, and epoxy is used to seal the entire fuel cell structure (b).

The C-MEMS fuel cell could be further improved by engineering the carbon electrodes further. This may involve at least three strategies: incorporation of SWNTs and/or multiwall carbon nanotubes (MWNTs), the use of carbon horns, and the use of carbon fractals. Compared with other forms of carbon electrodes (such as graphite and bare glassy carbon electrode), MWNT microelectrodes show better performances because of the dimensions of the MWNTs and their electronic structure.[24] As we pointed out above, fuel cell electrodes, made of carbon nanohorns are expected to help improve the cells' power-generation capacity even further and extend their lifetime, because platinum catalyst nanoparticles disperse better among carbon nanohorns and do not aggregate.[24] The concept of implementing carbon fractal electrodes for both batteries and fuel cells is introduced below (see "Fractals and Fractal Dimensions").

Finally, in the microbiofuel cell context, we briefly highlight hybrid insect-MEMS (HI-MEMS; also called *cyborg bugs*, or *cybugs*), a concept sponsored by the Defense Advanced Research Projects Agency (DARPA). The MEMS on the insect are coupled to the insect anatomy and nervous system and draw power from the insect to control its locomotion by obtaining motion trajectories either from global positioning system (GPS) coordinates or by using radiofrequency (RF)-, optical-, or ultrasonic-based remote control. DARPA has military applications in mind for HI-MEMS, with visions of insect swarms with various embedded MEMS sensors (such as cameras, microphones, and chemical sniffers) penetrating enemy territory (e.g., the Crawford Ranch). The HI-MEMS battalions would perform reconnaissance missions beyond the capabilities of soldiers. But robo-insects, equipped with chemical sensors, might also have civil applications, perhaps to explore a hazardous chemical spill zone and report back about it to humans in a

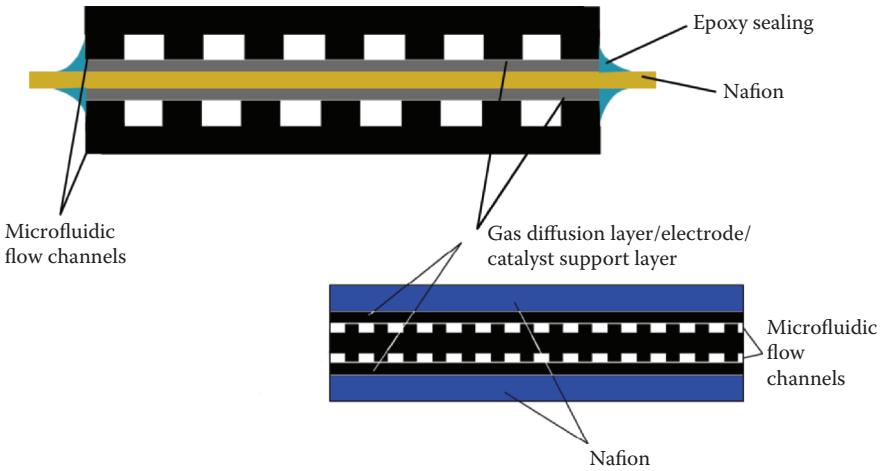


FIGURE 9.18 Assembly of carbon-MEMS fuel cell.

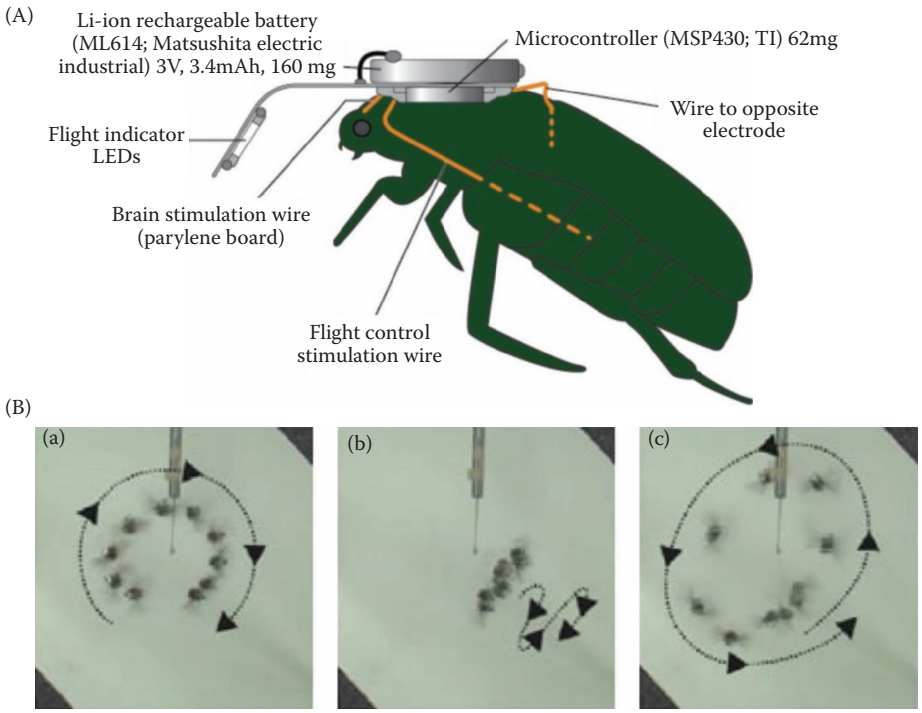


FIGURE 9.20 (A) Cyborg beetle system. The microcontroller controls flight by stimulating brain, wing, and other nerves. An array of ten light-emitting diodes hanging in front of the beetle's eyes is lit to control flight without direct nerve stimulation. The beetle weighs 1–2 g, and the total control system weighs 422 mg. (From Nikkei Electronics Asia, April 2008.) (B) Electrical stimulation of wing muscles on either side initiates a turn. Beetles mounted on a long string (10 cm) were programmed with a continuous sequence of left, pause, right, pause instructions; each instruction lasted 2 s. Left flight muscle stimulation generates a right turn (a), followed by a pause during which the beetle zigs and zags (b), followed by right muscle stimulation that generates a left turn (c). Each photograph consists of ten frames; frames were taken every 0.2 s. (From MEMS 2008, an international academic conference on MEMS that took place January 13–17, 2008, in Tucson, AZ.)

safe zone. A recent embodiment of the HI-MEMS concept comes from a University of Michigan team that successfully created a cyborg unicorn beetle microsystem (Figure 9.20A).[25] The University of Michigan researchers equipped a North American unicorn beetle with an IC, button battery, and a flight direction indicator consisting of ten light-emitting diodes (LEDs) and succeeded in controlling its flight externally. Flight control is handled by inserting electrodes into a portion of the brain immediately behind the beetle's eye and into wing muscles, and injecting AC pulses. The LED directional controller constitutes an independent control system and utilizes the beetle's proclivity to follow a blinking light. The cyborg beetle is able to take off and land, turn left or right, and demonstrate a number of other flight control behaviors. Electrical stimulation of wing muscles on either side initiates a turn (Figure 9.20B).

The DARPA program director, Dr. Amit Lal, got the idea for remote-controlled insects from the 1990 science fiction novel *Sparrowhawk* by Thomas A. Easton. In that novel, Easton writes about genetically engineered animals that are greatly enlarged and then outfitted with implanted control structures.

Heat Engines in MEMS

A heat engine converts heat to mechanical energy. Examples of everyday heat engines include the steam engine, the diesel engine, and the gasoline (petrol) engine in an automobile. Heat engines are divided into two groups: external-combustion engines and internal-combustion engines. External-combustion engines produce hot gases that transfer heat to another fluid, and the energy in this fluid, in turn, is changed into mechanical energy. External-combustion engines include gas and steam turbines and reciprocating steam engines. Internal-combustion engines produce hot gases whose heat energy is changed directly into mechanical energy.

Several groups involved in power MEMS are investigating scaled-down versions of well-established macroscale combustion devices. The idea is that since batteries have low specific energy, and liquid hydrocarbon fuels have a very high specific energy, a miniaturized power generating device, even with a relatively inefficient conversion of hydrocarbon fuels to power would result in increased lifetime and/or reduced weight of an electronic or mechanical system that currently requires batteries for power. A gas turbine powering a generator, for example, could in principle pack thirty times the energy of any battery. Refueling would replace recharging. There are quite a few difficulties with this approach, though, one being that flames extinguish because of heat losses if the dimensions of the combustion chamber are too small. Furthermore, even if flame quenching does not occur, heat and friction losses become increasingly important at smaller scales, since the heat release due to combustion and thus power output scales with the volume of the engine, whereas the heat and friction losses scale with the surface area.

As an example application of microcombustion engines, consider satellite missions that deal with smaller and smaller satellites, so that launch systems and altitude control thrusters of reduced mass and volume are now needed. Thrusters are devices in which propellant is burnt or disintegrates to produce a controlled force in a selected direction. These small rocket engines are expected to produce thrust of the order of tens of newtons at a thrust-to-weight of over 1000, i.e., 10 times the thrust-to-weight of conventional chemical liquid bipropellant engines. A simple MEMS thruster fabricated of a silicon/glass structure is shown in Figure 9.21.[26]

The nozzle, plenum, chamber, and injector of this MEMS thruster are all etched in silicon, and a glass ceiling is bonded to the etched silicon part. Propellants that have been used in this structure include H_2O_2 and N_2H_4 with silver and platinum on Al_2O_3 as catalysts. Hydrogen peroxide is fed through the plenum and injector and reacts in the presence of the catalyst to produce heat and gases in the chamber. The gases expand through the nozzle to produce thrust.

Another postage-stamp-size microthruster, fabricated at Massachusetts Institute of Technology (MIT), was made of six layers of silicon fused together

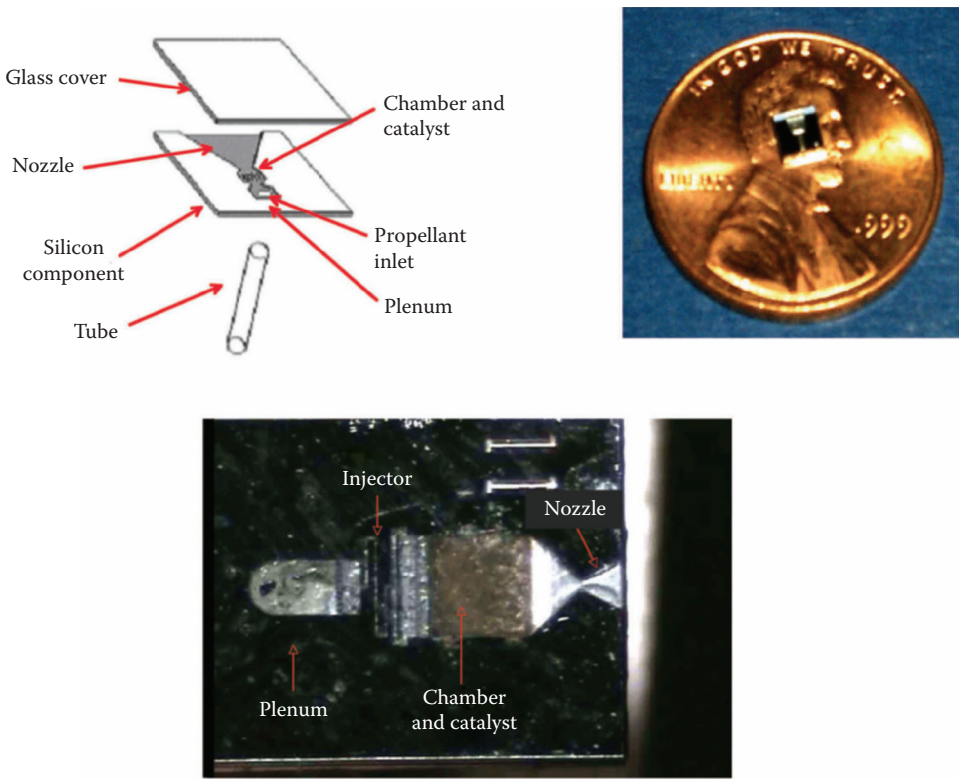


FIGURE 9.21 Simple silicon/glass thruster structure. (From a talk by Apurva Varia, NASA Goddard Thermal Model of MEMS Thruster.)

to make a sandwich 3 mm thick and was projected to produce up to 15 N of thrust by burning oxygen and methane. To prevent melting, ethanol coolant circulates around the tiny flat thrust chamber. The thrust level from these micromachined devices is expected to be very high in relation to their mass. MIT's Adam London thinks that a two-stage, 80-kg rocket might be sufficient to put a Coke-can-sized payload in orbit using 800 of his Si thrusters.[27] A more recent MIT microrocket engine thrust chamber and nozzle design measures $18 \times 14 \times 3$ mm and builds on London's earlier gaseous propellant work, expanding the operating envelope of these motors to higher thrust levels.[28] Directions for improved specific impulse engines include increasing the engine size by a factor of 2–4 and research on hydrogen peroxide as a coolant. Another approach is to use liquid hydrocarbons or solid propellants. With the combustor working at 1500–1600 K, Si with a melting point of 1685 K is out as a building material. SiC and perhaps quartz are alternative candidate materials.

MEMS thruster arrays for satellites offer several advantages over conventional thrusters: they have no moving parts; utilize a variety of propellants; are scalable; eliminate the need for tanks, fuel lines, and valves; and fully integrate the structure of the satellite with the propulsion to power it. The Honeywell-Princeton MEMS Megapixel Microthruster Array (MMMA) consists of a 256 × 256 array of pixels, each one capable of being fired as a separate thruster (Figure 9.22). The thruster operates by first igniting a small amount of lead styphanate, which in turn ignites the main propellant of a nitrocellulose mixture. This is then exhausted from the fired pixel out a channel to produce thrust. Thrusters in this class give very small and precise impulse bits (on the order of tens of mN-s). Another microthruster array, developed by TRW, Caltech, and the Aerospace Corp., is fabricated as a three-layer silicon and glass sandwich, with the middle layer consisting of multiple small propellant cells sealed with a rupturable diaphragm on one side and an ignitor on the other. Each cell is a separate thruster, and when ignited, delivers one impulse bit. Delivering propulsion in discrete increments by igniting thrusters in controlled sequences has lent the technology the name "digital propulsion."

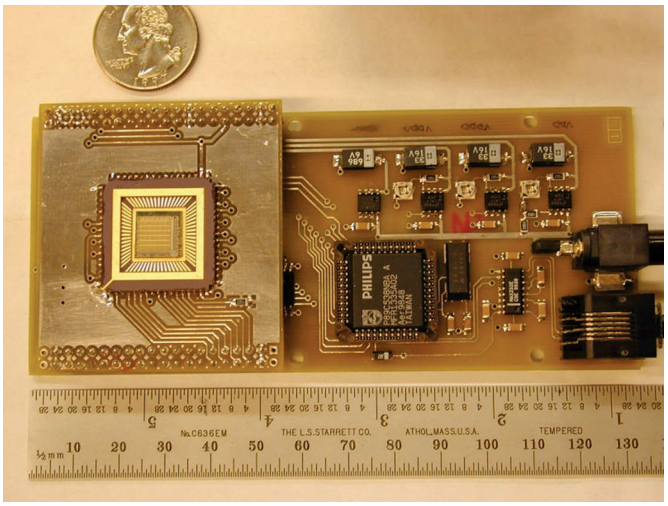


FIGURE 9.22 The MEMS Megapixel Microthruster Array ignitor thruster with its firing control board (http://alfven.princeton.edu/projects/MEMS1.htm).

The MEMS thrusters discussed above are all internal-combustion devices. External-combustion MEMS devices, with power generation by the alternate heating and cooling of a working fluid or a solid (e.g., shape memory alloys) integrated on a chip, have been attempted for driving a load as well. The heating in such an engine results typically from passing a current through a resistor. It would be preferable to use infrared radiation instead, since in that case no leads need to connect to the chip. It has been projected that a gas-based external heat engine of $5 \times 5 \times 5$ mm^3 might provide an output of 10–100 W/kg. With actuators based on shape memory alloys, an output of up to 1 kW/kg is feasible, with an efficiency ten times lower.[6] Problems associated with crafting MEMS external heat engines include the thermal isolation of heating and cooling sections, minimization of friction, and the difficulty of implementing a flywheel. Some of these problems were successfully addressed by Sniegowski and coworkers,[29,30] who demonstrated a surface micromachined microengine capable of delivering torque to a micromechanism. Angular velocities of 600,000 rpm were registered for this engine, driven by steam.

Solar Cells

Introduction

The conversion of light (photons) to electricity (voltage) in a solar cell is called the *photovoltaic* (PV)

effect. The most iconic form of solar power, the solar cell, consumes no fuel, has a wide power-handling capacity, has a high power-to-weight ratio, produces no pollution, and is more and more often made from inexpensive polycrystalline or amorphous materials, with even polymer-based solar cells being introduced. The operating principle of a solar cell was covered in Volume I, Chapter 4 (Figure 4.42), and in Table 9.5 we present a short historical overview of some important solar cell related milestones since the original observation of the photoeffect by Antoine Becquerel in 1839. As with batteries, progress in improving solar cell efficiency has been rather slow (there is no Moore's law curve for solar cells, either). Several important measures are used to characterize solar cells, the most obvious one being the total amount of electrical power produced for a given amount of solar power input. Expressed as a percentage, this is known as the solar conversion efficiency. Electrical power is the product of current and voltage, so the maximum values for these measurements are important as well: J_{sc} (short circuit current) and V_{oc} (open circuit voltage), respectively. Finally, in order to understand the underlying physics, the *quantum efficiency* is used to compare the chance that one photon (of a particular energy) will create one electron. A solar cell must be capable of producing electricity for at least twenty years without a significant decrease in efficiency, so lifespan is another crucial characteristic.

Table 9.5 Milestones in Solar Cell Development

1839: Discovery of photovoltaic effect between two electrodes in a liquid (Antoine Cesar Becquerel)
1876: Photovoltaic effect in Se (Heinrich Hertz)
1883: First solar electric module (Charles Fritts)
1906: Pochettino studies the photoconductivity of anthracene
1930: Research on Cu_2O/Cu solar cell
1941: Patent on Si solar cell (R. Ohl in Bell Labs)
1954: Crystalline Si solar cell (Bell Labs); 4% efficiency
1958: Solar panels as backup and power assist in space (*Vanguard 1*); 5 mW
1970: Zhores Alferov and his team in the USSR created the first GaAs heterostructure solar cells
1973: Oil crisis—solar energy is hot, subsequent Republican United States president kills Carter's efforts in favor of oil
1980: Solar cell using CdTe, $CuInSe_2$,TiO_2, etc.
1991: Liquid-junction solar cells with photosensitized anode invented by Michael Grätzel and Brian O'Regan at the École Polytechnique Fédérale de Lausanne, known as Grätzel cells
1995: Yu and Hall make the first bulk polymer/polymer heterojunction photovoltaic cell
2000: Peters/van Hall used oligomer-C_{60} diads/triads as the active material in photovoltaic cells (see Figure 3.15, for an example coming from the Stoddard-Zinc group)
2001: Schmidt-Mende made a self-organized liquid crystalline solar cell of hexabenzocoronene and perylene

Because of the soaring oil prices, solar energy is a hot topic again, as it was under President Carter (even Republicans now concede that cheap oil is not a US birthright). The photovoltaic cell is currently the fastest-growing form of alternative energy, increasing by 50% a year, and where a kWh of PV electricity cost 50 cents in 1995, it fell to 20 cents in 2005. BBC Research, which charts technology markets, expects the global solar market to grow to $32 billion by 2012, with thin-film solar cells expanding 45% a year.

At the earth's surface, the average solar energy is $\sim 4 \times 10^{24}$ J per year, and the 2001 global energy use was $\sim 4 \times 10^{20}$ J per year (increasing ~2% annually). In the United States alone, the average power requirement is 3.3 TW. With a 10% efficient solar cell, this means that we would need 1.7% of the land area of the United States devoted to PV to quench this thirst; this roughly corresponds to the area occupied by the interstate highways.

As we saw in Volume I, Chapter 4, a traditional semiconductor solar cell is made from an n-type and a p-type semiconductor put in contact with each other. The result is a region at the interface, the p-n junction, where charge carriers are depleted and/or accumulated on each side of the interface. The transfer of charge carriers produces a potential barrier, typically 0.6–0.7 V high. In the case of silicon (bandgap = 1.1 eV), sunlight provides enough energy to promote electrons out of the valence band into the conduction band, and when a load is placed across the cell, these electrons flow from the p-type to the n-type material, move through the external circuit, and then stream back into the p-type material, where they recombine with the valence-band holes they left behind. How sunlight creates an

electrical current in a solar cell is detailed a bit further in Figure 9.23.

If the incoming photon energy $E_\gamma < E_g$, then the photon is not absorbed; with $E_\gamma > E_g$, the photon is absorbed and the energy $E_\gamma - E_g$ is lost via thermalization. Additional losses occur through radiative recombination, trap states, and Auger recombination. If the carrier lifetime is less than the recombination lifetime then the charge is swept away by the inbuilt electric field (E). This depends on carrier mobility, defect density, carrier diffusion length, and exciton energy.

The conventional approach to PVs, as sketched in Figure 9.23, comes with several disadvantages. First, the semiconductor bandgap dictates that only photons with the bandgap energy or more contribute to producing a current. Higher-energy photons, at the blue end of the spectrum, have more than enough energy to cross the bandgap, but much of it is wasted as heat.

A cascade or tandem cell can achieve a higher total conversion efficiency by capturing a larger portion of the solar spectrum. In a typical multijunction cell, individual cells with different bandgaps are stacked on top of one another. The cells are stacked in such a way that sunlight falls first on the material having the largest bandgap. Photons not absorbed in the first cell are transmitted to the second cell, which then absorbs the higher-energy portion of the remaining solar radiation while

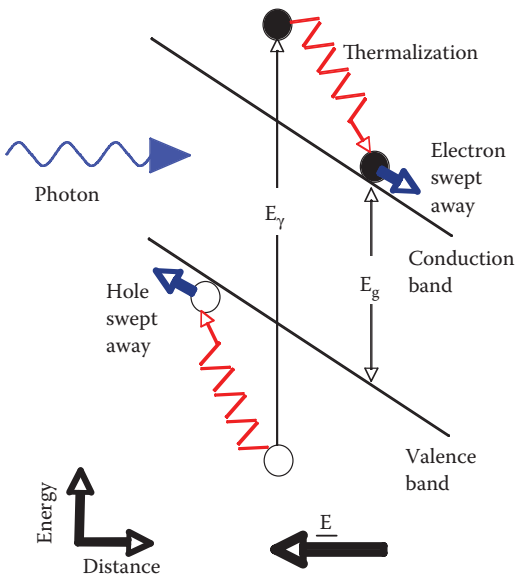


FIGURE 9.23 Energy diagram of a solar cell.

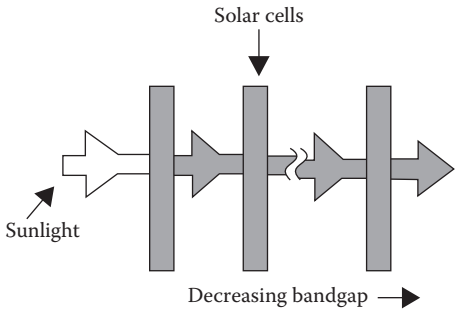


FIGURE 9.24 Multijunction, cascade, or tandem solar cell.

remaining transparent to the lower-energy photons. These selective absorption processes continue through to the final cell, which has the smallest bandgap (Figure 9.24). The ideal bandgap combination for a tandem solar cell is typically 1.75 and 1.1 eV. Such bandgap combinations can be attained using lattice-matched III–V semiconductor material combinations. The best monolithic combination is represented by the GaInP/GaAs tandem cell, possessing an AM 0 efficiency of 26.9%. The ideal efficiency for an infinite stack is 86.8%, and for a GaInP/GaAs/Ge triple-junction cell an efficiency of 32% has been measured.

Multijunction solar cells dramatically improve efficiency, but these cells are still very high-cost and suitable only for large commercial deployments (combinations include GaAs/InGaP, Ge/InGaAsN, InP/InAlAs, InP/InGaAs, and GaN/InGaN). In general terms, the types of cells suitable for rooftop deployment have not changed significantly in efficiency, although costs have dropped somewhat because of increased supply.

A higher light intensity can be focused on the solar cells by the use of mirror and lens systems, perhaps even a system that tracks the sun. This is perhaps too expensive a solution for rooftop implementation but could be useful in MEMS devices.

Another problem associated with conventional solar cells is that in order to capture enough photons in the p-type layer, that part has to be quite "thick." However, a thick p-type layer also increases the chance that electrons and holes will recombine before reaching the p-n junction. These effects combine to produce an upper limit of about 20% on the efficiency of silicon solar cells. The current problem with solar cells is not really the efficiency (20%), but rather the cost of production. Though it has come

down by a factor of 100 in fifty years, it still has to go down by a factor of 2–5 to be competitive with fossil plants.

The best known solar cells today remain based on Si, and they are either monocrystalline, polycrystalline, or hydrogenated amorphous (α-Si:H) based. Polycrystalline films are poured and are cheaper and simpler to make than monocrystalline silicon that is Czochralski grown (see Volume I, Chapter 4). The efficiency of polycrystalline cells is, however, somewhat less than that of monocrystalline cells. Hydrogenated amorphous Si (α-Si:H) cells are relatively cheap, but they come with the lowest efficiency (see Table 9.6). Although its semiconducting properties are inferior to those of single-crystal Si, we learned in Chapter 5 that α-Si:H, manufactured by plasma-enhanced chemical vapor deposition (PECVD) from silane, can be doped both n- and p-type (in contrast to α-Si) (see Table 5.8 for state-of-the-art parameters of PECVD α-Si:H). Exposure of α-Si:H to light increases photoconductivity by four to six orders of magnitude, and the absorption coefficient of α-Si:H is more than an order of magnitude larger than that of single-crystal Si near the maximum solar photon energy region of 500 nm. Accordingly, the optimum thickness of the active layer in an α-Si:H solar cell can measure 1 μm, much smaller than that of single-crystal Si solar cells. In Figure 5.15, we compared polycrystalline-based solar cells with ultrathin amorphous Si on glass and ultrathin amorphous Si on flexible substrates.

Solar cell technology represents the most MEMS-compatible technology for power integration. Lee et al.[31] note that hydrogenated amorphous silicon solar cells are a particularly attractive means to realize an onboard MEMS power supply, and they implemented a high-voltage integrated solar cell array[31] as an electrostatic MEMS power supply. By interconnecting 100 individual solar cells in series, this team measured an open circuit potential as high as 150 V under AM 1.5 conditions, a voltage high enough to drive onboard electrostatic actuators. The conversion efficiency in that effort was only 0.2%, however. Sakakibara et al.[32] were able to generate more than 200 V with a similar solar cell on an area of 1 cm^2 and obtained a conversion efficiency of 4.65%. In both cases, amorphous silicon was used in a triple-stacked PV structure generating up to 2.3 V per cell. To obtain a very dense packing of array elements and to make the series connection of the solar cells, the latter group used focused laser beams for patterning electrodes and PV materials.

Since solar light is only intermittently available, electric storage elements need to be implemented as well. Along this line, Kimura et al.[33] fabricated a miniature optoelectric transformer consisting of a p-n junction photocell and a multilayer spiral coil transformer.

III/V solar cells are cells with a high efficiency, but they are made of very expensive semiconducting materials. A single junction cell based on GaAs comes with an efficiency of 25.1%. In 1970, Zhores Alferov and his team in the USSR created the first GaAs heterostructure solar cells. Today, the combination of GaAs with germanium and indium gallium phosphide is the basis of a triple-junction solar cell that holds the record for solar light conversion efficiency of over 32% and can operate also with light as concentrated as 2000 Suns.* This type of solar cell powers the robots *Spirit* and *Opportunity*, which are exploring the surface of Mars.

Newer solar cell materials include thin films of copper indium diselenide ($CuInSe_2$), copper indium gallium diselenide (CIGS), and cadmium telluride (CdTe). Commercial leaders in this area are a firm called First Solar, which uses cadmium telluride, and Miasolé, which manufactures a thin-film PV cell based on an ultrathin layer of photoactive CIGS on a stainless steel foil only 50 μm thick. The latter can easily be used in PV modules or incorporated into building materials such as membrane roofing. The efficiencies and bandgaps of GaAs and two of

Table 9.6 Silicon-Based Solar Cell Efficiency in %

Type	Efficiency in Laboratory	Efficiency in Production
Monocrystalline	24	14–17
Polycrystalline	18	13–15
Amorphous	13	5–7

* Where 1 Sun equals 1000 W/m^2 at the point of arrival. One Sun is a unit of power flux.

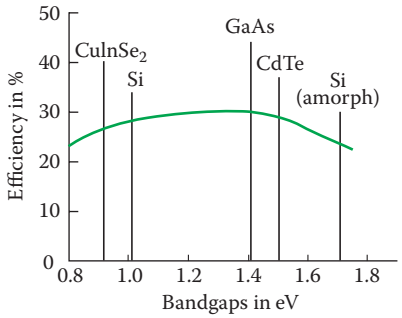


FIGURE 9.25 Bandgaps and efficiencies of the most important solar cell materials. The bars represent theoretical efficiencies, and the smooth line represents practical efficiencies.

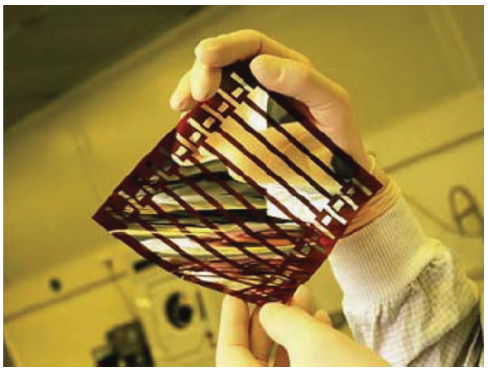


FIGURE 9.26 Polymer or organic photovoltaics.

the newer solar materials are compared with those of single-crystal and amorphous Si in Figure 9.25.

One development that made organic PVs an attractive option is the availability of light-sensitive conjugated polymers: polymers with alternating single and double carbon-carbon (sometimes carbon-nitrogen) bonds. In the 1970s, it was discovered that chemical doping of conjugated polymers increased electronic conductivity several orders of magnitude. Since that time, electronically conducting materials based on conjugated polymers have found many applications, including sensors, LEDs, and solar cells. Advantages of polymer or organic PVs (OPVs) include the fact that they are easily processed over large areas using spin-coating, doctor blade techniques (wet-processing), evaporation through a mask (dry processing), and printing (Figure 9.26). The usage of roll-to-roll printing, especially, guarantees a favorable cost (see Figure 1.5). They also have a low weight, exhibit mechanical flexibility and transparency, and have a bandgap that can easily be tuned chemically by incorporation of different functional groups. Efficiencies stand at 5%, but there are no obstacles identified for 10%. The low cost of plastic PV may enable solar power in applications where it was previously uneconomic (e.g., solar tents, sails, packaging, etc.). For rigid cells, ITO (Indium tin oxide)-coated glass usually serves as the substrate for these organic cells.

Figure 9.27 summarizes the best research solar cell efficiencies for a broad range of materials from universities, research institutes, and private companies. Besides PV converters for solar light and laser

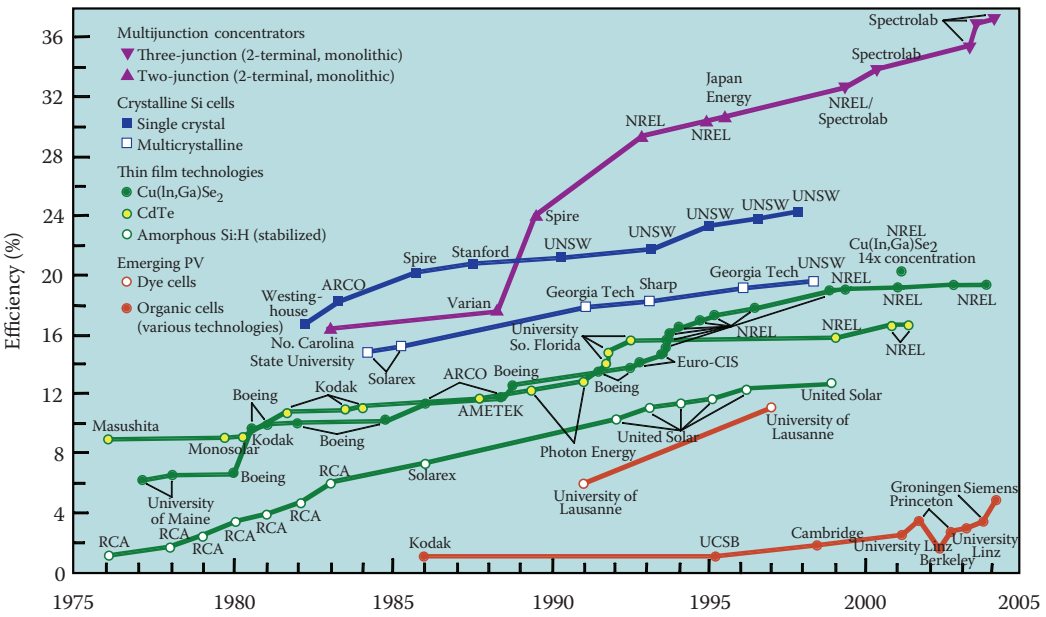


FIGURE 9.27 Research solar cell efficiencies. (From Art Nozik, DOE Solar Energy Workshop, 2005.)

light, microwaves could be used to power microsystems. In the latter, extremely small receivers and converters need to be constructed.

MEMS and NEMS in Solar Cells

Introduction

Micro- and nanotechnologies are crucial to increasing the efficiency of existing solar cells, and NEMS is enabling totally new solar cell designs as well. MEMS techniques allow for the fabrication of lenses (including Fresnel lenses), mirrors, photonic crystals, and cavities and the texturing of solar cell surfaces to enhance the interaction of incoming photons with the solar cell active junctions. As an example of a nanoenhancement of existing solar cells, consider the use of quantum dots (QDs) to replace dyes in luminescent concentrators.

NEMS may make a much bigger impact, though, on future solar cells. QDs, for example, exhibit optical properties that are absent in the bulk material due to the confinement of electron-hole pairs, in a region of a few nanometers and the wavelength at which a QD absorbs or emits radiation can be adjusted at will (see Volume I, Chapters 3 and 5). A first advantage of QDs, then, stems from the fact that an optimum bandgap, corresponding to the highest possible solar-electric energy conversion, can be achieved by using a mixture of QDs of different sizes for harvesting the maximum amount of the incident light (E_γ). A second advantage stems from the increased possibility of multiphoton absorption by nanoparticles. The formation of multiple excitons per absorbed photon takes place only when the energy of the absorbed photon is far greater than the semiconductor bandgap (E_g). Multiple photon absorption does not readily occur in bulk semiconductors, where the excess energy quickly dissipates as heat before it can create other electron-hole pairs. However, in semiconducting QDs, the rate of energy dissipation is reduced dramatically, and with the charge carriers confined within a nanovolume, their interactions are enhanced, increasing the probability for multiple excitons per incoming photon. A third advantage of QDs is that in contrast to traditional crystalline semiconductor materials that are rigid, QDs can be molded into a variety of different forms, such as sheets or three-dimensional arrays. Quantum dots can easily be combined with organic polymers or dyes or made into porous films, and in colloidal form, they can be processed to create junctions on inexpensive substrates such as plastics or thin metal foils. A fourth advantage of QDs is that it is quite easy to decouple the absorption and charge separation process. Decoupling charge separation from light absorption results in many more engineering options to make the solar cell of the future. For example, the charge separation may be achieved in a semiconductor QD, while the light absorption may be in an organic dye adsorbed on its surface.

The broader quantum solar cell strategy is thus to sensitize a first material with bandgap E_g with another material of bandgap (E_q) smaller than E_g (i.e., $E_q < E_\gamma < E_g$), and this second semiconductor material can be a dye (monomeric), a photoactive biopolymer, a QD, or a quantum well. QDs have the advantage over dyes as they degrade less in sunlight. In its simplest form, a quantum well solar cell (QWSC) consists of a p-i-n solar cell with a multiple quantum well (MQW) system added to the intrinsic (i) region. The idea, as with a tandem solar cell, is that the quantum wells absorb light of energy below the bulk bandgap, E_g.

MEMS in Solar Cells

MEMS techniques allow for the fabrication of lenses (including Fresnel lenses), mirrors, photonic crystals, cavities, and solar cell surface textures to enhance the interaction of incoming photons with the solar cell. A simple method to put MEMS to work in a conventional solar cell is by surface structuring the cell so that the light hits the PV several times, as illustrated in the pyramid structuring in Figure 9.28.[34]

NEMS in Solar Cells

We present examples of quantum technology benefiting solar cells by highlighting nanoporous semiconductors in the liquid-junction Grätzel solar cell, two-photon excitation in hybrid organic-QD PV devices, and increased efficiency in a QWSCs.

In Volume I, Chapter 7, we learned that in liquid-junction solar cells, to avoid corrosion of the cell, one chooses a redox system where the maximum density

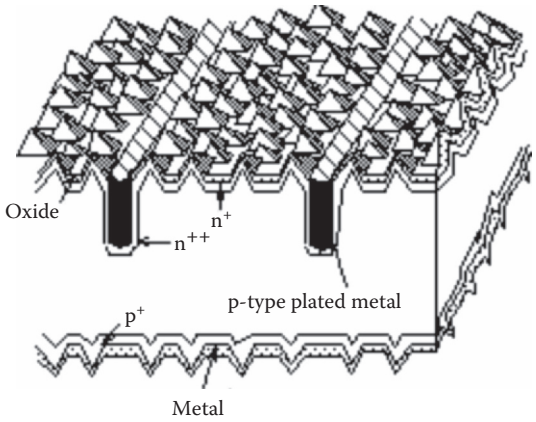


FIGURE 9.28 Surface structuring of a solar cell surface.[34]

of states of the reductant (the peak of the Gaussian in Volume I, Figure 7.62) is at the same level as the valence band edge of the semiconductor.[35] In this configuration, empty levels (holes) in the valence band, generated by shining light on the electrolyte/semiconductor interface, react with the reductant rather than with the semiconductor itself. In such liquid-junction solar cells, appropriate reductants thus compete effectively with the oxidation/corrosion of the semiconductor because they react more avidly with the holes in the valence band than the material itself.[36] The maximum voltage generated by such a cell, in theory, is simply the difference between the Fermi level of the semiconductor and the redox potential of the electrolyte.

The best known liquid-junction solar cell today is the dye-sensitized Grätzel design.[37–39] The Grätzel cell is significantly different from a conventional semiconductor solar cell, in which the semiconductor acts as the source of photoelectrons and also provides the potential barrier to separate the charges to generate the current. In contrast, in the Grätzel dye-sensitized solar cell, the semiconductor is used solely for charge separation, and the photoelectrons are provided from a separate photosensitive dye. As shown in Figure 9.29 the solar cell consists of a TiO_2 layer forming a nanoporous structure with a dye (e.g., ruthenium-polypyridine) spread throughout its surface. The dye molecules are small, and in order to capture a reasonable amount of the incoming light, the nanoporous structure is used as a scaffold holding large numbers of the molecules in a 3D matrix, vastly increasing the number of molecules for a given surface area. The

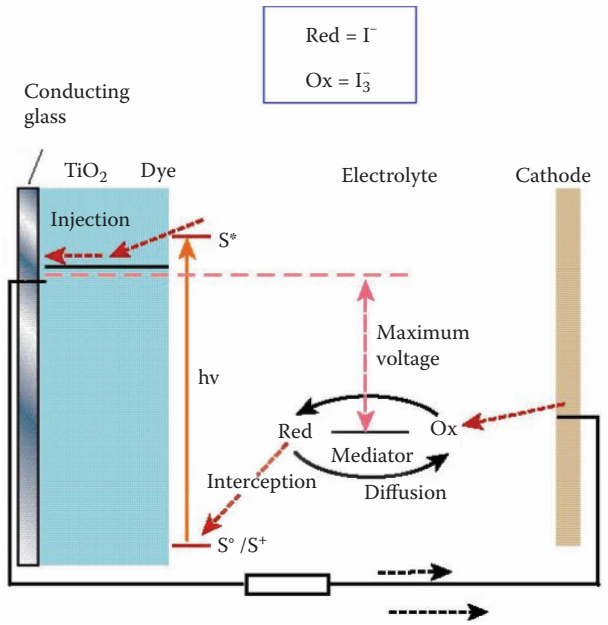


FIGURE 9.29 A liquid-junction solar cell: the dye-sensitized Grätzel design. The *S* energy levels represent the photosensitive dye levels.

charge separation is provided by the semiconductor-liquid junction contact. The TiO_2 layer sits on a transparent anode made of fluorine-doped tin oxide (SnO_2:F) deposited on the side of the glass plate facing the TiO_2 layer.

The electrolyte solution also holds a redox couple (e.g., I^-/I_3^-) mediator, located in the space between the dye-coated TiO_2 and a cathode, typically a thin film of platinum metal. Photons enter the cell through the transparent SnO_2:F window, and if they have enough energy, they are absorbed by the dye, creating an excited dye state. From this excited state an electron is "injected" into the conduction band of the TiO_2. This way, the dye molecule is oxidized, and it would decompose if it didn't quickly react with iodide in the electrolyte oxidizing it to form triiodide (dye regeneration reaction):

$$3I^- \rightarrow I_3^- + 2e^- \qquad \text{Reaction 9.6}$$

The reaction with iodide occurs very quickly compared with the recombination of the injected electron with the oxidized dye molecule, effectively preventing short-circuiting the solar cell. The injected electron then travels to the cathode via the external circuit, and the triiodide recovers its missing electron by diffusing through the solution to the

cathode, where it is reduced back to iodide (redox regeneration reaction):

$$I_3^- + 2e^- \rightarrow 3I^- \qquad \text{Reaction 9.7}$$

The electron injection process does not introduce a hole in the TiO_2, only an extra electron, and recombination of this electron directly from the TiO_2 to the electrolyte is not possible because of differences in energy levels (see Volume I, Chapter 7: if there are no overlapping energy levels, no reaction can occur). This is a very important property of the dye-sensitized Grätzel design: the electron-hole recombination that plagues the efficiency of traditional cells does not exist here. The potential used for external work in this solar cell is given by the following equation:

$$V_{ext} = E_F - V_{redox} \qquad (9.1)$$

The quantum efficiency of a Grätzel-cell is very high because the huge surface area of the TiO_2 nanostructure makes the chance that a photon will be absorbed very high, and the dyes are very effective at converting photons to electrons. Most of the losses are due to conduction losses in the TiO_2 and the clear anode, and the optical losses stem from the front electrode as well. The overall quantum efficiency is about 90%, with the "lost" 10% being largely accounted for by the optical losses in the anode. Although the dye is highly efficient in turning photons into electrons, only those electrons with enough energy to cross the TiO_2 bandgap result in current being produced. The bandgap is larger than that of silicon, which means that fewer of the photons in sunlight are usable for power generation. In addition, the electrolyte limits the speed at which the dye molecules can regain their electrons and become available for photoexcitation again. These factors limit the current generated by a liquid-junction solar cell; for comparison, a traditional silicon-based solar cell offers about 35 mA/cm², whereas a Grätzel cell offers about 20 mA/cm². The closest competitors in price-performance terms are the various thin-film approaches, which are currently somewhat more developed commercially.

The dyes used in early experimental Grätzel cells were sensitive only in the high-frequency end of the solar spectrum (ultraviolet and blue). Newer versions have a much wider frequency response; notably, the triscarboxy-terpyridine ruthenium complex $[Ru(2,2',2''\text{-}(COOH)_3\text{-terpy})(NCS)_3]$ is efficient right into the low-frequency range of red and infrared (IR) light. The wide spectral response results in the dye having a deep brown-black color, and it is referred to simply as *black dye*.[39] These polymers also are also sensitive to infrared are called *thermovoltaics*, and they generate electricity from both light and heat. Current Grätzel cells demonstrate a conversion efficiency of about 11%, whereas common low-cost commercial panels operate between 12% and 15%. Flexible thin-film cells are typically around 8%. Perhaps the major disadvantage of a liquid-junction solar cell is the use of liquid electrolytes, which brings leakage and corrosion problems, as well as temperature stability problems. At low temperatures, the electrolyte can freeze, ending power production and potentially leading to physical damage. Higher temperatures cause the liquid to expand, making sealing the cells a serious problem.

To avoid the above problems of liquid-junction solar cells, several research groups aim to produce a hybrid organic-QD PV device as the solid analog to the Grätzel cell. Researchers at the University of Toronto mixed lead sulfide (PbS) QDs with the polymer poly(2-methoxy-5-[2'-ethylhexyloxy-p-phenylenevinylene]) (MEH-PPV). Wrapping MEH-PPV around the QDs shifted the polymer's absorption into the infrared: on its own the polymer absorbs between ~400 and ~600 nm, and the QDs have absorption peaks that can be tuned from ~800 to ~2000 nm (see Figure 9.30).[40] The best efficiency registered thus far with this approach is 1.6%.[41]

In 2000, Nozik from the National Renewable Energy Laboratory in Golden, Colorado, predicted that theoretically, QDs could increase the efficiency of solar cells through multiple exciton generation.[42] "We have shown that solar cells based on QDs theoretically could convert more than 65 percent of the sun's energy into electricity, approximately doubling the efficiency of solar cells," said Nozik.[42]

In 2004, Schaller and Klimov from the Los Alamos National Laboratory in New Mexico were the first to demonstrate this phenomenon experimentally, using QDs made of lead selenide.[43] Multiple excitons

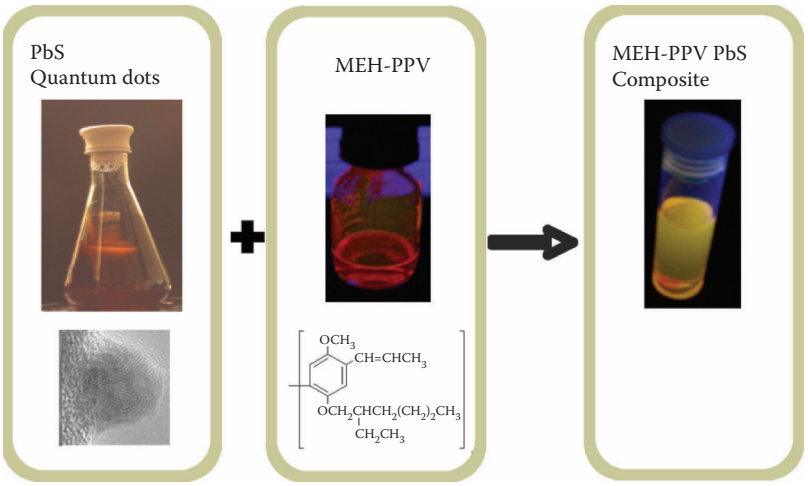


FIGURE 9.30 Lead sulfide (PbS) quantum dots are mixed with poly(2-methoxy-5-[2′-ethylhexyloxy-p-phenylenevinylene]) (MEH-PVV) to make a MEH-PPV PbS composite.[40]

start to form simultaneously as soon as the photon energy reaches twice the bandgap, and a group of researchers led by Nozik demonstrated that the absorption of a single photon by 2.9 nm PbSe QDs yielded not one exciton but three when the energy of the photon absorbed is four times that of the bandgap[43] (a quantum yield of 300%). Quantum dots made of PbS also showed the same phenomenon.

Quantum wells, as we saw in Volume I, Chapters 3 and 5, are ultrathin layers (nanostructures) of a narrower bandgap material (e.g., a GaAs well) sandwiched between regions of wider bandgap (e.g., AlGaAs barrier). They are made possible by the development of modern crystal growth techniques such as molecular beam epitaxy (MBE) and metal-organic vapor phase epitaxy (MOVPE). We also saw how quantum well lasers work at high forward bias and how quantum well photodiodes operate at reverse current and reverse bias. Quantum well solar cells (QWSCs) work in the forward bias and reverse current quadrant of the current-voltage plot, where power can be extracted.[44] In Figure 9.31 a first material with bandgap E_g is contacted with another material of bandgap (E_q) smaller than E_g, i.e., $E_q < E_\gamma < E_g$. The idea, as with a tandem solar cell, is that the quantum wells absorb light of energy below the bulk bandgap, E_g.

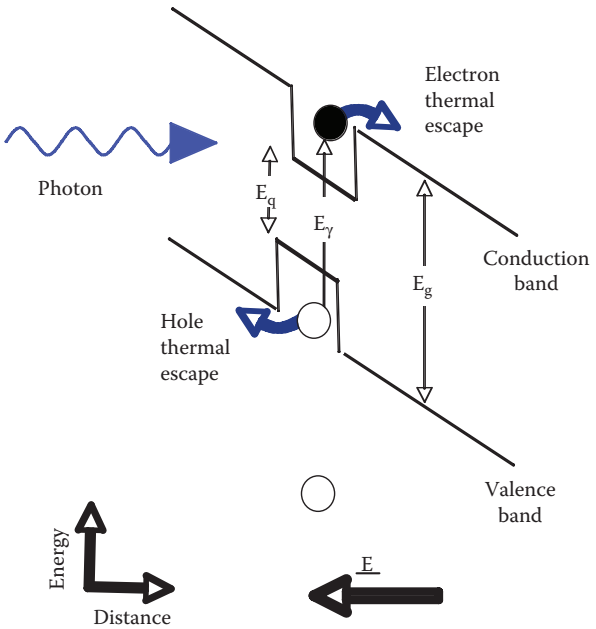


FIGURE 9.31 Schematic illustrating the quantum well solar cell.

In the simplest QWSC, several quantum wells are added to the *i*-region of a conventional p-i-n solar cell of wider bandgap (E_g). The quantum wells do absorb light of energy below the bulk bandgap E_g. If carriers escape from the wells before recombining, they enhance the output current, and short-circuit current (I_{sc}) can be strongly enhanced. The open-circuit voltage (Voc) is reduced, but the cell's efficiency is considerably enhanced.

Energy Scavenging (Harvesting)

MEMS devices either feature onboard power sources such as batteries, capacitors, fuel cells, radioactive elements, etc.; are powered by energy that is beamed in on purpose; or harvest energy from their own

environment. Energy harvesting is the process of capturing minute amounts of energy from one or more naturally occurring energy sources and storing that energy for later use. Harvestable energy sources for a MEMS device include mechanical (from sources such as vibration, mechanical stress, and strain, captured by ratchets and springs), thermal (waste energy from furnaces, heaters, and friction, captured by thermocouples, bimorphs, and thermoelectrics), light (captured from sunlight or room light via photosensors, photodiodes, or solar panels), electromagnetic (from inductors, coils, and transformers, captured by coils), natural (wind and ocean currents), the human body (a combination of mechanical and thermal energy naturally generated from bio-organisms or through actions such as walking and sitting), and other energy from chemical and biological sources. All these energy sources are virtually unlimited and essentially free, if they can be captured at or near the system location.

Naturally occurring temperature gradients are one potential source of power. Stark and Stordeur,[45] for example, demonstrated a thermoelectric device that can produce 15 mW of power from a 10°C temperature difference. The thermal gradient principle is employed in satellites, which are hot on the side turned toward the sun but cold on the side facing away from the sun. A satellite can be powered by the energy thus generated for up to thirty years. The principle also operates in reverse: thermoelectric elements can produce cold from electricity and are used these days to cool the interiors of computers. In the 1960s, it was believed that if transistors could replace vacuum tubes for electronics, then surely thermoelectrics could replace refrigerators. Requirements for a good thermoelectric (TE) material are a high figure of merit, ZT; this requires a high Seebeck coefficient, S, a high electrical conductivity, σ, and a low thermal conductivity, κ (see Equations 8.70–8.73). Unfortunately, ZT has to be improved from ~0.9 to 3 or higher to become commercially attractive. (ZT needs to improve to over 1.3 at 300 K for a major impact in electronics cooling and to around 2.5 for a revolutionary impact in air conditioning and power from waste heat.)

Some researchers have focused on scavenging power from the human body. Shenck and Paradiso,[46] for example, did create shoe inserts capable of generating 8.4 mW of power under normal walking conditions. [See also MIT Media Lab's Parasitic Power Harvesting project for devices built into a shoe (http://www.media.mit.edu/resenv/power.html).]

Mechanical vibrations are another promising power source. Scavenging the power from commonly occurring vibrations for use by electronics is both feasible and attractive for certain applications. Three popular methods to harvest energy from vibrations are piezoelectric, capacitive, and inductive. In the piezoelectric method, the strain in a piezoelectric material causes a charge separation (voltage across capacitor). Ottman et al.[47] have designed and optimized power circuitry to be used with a piezoelectric vibration-based generator. In a capacitive pick-up, a change in capacitance causes either a voltage or charge increase. Roundy et al.[48] have published electrostatic MEMS designs with simulated power outputs in the range of tens of microwatts per cubic centimeter. With an inductive pick-up, a coil moves through a magnetic field causing current in a wire. Amirtharajah, Chandrakasan, and coworkers[49] demonstrated an electromagnetic vibration-to-electricity converter with a 160-cm^3 volume ($4 \times 4 \times 10$ cm) that produces 400 mW of power (or 2.5 mW/cm^3). Table 9.7 compares the features of different vibrational harvesting techniques. Currently, a piezoelectric approach with PZT is the most popular because of its compact configuration and compatibility with MEMS, but its inherent limitations, including aging, depolarization, brittleness, and high-output impedance, limit many applications.[50]

As Goemans points out, the possibility of converting motion into electrical energy can be very attractive for cases where battery replacement is unacceptable, kinetic energy is abundantly available, and space is not too limited.[6] He lists biomedical implants, tire pressure monitoring systems, and electronic locks as potential application areas.[6] Some power harvesting systems are compared with batteries, fuel cells, solar power, and nuclear power in Table 9.8.

Table 9.7 Summary Comparison of the Different Vibrational Types of Harvesting Mechanisms[50]

Type	Advantages	Disadvantages
Electromagnetic	• No need of smart material • No external voltage source	• Bulky size: magnets and pick-up coil • Difficult to integrate with MEMS • Maximum voltage of 0.1 V
Electrostatic	• No need of smart material • Compatible with MEMS • Voltages of 2~10 V	• External voltage (or charge) source • Mechanical constraints needed • Capacitive
Piezoelectric	• No external voltage source • High voltages of 2~10 V • Compact configuration • Compatible with MEMS • High coupling in single crystals	• Depolarization • Brittleness in PZT • Poor coupling in piezo film (PVDF) charge leakage • High-output impedance
Magnetostrictive	• Ultrahigh coupling coefficient >0.9 • No depolarization problem • High flexibility • Suited to high-frequency vibration	• Nonlinear effect • Pick-up coil • May need bias magnets • Difficult to integrate with MEMS

Table 9.8 Energy Harvesting Compared with Batteries, Fuel Cells, and Nuclear Power

	Power (Energy) Density	Source of Estimates
Batteries (zinc-air)	1050–1560 mWh/cm^3 (1.4 V)	Published data from manufacturers
Batteries (lithium ion)	300 mWh/cm^3 (3–4 V)	Published data from manufacturers
Solar (outdoors)	15 mW/cm^2 – direct sun 0.15 mW/cm^2 – cloudy day	Published data and testing
Solar (indoor)	0.006 mW/cm^2 – my desk 0.57 m/Wcm^2 – 12 in. under a 60 W bulb	Testing
Vibrations	0.001–0.1 m/Wcm^3	Simulations and testing
Acoustic noise	3E-6 mW/cm^2 at 75 Db sound level 9.6E-4 mW/cm^2 at 100 Db sound level	Direct calculations from acoustic theory
Passive human powered	1.8 mW (shoe inserts >> 1 cm^2)	Published study
Thermal conversion	0.0018 mW – 10 deg. C gradient	Published study
Nuclear reaction	80 mW/cm^3 1E6 mWh/cm^3	Published data
Fuel cells	300–500 mW/cm^3 ~4000 mWh/cm^3	Published data

Fractal Power

Introduction

In this section, we show how the use of fractal-like electrodes in batteries and fuel cells minimizes the internal resistance while maximizing surface-to-volume ratios. We also demonstrate how C-MEMS techniques show great promise for use in the fabrication of such fractal electrodes for electrochemical applications.

Fractals and Fractal Dimensions

A fractal is a rough or fragmented geometric shape, each part of which is, at least approximately, a smaller copy of the whole. Another definition holds that a fractal is an irregular geometric object with an infinite nesting of structure at all scales. Fractals are said to be self-similar because no matter what the scale of observation, whether we zoom in or out, we appear to be seeing the same object. Fractals came to the public attention through the stunning abstractions generated in computer screen savers by the famous Mandelbrot set. Mandelbrot* introduced the idea of

* As a child in France, Mandelbrot wondered how to use the smooth regularities of Euclidean shapes to model the complexity of the world he saw around him. Where were the circles in nature? Where were the parallel lines and infinite planes? He concluded "clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth nor does lightning travel in a straight line."[51]

fractals in a paper that asked, how long is the coastline of Great Britain?[51] The coastline is irregular, so a measure with a straight ruler provides an estimate only. The estimated coastal length, *L*, equals the length of the ruler, *s*, multiplied by *N*, the number of such rulers needed to cover the measured object. Mandelbrot realized that the smaller his measuring stick, the longer the coast would turn out to be and that using molecular distances, the coast would have an infinite length. This means, Mandelbrot conjectured, that the Euclidean dimension, *D*, of a coastline is between the one dimension of a straight line and the two of a plane.

If we take an object residing in Euclidean dimension *D* and reduce its linear size by 1/*r* in each spatial direction, its measure (length, area, or volume) would increase by *N* times the original, with *N* given by the following equation:

$$N = r^D \tag{9.2}$$

This is pictured in Figure 9.32.

The German mathematician Felix Hausdorff introduced a more general treatment of dimension *D*. He pointed out that taking the log of both sides of Equation 9.2 one obtains the following:

$$\log(N) = D \log(r) \tag{9.3}$$

And if we solve for *D*, we get the following:

$$D = \log(N)/\log(r) \tag{9.4}$$

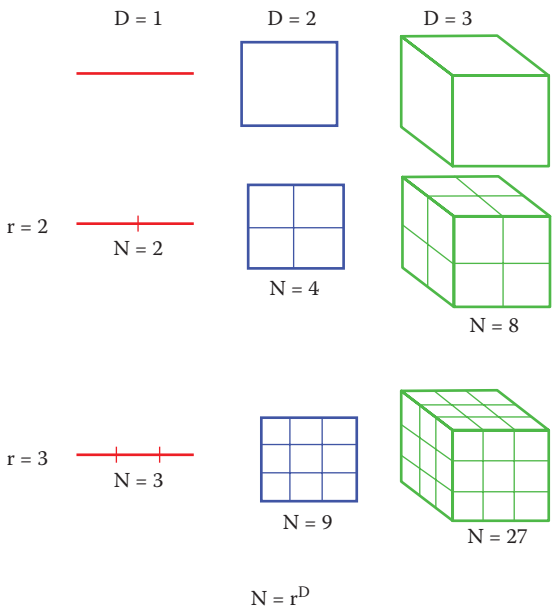


FIGURE 9.32 Fractal dimensionality: the Hausdorff dimension.

The point that emerges is that if dimensionality, *D*, is examined this way, *D* need not be an integer, as it is in Euclidean geometry. It could be a fraction, as it is in fractal geometry. This approach has proved very useful for describing all types of natural objects and for evaluating trajectories of dynamic systems.

It was Lewis Fry Richardson who first noted the regularity between the length *L* of coastlines, rivers, continental boundaries, etc., and the scale size *r*: when plotted in a log-log plot, a linear relation is obtained between log [*L*(*r*)] and log(*r*). Mandelbrot assigned the term (1 – *D*) to the slope of such a log-log plot, or:

$$\log[L(r)] = (1 - D)\log(r) + b \tag{9.5}$$

where *D* is the fractal dimension and *b* a constant.

For Great Britain, the slope 1 – D = –0.24, approximately, or D = 1.24, a fractional value. The coastline of South Africa is very smooth, virtually an arc of a circle, and the slope is very near zero, and this gives a fractional dimension D = 1 – 0 = 1. This makes sense because the coastline is very nearly a regular Euclidean object, a line, which has dimensionality of 1. In general, the "rougher" the line, the steeper the slope and the larger the fractal dimension (the fractal dimension for the Nile delta is 1.4; that of the Amazon basin is 1.85).

Hausdorffian dimensionality is further illustrated by considering the Koch curve shown in the inset of the header of this chapter. To make a Koch curve, a one-dimensional (1D) line is buckled to take on some fractional dimension between 1D and 2D. We begin with a straight line of length 1, called the *initiator*. We then remove the middle third of the line, and replace it with two lines that each have the same length (1/3) as the remaining lines on each side. This new form is called the *generator*, because it specifies a rule that is used to generate a new form. The rule says to take each line and replace it with four lines, each one-third the length of the original. We can do this iteratively without end; the length of the curve increases with each iteration, and it asymptotically approaches infinite length. If we treat the Koch curve as we did the coastlines and continents, the relation between log[*L*(*r*)] and log(*r*) for the Koch curve gives us D = log(4)/log(3) = 1.26. It is a number that the Koch curve approaches

as it becomes ever more finely broken over all possible scales. A huge number of everyday phenomena do seem ruled by such power-laws rather than Gaussian statistics. Some are well-known examples, such as earthquakes and species extinctions. Fractal organization is to be found in all kinds of dissipative structure—structures that "chaotically" branch in order to dump a load of energy as fast as possible (see Chapter 3, "Self-Organization").

Fractals in Nature

Nature always strives for the highest efficiency in all its organisms, and fractal-like networks are nature's answer to this maximization problem, literally "giving life an additional dimension."[52] Fractals in nature represent the geometry of irregular shapes and chaotic systems: trees, rivers, brains, lungs, the cardiovascular system, clouds, and coastlines all can be described by fractals.

In nature, the fundamental equation that describes scaling of various variables with relation to body mass, we saw in Chapter 7, is an allometric equation of the following form:

$$V = aM^b \tag{7.1}$$

where V is the variable in question, a is a constant, and b is the scaling exponent (in Chapter 7 we used $y = bx^a$). The fascinating empirical observation that has perplexed many scientists and has been a source of much debate is the fact that the exponents (b in Equation 7.1) of many variables, including cellular metabolism ($b = -1/4$), heartbeat ($b = -1/4$), maximal population growth ($b = -1/4$), lifespan ($b = 1/4$), blood circulation ($b = 1/4$), embryonic growth/development ($b = 1/4$), metabolic rates of entire organisms ($b = 3/4$), cross-sectional areas of mammalian aortas ($b = 3/4$), and cross-sectional areas of tree trunks ($b = 3/4$) are multiples of 1/4 instead of 1/3.[52] Pure geometric scaling of area (l^2) and volume (l^3) leads to a scaling constant of 2/3, but as Max Kleiber discovered in the early 1930s, the metabolic rates of entire organisms (which should scale with area) scale with respect to body mass (which should scale with volume), with a scaling constant of 3/4. This relationship is now referred to as *Kleiber's law* (see Figure 7.4). We also saw that West et al.[52] proposed a general model for this extra "fourth dimension". By modeling the cardiovascular system as a fractal-like network that ends in terminal units (capillaries) of the same size, they have shown that minimizing the work that the heart performs in pulsatile systems leads to the proper scaling constants that are empirically observed.

Fractals in Electrochemistry

Just as nature has benefited from fractal structures in almost all of its organisms, biomimetic fractal designs in electrochemical devices such as power conversion devices and sensors can also lead to benefits in scaling.

Fractals are an optimal geometry for minimizing the work lost due to the transfer network while maximizing the effective surface area. In many electrochemical systems, it is advantageous to have a large surface-to-volume ratio, while needing to transfer the signal or power effectively to an electrical network. Electrochemical energy conversion devices, such as fuel cells and some types of batteries, as well as sensors, such as glucose sensors, are examples of such applications. Miniaturization of these electrochemical devices to create miniature electrochemical devices poses several problems due to scaling. One issue is that there is a problem of increasing internal resistance in miniature 3D devices. In cases where an electrode is composed of materials that are excellent electrical conductors, such as metals, the internal resistance contribution of the electrode is minimal and can be ignored, but if the electrode is composed of carbon or another material that is a optimal electrochemical material, but not necessarily an optimal electrical material, the internal resistance contribution due to the electrodes can be significant.[53]

The resistance of a bulk or porous network scales inversely to the length scale l if scaled isometrically, as illustrated in Figure 9.33a. Figure 9.33b illustrates the scaling involved when scaling area with a constant thickness. Since the volumes of the structures can be written as follows:

$$V = l^3 \qquad \text{for Figure 9.33a} \tag{9.6}$$

$$V = tl^2 \propto l^2 \qquad \text{for Figure 9.33b} \tag{9.7}$$

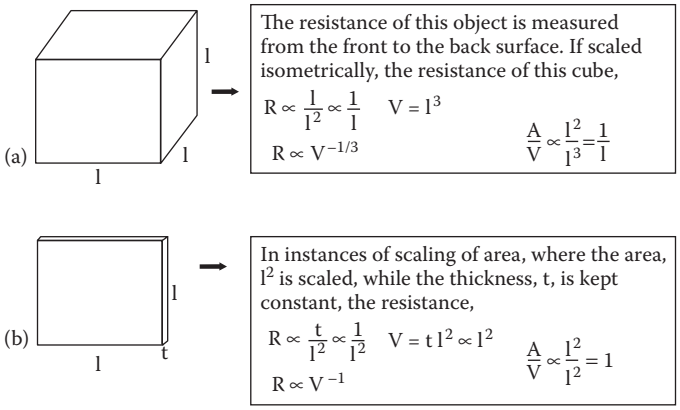


FIGURE 9.33 Scaling of a cube (a) and a sheet (b).

From this figure it can be concluded that resistance of a bulk material scales as follows:

$$R \sim V^{-1/3} \qquad \text{for Figure 9.33a} \qquad (9.8)$$

when scaling isometrically and as:

$$R \sim V^{-1} \qquad \text{for Figure 9.33b} \qquad (9.9)$$

when scaling area only.

The surface area and the scaling of the surface-to-volume ratios of the two cases in Figure 9.33 are as follows (assuming $t \ll l$).

$$\frac{A}{V} \propto \frac{l^2}{l^3} = \frac{1}{l} \qquad \text{for Figure 9.33a} \qquad (9.10)$$

and:

$$A = 4lt + 2l^2 \approx 2l^2 \text{ and } \frac{A}{V} \propto \frac{l^2}{l^2} = 1 \qquad \text{for Figure 9.33b} \qquad (9.11)$$

The surface-to-volume ratio of a bulk material is inversely proportional to the length scale ($1/l$) when scaling isometrically in three dimensions and is invariant (1) when scaling in two dimensions.

In the case of conventional Li-ion batteries and fuel cells, where the dimensions of the fundamental building blocks (the anode/electrolyte/cathode sandwich) are smaller than the final device dimensions and need to be scaled up, areal scaling is advantageous. As can be seen from Equations 9.8 to 9.11, 2D scaling from a small fundamental building block results in lower internal resistance and higher effective area compared with an isometric three-dimensional scaling approach. Indeed, commercial Li-ion batteries and fuel cells are composed of many layers of anode, cathode, and electrolyte that are rolled, folded, or stacked on top of each other. In the case of miniature power devices, where conventional methods of rolling, folding, and stacking are not applicable, microfabrication deposition techniques have been used to create thin-film batteries. Although applicable in some circumstances, there are instances where a 3D package is needed. As mentioned by Long et al.[7], the limiting factor of increasing the height of 3D miniature batteries is the increased internal resistance of the high-aspect-ratio structures. The use of better electronic and ionic conductors, in addition to the optimal geometries, determines the practical upper limit of these devices. This author believes that fractal networks can prove to be a more favorable architecture for miniature electrochemical devices. Other than the purely electrical advantages, fractal architectures provide a middle ground between high-surface-area electrodes composed of nanomaterials and solid electrodes. Although the high surface area of nanomaterial electrodes does increase performance in many areas, much of the volume of the entire battery can be "wasted" because of the electrolyte around the electrode. In a fractal electrode, much of the volume that would be composed of electrolyte is replaced with electrode material.

The approach used by West et al.[52] to model the fractal nature of the cardiovascular system is mirrored and directly converted to that of a fractal network of electrically conductive material in the model of Park et al.[53] The purpose of this exercise was not to model a specific system but to provide an idealized zeroth-order approximation for a fractal electrochemical system.

There are three general assumptions made by West et al.[52] The first assumption is that a fractal-like branching network is required to supply a volume effectively. In an electrochemical device, it is important to maximize the use of space. For example, fuel cell engineers strive to maximize the density of triple phase boundaries (most of the time by increasing the effective surface area), and batteries and sensors benefit from large surface-to-volume ratios because, in many cases, the chemistry is largely a surface phenomenon (limited by a process near the surface).

In situations where the reaction is limited by mass transfer, the mass transfer problem can be alleviated by reducing the distance the reactants need to travel, i.e., by spreading the reaction sites throughout the entire volume of the device. Just as capillaries are spread out throughout an organism because each capillary can only supply a certain volume of space with oxygen and nutrients, it is beneficial to spread out the reaction surface throughout the volume in an electrochemical device. It is possible to spread out the effective surface area throughout a 3D space using a sponge-like or web-like material composed of fibers or particulates, but in many cases, it becomes difficult to transfer the charge from and to the reaction sites.

Space-filling fractal networks can work to ensure that there is efficient charge transfer from a huge effective surface area to a current collector. The second assumption is that the smallest fractal structure is a size-invariant unit independent of the total volume. There are physical and practical limitations in size for all materials, even though these limits may differ from material to material. Nanoparticles, nanofibers, nanotubes, and other nanoscale materials are all examples of a fundamental building block. The final assumption is that energy used to distribute resources should be minimized.

Figure 9.34 illustrates the fractal electrochemical network with three branches (fractal levels) and a branching factor, n, of two as assumed by our team for our fractal model. N_k is the number of elements at fractal level k, and each level has $N_k = n^k$ elements; s is the fractal level for the smallest elements. J_{limit} is the maximum current density of any element, and R_k, ρ, l_k, A_k, r_k, v_k are the resistance, resistivity, length, area, radius, and service volume, respectively, for a fractal element at level k. A cylindrical geometry for each element is assumed.[53]

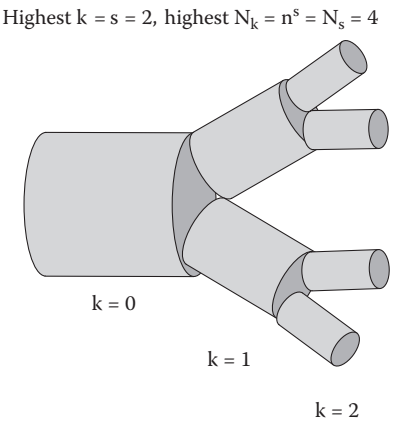


FIGURE 9.34 A simple fractal electrode.[53]

In electrochemical devices, there is loss of energy and an associated voltage drop due to the internal resistance. Minimization of this internal resistance is key to achieving efficient power transfer. This becomes more crucial as the current to the electrical load becomes larger because of Ohm's law:

$$E = IR \tag{9.12}$$

where E is the electrical potential, I is the current, and R is the resistance. In situations where the electrochemical system is limited by surface phenomena, there is a limiting interface current density J_{limit}. The total current, I_{total}, flowing through an electrode with a total surface area (total exchange area) of A_{total} operating in a surface-phenomena-limited regime is as follows:

$$I_{total} = J_{limit} A_{total} \tag{9.13}$$

We define the fractal level with the smallest elements as level s and N_k to be the number of elements at level k. We derived that the total exchange surface area, A_{total}, is proportional to the number of the smallest elements, N_s:

$$I_{total} \propto N_s \tag{9.14}$$

The total current is thus proportional to the number of the smallest elements within the fractal network. Obviously, maximizing the number of small elements within a fractal structure is desired. We also derived the following:

$$R_{total} \propto \frac{1}{N_s} \tag{9.15}$$

Again, as with Equation 9.14, maximizing the number of the smallest elements is desirable. This ensures that current is maximized, while the internal resistance is minimized. From Ohm's law and Equations 9.14 and 9.15, it can be concluded that if a fractal network geometry is used, the voltage drop across the entire structure is as follows:

$$E_{total\ internal\ voltage\ drop} = I_{total} R_{total} = \text{constant} \tag{9.16}$$

The power dissipation, in turn, scales with N_s:

$$P = E_{total}\, I_{total} = I^2_{total}\, R_{total} \propto N_s \qquad (9.17)$$

We also calculated that V_0, the volume of the liquid where effective exchange occurs with the electrode, the exchange volume can take place as given by the following (for details see Park et al.[53]):

$$V_0 \propto N_s \qquad (9.18)$$

Combining Equation 9.18 with Equations 9.14, 9.15, 9.16, and 9.17, one can then summarize the volumetric scaling of the electrical characteristics of a fractal network:

$$I_{total} \propto V_o \qquad (9.19)$$

and:

$$R_{total} \propto \frac{1}{V_o} \qquad (9.20)$$

So that $E_{total\ internal\ voltage\ drop} = I_{total}\, R_{total} = \text{constant}$ again, and:

$$P \propto V_0 \qquad (9.21)$$

We also obtained the following:

$$\frac{A_{total}}{V_0} \approx \text{constant} \qquad (9.22)$$

In other words the surface-to-volume ratio of a fractal electrode does not change as the volume changes.

Comparing Equation 9.20 with Equation 9.9, it can be concluded that the resistance of fractal electrodes scale similar to a thinly layered structure, not to an isometrically scaled volumetric geometry, and comparing Equation 9.22 with Equation 9.11, the surface-to-volume ratio of a fractal electrode does not change as the volume changes; again the fractal geometry scales more like a layered film than a volumetric electrode. The scaling of the electrical properties of a fractal electrode is summarized and compared with that of a 2D or 3D electrode in Table 9.9.

Table 9.9 Scaling of a Fractal Compared with Two-Dimensional and Three-Dimensional Scaling

Fractal	3D Scaling	2D Scaling
$I_{total} \propto V_0$		
$R_{total} \propto \frac{1}{V_o}$	$R \propto V^{-1/3}$	$R \propto V^{-1}$
$E_{total\ internal\ voltage\ drop} = \text{constant}$		
$P \propto V_0$		
$\frac{A_{total}}{V_0} \approx \text{constant}$	$\frac{A}{V} \propto \frac{l^2}{l^3} = \frac{1}{l}.$	$\frac{A}{V} \propto \frac{l^2}{l^2} = 1$

C-MEMS Fractals

The author believes that C-MEMS[54–57] technology is an attractive path toward the fabrication of 3D fractal-like carbon electrodes. C-MEMS describes a fabrication technique in which carbon devices are made by treating a prestructured organic material with high temperatures (typically 900°C and higher) in an inert or reducing environment. The details of the fabrication process are described in Chapter 5.

Fabricating a multilevel fractal structure that is idealized by the model that we have described above (and shown in Figure 9.34) is not a trivial issue. One reason behind the difficulty of this task is that the tools that are used for 3D microfabrication do not apply over a large range of length scales. Hybrid 3D fabrication techniques that combine different fabrication methods are therefore required (see Figure 3.11). An example hybrid method is briefly described below for the construction of a 3D fractal-like C-MEMS structure.

Example 9.3: Doped Photoresist Method to Make Fractal-Like Carbon Structures

In this method, standard photolithography is used on a photoresist (SU-8) premixed with carbon nanofibers (CNFs). The CNFs (Hawaii Industrial Labs) are first mixed with the SU-8 100 resin (1% by weight). The doped photoresist is then spin-coated onto a silicon wafer to an approximate 100-μm thickness. The standard photolithographic process to pattern the desired cylindrical photoresist posts then follows. After development, the posts exhibit a few clumps of CNFs on their surfaces that are partially embedded inside the SU-8. Pyrolysis of these structures converts the entire structure into carbon, with a thick coat of clumps of CNFs now covering the side walls of the posts. The CNFs shown can withstand the pyrolysis process and stay attached to the carbon posts. The process of all the

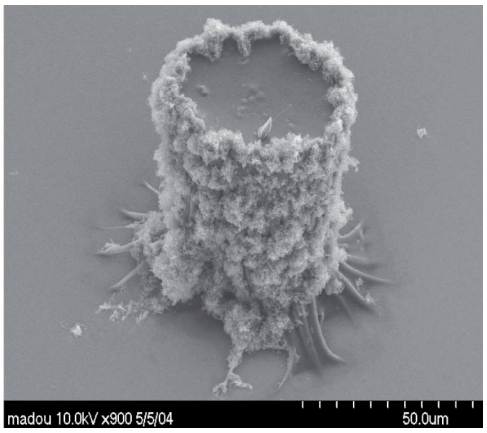


FIGURE 9.35 Scanning electron microscope image of a carbon post with nanofibers.

nanofibers migrating to the cylinder walls during pyrolysis is not yet understood. Figure 9.35 shows a scanning electron microscope image of this fractal-like structure. This is a direct attempt at mimicking the branching that we have mathematically modeled above (see Figure 9.34). Obviously this represents a rather poor fractal-like structure, with only two possible length scales, i.e., 100 μm (for the posts or tree stems defined by lithography) and nanometer-sized CNFs (the "branches").

For a sensor application, a single tree-like carbon microelectrode would already make for a big improvement over current sensor electrodes since it comes with the largest surface area while keeping the internal resistance minimal. Such a tree-like structure could be decorated with catalyst molecules like the decorations on a Christmas tree.

As illustrated in Figure 9.36, plant matter can be converted into carbon, much as charcoal is formed. One could then envision the use of natural occurring organic matter-based fractals to make carbon fractals.

In order to create a truly C-MEMS biomimetic energy conversion device, the anode and cathode, both fractal structures, must be intertwined while maintaining a pinhole-free electrolyte separation between the two in a manner similar to the blood vessels and air pathways in a mammalian lung. This may prove to be the most difficult task in arriving at a fully C-MEMS fractal battery or fuel cell.

Brains in Miniature Devices

Introduction

Before considering the implementation of computing power or brains in miniaturized systems, such as in microrobots, it is a good idea to briefly look at the state of the art in computer development in general and compare it with nature's arsenal using neurons and brains. After this introductory section, we consider the latest developments in artificial intelligence (AI). We discuss how both bottom-up and top-down approaches to AI might affect MEMS. Complexity theory and artificial life are considered next as two examples of bottom-up approaches to building AI that feature the self-organizing or emergent properties touched upon in Chapter 3. Artificial neural networks represent a bottom-up approach heavily dependent on learning cycles. Electronic noses and tongues have already been developed on the basis of such neural nets. As in the case of machining, we will learn that bottom-up approaches for embedding intelligence in MEMS are gaining more and more interest.

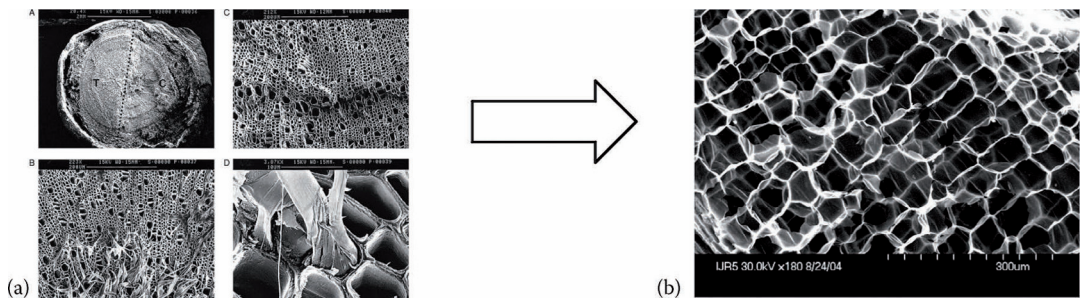


FIGURE 9.36 In the charcoal process, the structure of the plant cells is preserved. The high-temperature process is carried out in an oxygen-starved environment. (a) Twig of a tree with plant cells. (b) Charcoal. The original plant cell walls are preserved in the carbon.

State of the Art in Computer Development

IBM's ASCI White is a 12.3 trillion calculations per second machine the size of two basketball courts; it weighs 106 tons and cost $110 million. The machine runs programs on 8192 central processing units (CPUs) in parallel, resides at the Lawrence Livermore National Laboratory, and is the property of the US Department of Energy. Today, the most ambitious development effort to reach petaflop computers (*flops* is an acronym for floating point operations per second) is being carried out by a consortium headed by Caltech. The new system will have 150-GHz chips, cooled to near absolute zero. Because of their size and cost, such systems obviously remain out of range for MEMS. Closer to home are PCs. Intel's 0.18-μm manufacturing process is used to produce Coppermine Pentium III chips for PCs, which, when working at full speed, are as fast as 1.5 GHz. As we have seen in Volume II, Chapters 1 and 2, Moore's law will continue to hold for the next three to four MOSFET generations, reducing any mainframe to PC status eventually. In Volume I, Chapter 4, we saw how the fundamental thermodynamic limit of a MOSFET switch in a computer is the energy required to produce a binary transition that can be distinguished. The von Neumann-Landauer formula sets that theoretical limit (at 300K) as:

$$E_{bit} > kT \ln 2 \approx 18 \text{ meV}$$

(Volume I, Equation 4.70)

If a MOSFET switch consumes an amount of free energy E and performs N useful operations, the energy efficiency, we saw, is given by the following:

$$\eta = E/N \qquad \text{(Volume I, Equation 4.71)}$$

with η in units of operations per unit energy, or operations/second/watt. Using the thermodynamic energy limit per bit, this, in theory, enables an energy efficiency of $\eta = 3.5 \times 10^{20}$ operations/second/watt. In practice, using the case of 90 nm VLSI technology as an example, E is on the order of 1 fJ (femtojoule) and $\eta = 10^{15}$ operations/second/watt. This means that theoretically we can increase computing efficiency by a factor of more than 100,000. MOSFET downscaling is hampered by energy level quantization, though. The quantum mechanical limits on MOSFET scaling, based on the Heisenberg uncertainty principle, we summarized as follows:

$$x_{min} = \frac{\hbar}{\Delta p_x} = \frac{\hbar}{\sqrt{2m_e\ E_{bit}}} = \frac{\hbar}{\sqrt{2m_e\ k_B T \ln 2}} = 1.5 \text{ nm}$$

(Volume I, Equation 4.72)

which puts a fundamental limit on the integration density. For a limit on switching speed, we calculated the following:

$$\Delta E \Delta t \geq \frac{h}{2\pi} = \hbar$$

$$\Rightarrow \tau \sim 0.04 \text{ ps or } 25 \text{ Tbits/second}$$

(Volume I, Equation 4.73)

Thus, quantum mechanics allows for a very small switch, but other factors, such as a material's temperature tolerance range, will limit before this fundamental limit sets in. Maintaining Moore's exponential increase in transistor count, however, will become progressively more difficult and expensive. Cost may ultimately become more significant than technical boundaries, since it dictates every investment in research and development. Memory storage devices are improving even faster than ICs, but they will also encounter similar obstacles. This has fueled the drive to look beyond silicon and current transistor logic to totally new areas, such as optical, DNA, and quantum computing.

Radical modifications to current computer architecture are under consideration today. For example, it is the electron charge that is used in the conventional semiconductors that comprise the CPUs and dynamic random access memories (DRAMs) in computers today. For the nonvolatile memory, magnetic disk drives are used, and these rely on the spin state of electrons to store large amounts of information and in particular to preserve information when the computer is shut down. There is now the possibility of developing devices that simultaneously use the spin and charge of the electron to vastly increase the capability of information processing, in what are called *spintronics* (spin-based electronics) *devices*. All spintronic devices behave according to this simple scheme: 1) information is stored (written) into spins as a particular

spin orientation (up or down), 2) the spins, being attached to mobile electrons, carry the information along a wire, and 3) the information is read at a terminal. Spin orientation of conduction electrons survives for a relatively long time (nanoseconds, compared with tens of femtoseconds during which electron momentum decays), which makes spintronic devices particularly attractive for memory storage, magnetic sensors applications, and quantum computing, where electron spin represents a bit (called a *qubit*) of information. A qubit can exist not only in a state corresponding to the logical state 0 or 1 as in a classical bit, but also in states corresponding to a blend or superposition of these classical states. In other words, a qubit can exist as a zero, a one, or simultaneously as both 0 and 1, with a numerical coefficient representing the probability for each state. Four pairs of classical bits are needed to represent the same information as 1 pair of qubits. With eight qubits we can represent every number between 0 and 255 simultaneously. This allows for the compression of massive amounts of data. Quantum computing thus takes advantage of the concepts of superposition of states and entaglement introduced in Volume I, Chapter 3. This gives rise to many more states of the system, and thus in principle the ability to represent much more information without increasing the number of components—providing that it is possible to devise a way for the states of the system to interact quantum mechanically.[58]

Quantum computers with a handful of qubits have been made, but it seems hard to envision one with more than ten qubits. These early prototypes tend to suffer from decoherence after around 1000 operations. An interesting approach is to work with electrons on nanodots instead of on individual atoms as qubits. Nanowires are used to connect these nanodots and provide a physical entanglement. Providing entanglement through a physical connection between individual atoms is impossible, but between nanodots it can be done and it might avoid decoherence.

Research in these computing alternatives is well under way and even though non-silicon-based processors will not appear for many years to come, it seems quite certain that information technology will not be stopped when Moore's law ceases to be valid. For integrating computing power in MEMS, there are already good solutions today. The evolution of ICs and microprocessors has miniaturized the size of the computing aspect to the point where ubiquitous computing (ubicomp) has become possible, but sensors, output devices, data storage, and power supplies have a lot of catching up to do. It is precisely in this domain—the coupling of computation to the physical world—that micromachining will make its biggest contribution yet. To appreciate how to program memory embedded in MEMS with powerful top-down algorithms or simpler bottom-up learning routines, we will now look at how brains in nature work and consider the latest developments in AI, complexity theory, and artificial life.

The human brain and senses, just like nature's manufacturing methods, evolved in a biological context and are concerned mainly with survival and reproduction. Computing needs in MEMS are different, so we do not want to blindly mimic the brain but study it to learn and adapt strategies that fit our needs.

From Neurons to a Brain

It is important to realize that in nature, the simplest forms of life already exhibit complex reflexive behavior without any nerve cells or a brain at all. Even nature's reflexive behavior is often more complex than what even our "smart" robots can do today. Thus, the topic deserves some attention. A very simple brainless and nerveless organism, such as a paramecium, moves toward food or a comfortable temperature zone, orients itself in gravitational and centrifugal fields, responds to sound waves, swims upstream in a water current and toward a cathode at low applied currents and toward the anode at higher currents, moves away from damaging UV radiation and away from a touch and unfamiliar chemicals. Insects with only a few hundred thousand neurons also manage to fly around in real time and manage to avoid obstacles, although their computers are immensely slow, so they obviously must be able to manage their intelligence in some other ways than our large computers do.

In biology, nerve cells connect sense organs to the brain, and the brain processes incoming signals into perceptions. Nerve cells also run the other way—from the brain to the sense organs, enabling the brain to fine-tune its senses to pick up whatever it considers most important; in other words, perception is a two-way street. The brain itself relies on 10^{10} interconnected neurons (Figure 9.37) in 3.5 pounds or 15 cm^3 of gray matter. These neurons act like armies of switches, interconnected at 10^{14} synapses and collectively firing 10^{16} times a second while consuming less power than an ordinary light bulb, making today's attempts at artificial neural networks (see below) a pale version of their natural counterpart. An individual nerve cell does not posses a smidgen of intelligence; the latter emerges only from the interaction of billions of nerve cells. Nerve cells form long, thin fibers that carry electrical signals. The signals involve ions, not free electrons. The ions do not travel along the nerve cell but across its cell membrane. Electrical activity at one position along the nerve triggers activity in the next position. A charge pulse travels along the axon of a neuron (Figure 9.37) at a speed of 300 feet per second (200 miles per hour). This is slow compared with the speed of an electrical pulse in a wire and a million times slower than that in a Si-based switch (milliseconds versus nanoseconds).[59] Nature obviously does not get its advantage from speed; people are not able to learn, and later remember, more than about 2 bits per second for any extended period. If you could maintain that rate for 12 h every day for 100 years, the total would be about three billion bits–less than what we can store today on a regular 5-in. compact disc (CD).[60] The merits of our brain stem from its 3D nature and interconnections of neurons, as well as their individual complexity. The number advantage of human brains might disappear soon enough. As depicted in Volume II, Figure 2.1, the total number of transistors on a 2D chip will reach the number of neurons in the human brain by the year 2010; computers, some predict, will achieve parity in memory capacity with the human brain shortly after (in Volume II, Figure 2.1, the year 2020 is projected for this event).[61] The area density of the neurons was actually surpassed in the mid-1980s.[62] Neurons are not limited to communicating in an on/off or fire/no-fire mode; they send information in the form of rate of firing, a myriad complicated mechanisms increase or decrease signal strengths at synapses, neurons constantly adapt and change (growth and movement of axons and dendrites), and the brain does not employ a single generic neuron.[59] So not only does nature have an advantage in total number of interconnected switches, the switching behavior of a neuron is also significantly more complex than that of a computer switch.

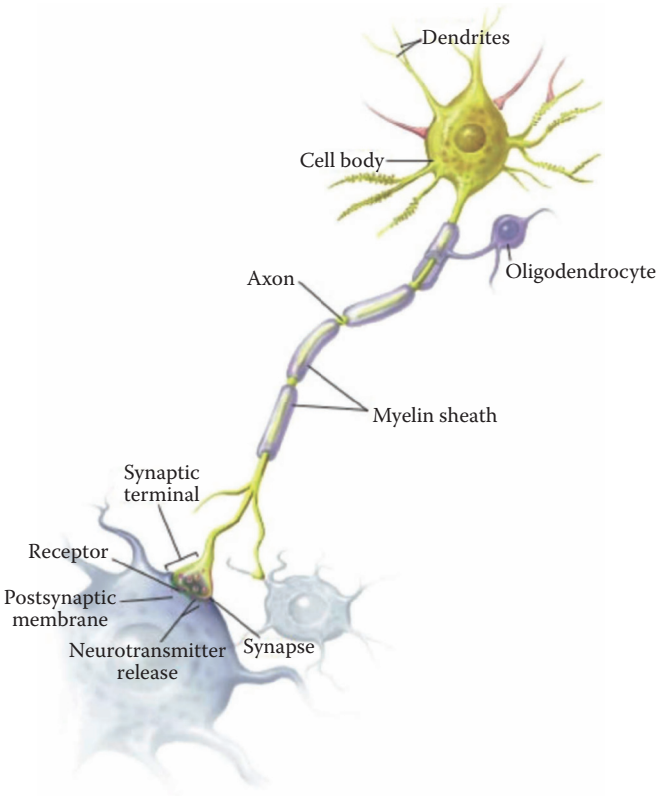


FIGURE 9.37 A neuron cell.

DNA constitutes the long-term read-only memory storage at the heart of life as it evolved on planet earth. A DNA strand with about 2×10^8 nucleotides is tightly packed into a volume of 500 μm^3 (see Figure 4.1). With the assumption that a set of three nucleotides is analogous to a byte, this corresponds to a memory volume density of 1.2 $Mb/\mu m^3$ and about 1 $Kb/\mu m$ in linear density, indicating that memory chips based on DNA as the active elements could have extremely high density. Transcription, or the process of reading

the DNA code, base by base, to make proteins, was illustrated in Figure 2.12. Both the packing density and the transcription process are perfect illustrations of the power of nanochemistry at work in nature. The evolutionary development of brains is very slow, though, and that of *Homo sapiens*—evolution's crown jewel (despite Howard Stern and Rush Limbaugh)—emerged only tens of thousands of years ago. According to Kurzweil's law of accelerating returns (Volume II, Chapter 2), the next salient events (e.g., "intelligent machines making other intelligent machines") will emerge much faster than mutations in our DNA can accommodate. For these next salient events, nature's DNA process is too slow: human intelligence has taken over already, and technological innovations will accumulate faster and faster.

DNA computing utilizes the information coding sequence of DNA and the interactions between complimentary strands to perform complex computations. DNA computing was first demonstrated by Adleman of the University of Southern California, in 1994.[63] The computation in Adleman's experiment ran along at 1014 operations per second, a rate of 100 teraflops, or 100 trillion floating point operations per second. Contrast this with the world's fastest supercomputer, NEC's Earth Simulator, running at just 35.8 teraflops. Obviously, computing with DNA has some advantages over current machines. Whereas today's computing technology rests on a highly linear principle of logic and one computation must be completed before the next can begin, the use of DNA means that an enormous number of calculations can take place simultaneously: the largest supercomputers have a few hundred processors run in parallel, but in some first DNA experiments there were 1 trillion processes running in parallel. This parallel power is many times faster than that of traditional machines—a mix of 1,018 strands of DNA could operate at 10,000 times the speed of today's advanced supercomputers.

In such a computer, the hardware consists of enzymes (restriction nuclease and ligase), a double strand of DNA as input data, and a few short DNA molecules as transition rules or software. The solutions come as pieces of DNA that must be isolated to be read. The process uses only one-ten-billionth of a watt versus tens of watts for an electronic microprocessor. There are some very serious limitations though. Although DNA can store a trillion times more information than current storage media, the way in which the information is processed necessitates a massive amount of DNA if larger-scale problems are to be solved. Furthermore, although the computation process takes place at an awesome speed, the "printout" of the result is excruciatingly sluggish, and involves many steps—it took Adleman a week of laboratory work to extract the potential solutions from his DNA cocktail.

Artificial Intelligence

Two Contenders: Top-Down and Bottom-Up

Introduction

Artificial intelligence is the field of research that attempts to embed intelligence into machines; it includes knowledge-based systems, pattern recognition, automatic learning, natural-language understanding, robotics, etc. Two distinct camps within the AI community today are the very ambitious and traditional, analytical top-down approach and the more modest, natural bottom-up approach. It is interesting to note that the same top-down and bottom-up duality showed up in manufacturing a couple of decades later. Traditionalists represent a rational approach to AI, building intelligence from logical first principles and complex algorithms. Research fields influencing this approach are semiconductor technology, cognitive psychology, and information science. Intelligence has proved very difficult (perhaps too difficult) to describe this way, and although the naturalist approach represents a lower level of computing, it has recently been more successful than traditional AI. Nature-inspired AI emphasizes neural network computing, complex system interactions, and self-organization. The natural approach to AI draws from research such as neuroscience, complex adaptive systems, and molecular biology. As in many research fields, rational statements about the relative merits of these two approaches are hard to find. Comparisons are based not so much on proven scientific facts as on the

powerful egos, reputations of their institutions, and the old-boy networks propagating their views and influencing government funding agencies; the latter often listen only to the "loudest" voice. From the status of the two contending approaches discussed below, it appears that the natural AI approach has been the more successful. The two AI camps are compared in Table 9.10.

Traditional AI

Traditional AI got its unofficial baptism in 1956 at a small conference in Dartmouth, New Hampshire. Present at the time were luminaries such as Marvin Minsky, John McCarthy, and Edward Feigenbaum, and although they were, as many visionary scientists tend to be, over-the-top optimistic, the fruits of AI are now being realized.[59] An example of the optimism reigning in the early days of AI was the prediction about the early successful realization of the Turing test. Alan Turing, mathematician and computer scientist, had predicted, in a 1950 paper, that by the year 2000 a human judge interviewing a computer and a human foil, both hidden from view, would not be able to differentiate the computer from the human when questioning both via a terminal. Although this has not yet materialized, Ray Kurzweil maintains that Turing was correct and might simply be off by about twenty years.[61]

Surprisingly, breakthroughs in AI did not come from software programs reasoning in a precise, highly ordered, step-by-step fashion, as the logicist originators of AI had predicted but from simpler, so-called *expert systems* that made headway in the early 1980s. The latter sort of AI represents a rather practical or heuristic approach to addressing questions in very specific fields, such as the stock market, geophysics, or diagnosis of trouble with humans or cars. These simpler programs rely on rules learned by experience or training fed into the computer. An expert system will typically have a large number of *if-then* rules, and because the computer can quickly analyze thousands of these rules, it can outperform its human experts easily in more and more fields. One prominent example of AI's accomplishments is IBM's Deep Thought computer. Deep Thought beat grand chess master Brent Larsen in 1988 and sailed past world champion Garry Kasparov in May 1997.[59] The chess program has a human-like ability to zero in on a few most promising pattern of moves but does not really have a deep understanding of chess strategy. This makes it more akin to the natural AI approach we review next. Another example is the huge data and rule base of Lenat's Cyc 9 (*Cyc* from *encyclopedia*).[59] Besides information, Cyc again incorporates many rules; half of its memory content is facts, the other half is rules. The highly touted, ten-year-long Japanese Fifth Generation Project, a program from the 1980s meant to catch Japan up with the United States in software development and based on traditional AI, is widely seen as a failure now. The logical and heuristic approach simply proved to be inadequate. Interestingly, the Sixth Generation Project, also known as the Real-World Computing Project, started in Japan in 1992, aimed to achieve human-brain-like computing capabilities by the year 2002. Also this milestone was not met. Clearly, natural AI is today a more favored approach than traditional AI.

Table 9.10 Two Directions in Artificial Intelligence (AI) Research

AI	**Traditional**	Semiconductor technology Information sciences Cognitive psychology
	Reductionist	
	Analytical	
	Top-down (see nanomachining)	
	e.g., rule-based expert systems	
	Natural	Molecular biology Neurosciences Complex adaptive systems
	Complex interactions	
	Bottom-up (see nanochemistry)	
	Self-organization	
	e.g., neural networks	

Marvin Minsky remains a strong proponent of the traditional approach to AI, and in *The Society of Mind* he tries to put the field back on track.[64] In a typical broadside against the bottom-up school, Minsky says, "Why bother building a robot that's capable of getting from here to there, if one gets there it can't tell the difference between a table and a cup of coffee?"[65] The book proposes that one should take multiple approaches to solving a problem. Lenat's Cyc, he claims, does not have enough different approaches, and neural networks are just too dumb.

Natural AI

In the early 1980s, some scientists recognized that traditional AI had been trying to solve problems too much from a top-down mode, whereas a bottoms-up mode might be more appropriate. The very complex programs of traditional AI had not been adapting to or learning from the world around them, an innate capability of even the simplest animals. As in manufacturing today, it appeared that nature might provide some guidance. Neuroscientists speculated that the processing of conscious information takes up as little as a thousandth of the human brain's power and that most of the rest goes into lower-level aspects of survival.[61] In other words, there are deep connections between the need to get food and mate (i.e., survival) and intelligence. It is not so surprising, then, that *Homo sapiens'* DNA is 98.6% the same as that of the lowland gorilla and 97.8% the same as that of the orangutan. On the basis of these observations, the neuroscientist Wilson introduced the concept of creating computer simulations or even simple robots that could avoid danger, find food, and deal in general with a more or less complex environment.[66] Wilson called his computer simulations *animats*. Brooks and coworkers at MIT have been implementing similar ideas into a series of small robots that interact and learn from their environment rather than being loaded up front with a blueprint of every possible scenario the robots might encounter. Sensor data from these robots are shared with all of the different microprocessors on the machines. Maes and coworkers enhanced this type of thinking by adding programs that enable the robot to learn from its mistakes by implementing "positive" and "negative" feedback as well as attempting programs for "motivation," such as "aggression," "curiosity," and "hunger."[66] Brooks's robots are a far cry from von Neumann's self-replicating robots but probably closer to what we can expect to accomplish in this decade with smart MEMS.

Complexity Theory

The science of complexity is preoccupied with the investigation of emergent, self-organizing properties in complex systems. The term *emergent* is used for phenomena in which the whole is greater than the sum of the parts; the whole system exhibits collective behavior that seems not to be built into the individual components in an obvious or explicit manner. In Chapter 3 we saw how Ilya Prigogine with his theory of "dissipative structures" was the first to describe, in a rigid mathematical framework, the coexistence of structure and change, balance and flow, stillness and motion, stability and instability, order and disorder, equilibrium and nonequilibrium, and being and becoming characterizing complex systems.[67] The name *dissipative structures* conveys the notion of dissipation on the one hand and order or structure on the other. Prigogine's monumental and revealing theory is exemplified by living organisms maintaining themselves in a stable state far from equilibrium despite continual flow of matter and energy (metabolism) through their system. Systems far from equilibrium cannot be described by classical thermodynamics, which applies only to systems close to equilibrium. The linear mathematics applicable for small flows (also called *fluxes*) characteristic of systems close to equilibrium must be replaced by nonlinear expressions describing the large fluxes associated with nonequilibrium systems. Dissipative structures do not tend toward equilibrium; on the contrary, they exhibit instabilities, bifurcation points, leading to new forms of order that move the system farther and farther away from the equilibrium state with an accompanying increase in complexity. The greater the complexity, the higher the degree of nonlinearity. The behavior of such a nonequilibrium system cannot be described from the individual parts; instead it behaves as an interconnected whole with multiple feedback loops. Nonlinear equations describing dissipative systems usually have more than one

solution, and as a consequence, new situations may emerge at any moment. This contrasts with classical thermodynamics, where a system is completely predictable and pinned down by one solution only. From Prigogine we learn that the solution or path taken at an instability or bifurcation point in a dissipative structure depends on the previous history of the system and that its behavior is no longer universal but quite unique to the individual system itself. Moving away from equilibrium, we thus move away from the universal to the unique, from determinacy to indeterminacy, from individual components to an interconnected structure with links between structure and its history and a richness and variety as embodied by life itself.[68] Indeterminacy in dissipative structures is also introduced through small random variations, that is, "noise" from the environment. At bifurcation points, the path taken depends very sensitively on such noise, and the outcome becomes even more unpredictable, with new structures and higher order and complexity emerging spontaneously.

That order increases at bifurcation points does not contradict the second law of thermodynamics, as the increased order comes at the expense of greater disorder of the environment. Contrary to our intuition, chaotic systems are highly organized, exhibiting complex patterns dividing and subdividing again and again at smaller and smaller scales. In fluidic instabilities, turbulence will appear at sufficiently large velocities. In the case of chemical reactions, instabilities appear through repeated self-amplifying feedback. In both cases, self-organization, the spontaneous emergence of order, results from the combined effects of nonequilibrium, irreversibility, feedback loops, and instability. In fluidics, an example is the bathtub vortex, and in chemistry, examples include the so-called *chemical clocks* and *catalytic networks of enzymes*. The latter were shown to not only self-organize but also self-reproduce and evolve (see Chapter 3). In physics, an example is laser light, in which, under special circumstances, far from equilibrium and by "pumping" in energy from the outside, coherence spontaneously emerges from incoherent light. The most spectacular example of a self-organizing systems is the earth itself, as detailed in Lovelock and Margulis' Gaia theory.[69] Although often under attack, this theory is gaining more acceptance.

The reductionist approach of traditional AI fails when problems become extremely complex, and often no understanding is gained by breaking them down into their component parts. Earlier we saw how individual neurons or nerve cells work; not one exhibits any intelligence by itself, but intelligence in our brain emerges as a consequence of the interactions of billions of them. This is true for chemical and biological interactions that go on in living organisms, social structures, the weather, the economy, etc. Life itself, according to Kauffman, is nothing but an emergent property of a certain kind of complex system.[70] Although complexity theory seems to have made impact on market prediction, its impact on understanding life and evolution is still a matter of debate; this debate will probably go on for some time (see also Volume I, Chapter 7). Whereas natural AI is already useful in designing smart MEMS, complexity theory will start becoming important only when we work in the nanodomain and deal with self-organizing molecular systems.

Artificial Life

Artificial life, or a-life, is the study of lifelike creatures built by humans. This new science was officially inaugurated in Los Alamos, New Mexico, at a 1987 gathering organized by Christopher G. Langton.[66] Computer-generated entities or cellular automata behaving as self-operating machines, processing information about their surroundings, proceeding logically, and replicating themselves, were already predicted by the legendary mathematician von Neumann in the late 1940s. Today, many creatures of this type—really nothing more than software code—have been demonstrated, even on laptop computers. Some practitioners believe that by watching these creatures evolve on the computer screen and subjecting them to a variety of stresses, we can learn important lessons about the evolution of real-life systems more quickly and less expensively than with standard experimentation. Early a-life programs include *Boids* (http://www.lifeartificial.info/flocking-boid-behavior), which exhibit flocking behavior uncannily similar to that of real birds. The simple rules in this program control only individual Boids,

but somehow flocking group behavior emerges (see complexity theory).[71] *Network Tierra* by Thomas Ray features software simulations of organisms—*creatures* in which each *cell* has its own DNA-like genetic code.[71] *Tierra* can be downloaded from http://www.hip.atr.co.jp/~ray/tierra/tierra.html. *Tierra* organisms, in their simulated world, compete for a limited resource (energy) provided by the computer CPU. The seed electronic organism, the Ancestor in *Tierra*, has only three genes, and besides feeding it is designed to replicate itself and to mutate. Several sources of mutations are programmed in. There is also a *reaper* that kills off creatures that are less well adapted and those that are too old. To everyone's surprise, *Tierra* very quickly developed some lifelike evolution characteristics, such as the development of new species, parasites, defense mechanisms against parasites, symbiotic relationships, competition for food, and even something akin to punctuated equilibrium as proposed by Gould and Eldredge in 1972.[70] The discussion about how increasing complexity arose is an ongoing one. In the case of *Tierra*, there was no tendency toward increasing complexity; on the contrary, simpler creatures evolved over time. Emergence of complexity is a fact of life, though, and Ray is trying to induce it in *Network Tierra*. For evolution to proceed in this artificial world, it has become clear that a lifelike complexity must challenge the organisms. Ray is doing this now by letting his organisms grow on the Internet with its diverse electronic life. The organism, like a screensaver, becomes active only when the host computer is inactive. Ray speculates on the emergence of such intelligent life in *Tierra*-like environments; it will make the Turing test irrelevant. For interested readers, more on artificial life can be found at Avida Artificial Life Group (http://www.krl.caltech.edu/avida), Boids (http://hmt.com/cwr/boids.html), Complexity On-Line (http://www.csu.edu.au/complex), Dawkins (http://www.spacelab.net/~catalj), Primordial Soup Kitchen (http://www.psoup.math.wisc.edu/kitchen.html), Stuart Kauffman (http://www.santafe.edu.edu/People/kauffman), Swarm (http://www.santafe.edu/projects/swarm), and *Tierra* (http://www.hip.atr.co.jp/~ray/tierra/tierra.html). As in the case of complexity theory, artificial life is bound to become more important as we move from MEMS to NEMS.

Artificial Neural Network Software

As we have seen, neurons, to a first-order approximation, work by gathering signals from other neurons, weighing each by varying algebraic amounts, and then internally summing all the inputs. If this cumulative signal exceeds some internally stored threshold value, the neuron fires its own signal that is subsequently gathered in by other downstream neurons. John Hopfield, in 1982, invented a simple mathematical model of a network of nerve cells, called a *neural net*. As a quantum physicist Hopfield asked himself if a neuron in a brain might not act like an atom in a lattice. He reasoned that like an atom can take on discrete states, a neuron might also be found in discrete states: it can fire or not fire. His idea boiled down to the simple idea that just as atoms in a crystal arrange themselves to minimize the energy, so must a neural network minimize its "energy." This led to his celebrated 1982 paper, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." He showed that if many neural units were hooked together, they acquired computational abilities. The artificial neuron is a mathematical construct that emulates the more salient functions of biological neurons, i.e., the signal integration and threshold firing behavior. The most basic neural software consists of a layer of input and output *nodes*. When the input nodes receive a strong enough signal they "fire," sending their own signals to the output nodes attached to them. The input nodes in such a simple system represent an array of input stimuli, and a response is picked up from the output nodes. To make the neural network produce the right answer, an outside controller, usually a human or a computer programmed with the correct database, must adjust the switches between the input and output till the output produces the right answer. In neural network jargon, these settings are called *weights*. In the simplest case, we make the weights either one or zero; in general, though we set the weights to any of a range of signal pass-through strengths. Increasing the weight between an input and output node means that the input node signal will make a larger contribution to the likelihood that the connected output node will fire. This corresponds to an

excitatory signal in a nerve cell. A decrease in weight corresponds to an inhibitory signal and can stop a branch in the net from cascading throughout the net or shut off a nerve cell in the biological equivalent. The intelligence within the neural network is stored within these sundry algebraic connection weights. The process of teaching the network the appropriate weights is a complicated and often lengthy process. Frank Rosenblatt designed the first neural network—a so-called *perceptron*—that could be trained by consecutive small adaptations of the weighing factors.[59] This was achieved by comparing the results of consecutive network attempts at the right answer with the desired results and making small changes to the weighing factors after each comparison; more dramatic adjustments would have wiped out all previous training and give the wrong answers for any other pattern recognition.

Using a Hopfield-style neural net, Sahley and Gelperin worked out a model that faithfully captured the learning a common garden slug (*Limax maximus*) displayed when confronted with smells such as that of quinine (which it dislikes) and carrots (which it likes).[70] Their model neural net is shown in Figure 9.38.

Neural nets have been used in sensor array applications such as artificial noses and tongues.[72] In these instruments, a small number of odor and taste sensors are trained to smell and taste food and drinks containing hundreds of different chemicals. Making selective sensors for each separate chemical would be impossible, but a limited sensor array develops specific patterns (fingerprints) for complex chemical mixtures that correlate well with the human sensory evaluation. This approach has taken a rap for being a "black box"—it gives the right results but does not explain why. Indeed, neural net software packages develop connection traces internally that embody the rules behind the conceptual space on which they are training. For many chemical and biological applications, however, it is not important to know the identity of the individual components of a complex mixture, but it is important to sound an alarm if an unusual pattern is detected; for this application, natural AI seems the appropriate way to embed intelligence.

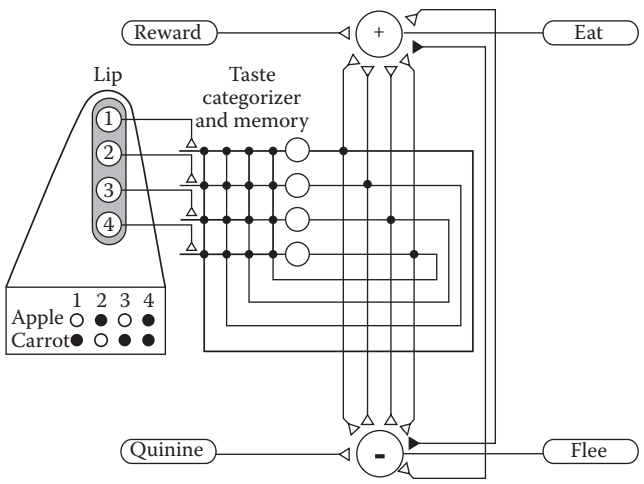


FIGURE 9.38 Model neural network for learning in the slug.

Acknowledgments

With special thanks to Ramses Madou and Drs. Rabih Zaouk and Benjamin Park.

Questions

9.1: What is meant by bottom-up and top-down artificial intelligence approaches? Which approach makes most sense for the development of smart MEMS?

9.2: Compare the function of a set of neurons in the brain with an artificial neural net. Find some real-life examples where neural nets are used today.

9.3: Explain how μ-3D batteries are better than thin-film batteries for powering a MEMS device.

9.4: Sketch a rough Ragone plot for capacitors, supercapacitors, batteries and fuel cells.

9.5: What MEMS power would you put into a microrobot? A microrocket? A microsubmarine? A microbutterfly?

9.6: Discuss some options for hydrogen storage. Why is it so important for the future of fuel cells?

9.7: As far as fuel cells are concerned what is the merit of biofuel cells?

9.8: Explain for what uses you might prefer a supercapacitor over a Li-ion battery.

9.9: What is the role of a bipolar plate in a fuel cell? How could C-MEMS make for better bipolar plates?

9.10: What are some of the promises and challenges involved with building micromachined combustion engines?

9.11: In what ways can MEMS and NEMS improve solar cells?

9.12: Explain with a diagram how a Grätzel solar cell works.

9.13: What are the fundamental limits in computing power? Use equations.

9.14: Explain how a fractal design of the electrodes might benefit a fuel cell.

9.15: Explain the scaling behavior of a cube and a sheet in terms of resistance and surface to volume ratio.

References

1. Malone, M. S. 2004. Moore's second law. *Wired*, April. http://www.wired.com/wired/archive/12.04/start.html?pg=2.
2. Chesnokov, S. A., V. A. Nalimova, A. G. Rinzler, R. E. Smalley, and J. E. Fischer. 1999. Mechanical energy storage in carbon nanotube springs. *Phys Rev Lett* 82:343–46.
3. Nalimova, V. A., D. E. Sklovsky, G. N. Bondarenko, H. A. Gaucher, S. Bonnamy, and F. Beguin. 1997. Lithium interaction with carbon nanotubes. *Syn Met* 88:89–93.
4. Gaucher, H., E. Frackowiak, S. Gautier, S. Bonnamy, and F. Beguin. 1999. Electrochemical storage of lithium in multiwalled carbon nanotubes. *Carbon* 37:61–69.
5. Bates, J. B., G. R. Gruzalski, and C. F. Luck. 1993. Rechargeable solid-state lithium microbatteries. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '93)*. Fort Lauderdale, FL, February 1993, pp. 82–86.
6. Goemans, P. A. F. M. 1994. Microsystems and energy: the role of energy. *Microsystem technology: Exploring opportunities,* ed. G. K. Lebbink. Alphen aan de Rijn/Zaventem: Samsom BedrijfsInformatie bv, pp. 50–64.
7. Long, J. W., B. Dunn, D. R. Rolison, and H. S. White. 2004. Three-dimensional battery architectures. *Chem Rev* 104:4463.
8. LaFollette, R. M., J. N. Harb, and P. Humble. 2001. Microfabricated secondary batteries for remote, autonomous electrical devices. *The Sixteenth Annual Battery Conference on Applications and Advances, 2001*. Long Beach, CA, January 9–12, 2001, pp. 349–54.
9. Koeneman, P. B., I. J. Busch-Vishniac, and K. L. Wood. 1997. Feasibility of micro power supplies for MEMS. *J Microelectromech Syst* 6:355–62.
10. Hart, R. W., H. S. White, B. Dunn, and D. R. Rolison. 2003. 3D microbatteries. *Electrochem Commun* 5:120–23.
11. Thomas, P., Y. Y. Fei, and X. D. Zhu. 2007. Adsorption and desorption of hydrogen on bare and Xe-covered Cu(111). *Appl Phys A* 86:115–21.
12. Sandronk, G. 1999. A panoramic overview of hydrogen storage alloys from a gas reaction point of view. *J Alloys Compounds* 293:887–88.
13. Schuth, F., B. Bogdanovic, and M. Felderhoff. 2004. Light metal hydrides and complex hydrides for hydrogen storage. *Chem Commun* 20:2249–58.
14. Orimo, S., and Y. Nakamori. 2004. Li-N based hydrogen storage materials. *Mater Sci Eng B* 108:48–50.
15. Dillon, A. C., K. M. Jones, T. A. Bekkedahl, and C. H. Kiang. 1997. Storage of hydrogen in single-walled carbon nanotubes. *Nature* 386:377–79.
16. Langmi, H. W., and G. S. McGrady. 2008. Ternary nitrides for hydrogen storage: Li–B–N, Li–Al–N and Li–Ga–N systems. *J Alloys Compounds* 466:287–92.
17. Park, S., D. Srivatava, and K. Cho. 2003. Generalized chemical reactivity of curved surfaces: carbon nanotubes. *Nano Lett*. 3:1273–77.
18. Li, J., T. Furuta, H. Goto, T. Ohashi, Y. Fujiwara, and S. Yip. 2003. Theoretical evaluation of hydrogen storage capacity in pure carbon nanostructures. *J Chem Phys* 119:2376–85.
19. Fernandez-Alonso, F., F. J. Bermejo, C. Cabrillo, R. O. Loutfy, V. Leon, and M. L. Saboungi. 2007. Nature of the bound states of molecular hydrogen in carbon nanohorns. *Phys Rev Lett* 98:215503.
20. Iijima, S. 1999. Nano-aggregates of single-walled graphitic carbon nano-horns. *Chem Phys Lett* 309:13.
21. Sakai, H., Y. Tokita, and T. Hatazawa. 2007. A high-power glucose/oxygen biofuel cell. *Fuel Cell Technology: Biofuel Cells, Enzymatic and Microbial, Division of Fuel Chemistry, 234th ACS National Meeting*. Boston, MA, August 19–23, 2007, p. 196.
22. Admin. 2007. http://www.ecoustics.com.
23. Meyers, J. P., and H. L. Maynard. 2002. Design considerations for miniaturized PEM fuel cells. *J Power Sources* 109:76–88.
24. Park, B., and M. J. Madou. 2006. Design, fabrication, and initial testing of a miniature PEM fuel cell with microscale pyrolyzed carbon fluidic plates. *J Power Sources* 162:369–79.
25. Sato, H., C. W. Berry, B. E. Casey, G. Lavella, Y. Yao, J. M. VandenBrooks, and M. M. Maharbiz. 2008. A cyborg beetle: insect flight control through an implantable, tetherless microsystem. *IEEE 21st International Conference on Micro Electro Mechanical Systems 2008*. Tucson, AZ, January 13–17, 2008, pp. 164–67.
26. Varia, A. 2003. Thermal model of MEMS thruster. *FEMCI Workshop 2003*. NASA Goddard Space Flight Center, May 7–8, 2003.
27. London, A. P. 2000. *Development and test of a microfabricated birpropellant rocket engine,* PhD diss. Cambridge, MA: Massachusetts Institute of Technology.
28. Protz, C. S. 2004. *Experimental investigation of microfabricated bipropellant rocket engines,* PhD diss. Cambridge, MA: Massachusetts Institute of Technology.
29. Sniegowski, J., and E. Garcia. 1995. Microfabricated actuators and their application to optics. *Micro-Optics/Micromechanics and Laser Scanning and Shaping (Proceedings of the SPIE)*. San Jose, CA, February 7–9, 1995, pp. 46–64.
30. Garcia, E., and J. Sniegowski. 1995. Surface micromachined microengine. *Sensors Actuators A* A48:203–14.
31. Lee, J. B., Z. Chen, M. G. Allen, A. Rohatgi, and R. Arya. 1995. A miniaturized high-voltage solar cell array as an electrostatic MEMS power supply. *J Microelectromech Syst* 4:102–08.
32. Salalibara, T., H. Izu, T. Kura, W. Shinohara, H. Iwata, S. Kiyama, and S. Tsuda. 1995. High-voltage photovoltaic micro-devices fabricated by a new laser-processing. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '95)*. Amsterdam, the Netherlands, January 29–February 2, 1995, pp. 282–87.

33. Kimura, M., N. Miyakoshi, and M. Daibou. 1991. A miniature opto-electric transformer. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '91)*. Nara, Japan, January 30–February 2, 1991, pp. 227–32.
34. 2008. School of Photovoltaic and Renewable Energy Engineering, University of New South Wales.
35. Mao, F., N. Mano, and A. Heller. 2003. Long tethers binding redox centers to polymer backbones enhance electron transport in enzyme "wiring" hydrogels. *J Am Chem Soc* 125:4951–57.
36. Madou, M. J., and S. R. Morrison. 1989. *Chemical sensing with solid state devices*. New York: Academic Press.
37. Gratzel, M. 2001. Photoelectrochemical cells. *Nature* 414:15.
38. Gratzel, M. 2005. Solar energy conversion by dye-sensitized photovoltaic cells. *Inorg Chem* 44:6841–51.
39. O'Regan, B., and M. Gratzel. 1991. A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films. *Nature* 353:737–40.
40. Greenham, N. C., X. Peng, and A. P. Alivisatos. 1996. Charge separation and transport in conjugated-polymer/semiconductor-nanocrystal composites studied by photoluminescence quenching and photoconductivity. *Phys Rev B* 54:17628–37.
41. Huynh, W., J. J. Dittmer, and A. P. Alivisatos. 2002. Hybrid nanorod-polymer solar cells. *Science* 295:2425–27.
42. Nozik, A. J. 2002. Quantum dot solar cells. *Pysica E* 14:115–20.
43. Pietryga, J. M., R. D. Schaller, D. Werder, M. H. Stewart, V. I. Klimov, and J. A. Hollinasworth. 2004. Pushing the band gap envelope: Mid-infrared emitting colloidal PbSe quantum dots. *J Am Chem Soc* 126:11752–53.
44. Branham, K., J. L. Marques, J. Hassard, and P. O'Brien. 2000. Quantum-dot concentrator and thermodynamic model for the global redshift. *Appl Phys Lett* 76:1197.
45. Stark, I., and M. Strodeur. 1999. New micro thermoelectric devices based on bismuth telluride-type thin solid films. *Eighteenth International Conference on Thermoelectrics.* Baltimore, MD, Aug. 29–Sept. 2, 1999, pp. 465–72.
46. Shenck, N. S., and J. A. Paradiso. 2001. Energy scavenging with shoe-mounted piezoelectrics. *IEEE Micro* 21:30–42.
47. Ottman, G. K., H. F. Hofmann, A. C. Bhatt, and G. A. Lesieutre. 2002. Adaptive piezoelectric energy harvesting circuit for wireless remote power supply. *IEEE Trans Power Electron* 17:669–76.
48. Roundy, S., B. P. Otis, Y. H. Chee, J. M. Rabaey, and P. Wright. A 1.9-GHz transmit beacon using environmentally scavenged energy. *ISPLED 2003*. Seoul Korea, August 25–27, 2003.
49. Meninger, S., T. O. Mur-Miranda, R. Amirtharajah, A. Chandrakasan, and J. Lang. 1999. Vibration-to-electric energy conversion. *1999 International Symposium on Low Power Electronics and Design*. New York: ACM Press, pp. 48–53.
50. Wang, L., and F. G. Yuan. 2007. Energy harvesting by magnetostrictive material (MsM) for powering wireless sensors in SHM. *SPIE Smart Structures and Materials and NDE and Health Monitoring, 14th International Symposium (SSN07)*, March 18–22, 2007.
51. Mandelbrot, B. B., and J. A. Wheeler. 1983. The fractal geometry of nature. *Am J Phys* 51:286–87.
52. West, G. B., J. H. Brown, and B. J. Enquist. 1997. A general model for the origin of allometric scaling laws in biology. *Science* 276:122–26.
53. Park, B., R. Zaouk, W. Chunlei, and M. J. Madou. 2007. A case for fractal electrodes in electrochemical applications. *J Electrochem Soc* 154:1–5.
54. Park, B. Y., R. Zaouk, C. Wang, and M. J. Madou. 2007. A case for fractal electrodes in electrochemical applications. *J Electrochem Soc* 154:1–5.
55. Wang, C., G. Jia, L. H. Taherabadi, and M. J. Madou. 2005. A novel method for the fabrication of high-aspect ratio C-MEMS structures. *J Microelectromechan Syst* 14: 348–58.
56. Wang, C., L. Taherabadi, G. Jia, M. Madou, Y. Yeh, and B. Dunn. 2004. C-MEMS for the manufacture of 3D microbatteries. *Electrochem Solid-State Lett* 7:A435–38.
57. Wang, C., and M. Madou. 2005. From MEMS to NEMs with carbon. *Biosens Bioelectron* 20:2181–87.
58. Wolf, S. DARPA—DSO. http://www.darpa.mil/dso/thrust/math/quist.htm.
59. Freeman, D. H. 1994. *Brainmakers.* New York: Touchstone.
60. Minsky, M. 1994. Will robots inherit the earth? *Scientific American*, 271 (October):108–13.
61. Kurzweil, R. 1999. *The age of the spiritual machine.* New York: Penguin.
62. Bashir, R. 2001. DNA-mediated artificial nano-biostructures: State of the art, future directions. *Superlattice Microstruct* 29:1–16.
63. Adleman, L. M. 1994. Molecular computation of solutions to combinatorial problems. *Science* 266:1021–24.
64. Minsky, M. 1987. *The society of mind.* New York: Simon and Schuster.
65. Kaku, M. 1997. *Visions: How science will revolutionize the 21st century.* New York: Oxford University Press.
66. Levy, S. 1992. *Artificial life.* New York: Vintage Books.
67. Prigogine, I., and I. Stengers. 1984. *Order out of chaos.* New York: Bantam.
68. Capra, F. 1996. *The web of life.* New York: Anchor Books.
69. Lovelock, J., and L. Margulis. 1974. Biological modulation of the earth's atmosphere. *Icarus* 21:471–89.
70. Stewart, I. 1988. *Life's other secret.* New York: John Wiley & Sons.
71. Morris, R. 1999. *Artificial worlds.* New York: Plenum Trade.
72. Toko, K. 2000. *Biomimetic sensor technology.* Cambridge, United Kingdom: Cambridge University Press.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Part III

Miniaturization Application

(a) The controller that is included as part of the popular *Guitar Hero* video game relies on (b) a three-axis motion-sensing accelerometer from STMicroelectronics. (From http://archive.electronicdesign.com/files/29/19285/fig_01.gif. With permission.)

Introduction to Part III

Chapter 10 MEMS and NEMS Applications

Introduction to Part III

In Chapter 10 on MEMS and NEMS applications we cover the market dynamics for MEMS and NEMS commercial products. We cover both lithography- and nonlithography-based MEMS. After an analysis of the total world market in MEMS, we consider some selected market categories, i.e., MEMS in the automotive market; MEMS in medical and biotechnology markets (bio-MEMS); MEMS in industrial automation, including environmental monitoring markets; and MEMS in information technology peripherals and telecommunications markets.

The market expectations for NEMS today are much greater than they ever were for MEMS. An important objective in the section on nanotechnology markets is to try to help the reader distinguish between hype and science/facts, perhaps helping to avoid yet another lurking technology bubble.

10

MEMS and NEMS Applications

Outline

Introduction

Preamble on the MEMS Market

Definitions

Overall Market for Micromachines

MEMS Market Segments

MEMS in Medical and Biotechnology Markets: Bio-MEMS

μ-TAS and LOC in Sample Preparation for Molecular Diagnostics

General Sample Preparation Protocols

Examples of Integrated Sample Preparation

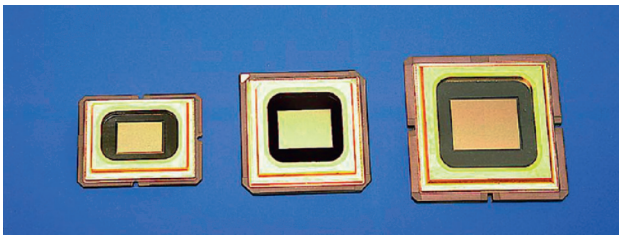
MEMS in Industrial Automation, Including Environmental Monitoring Markets

MEMS in IT Peripherals and Telecommunications Markets

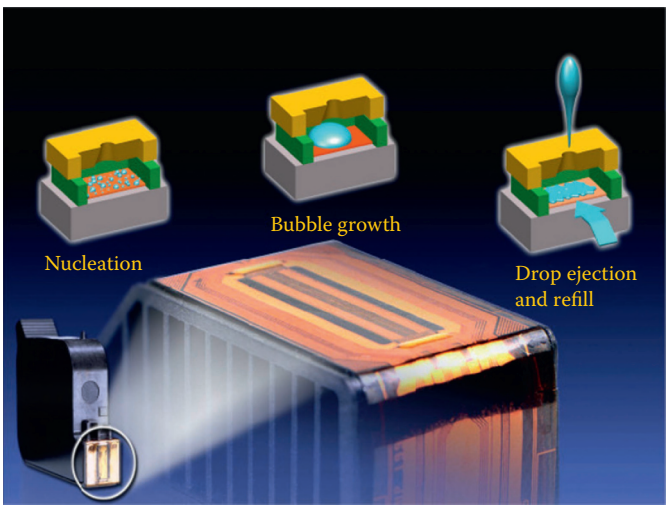
Nanotechnology Markets

Questions

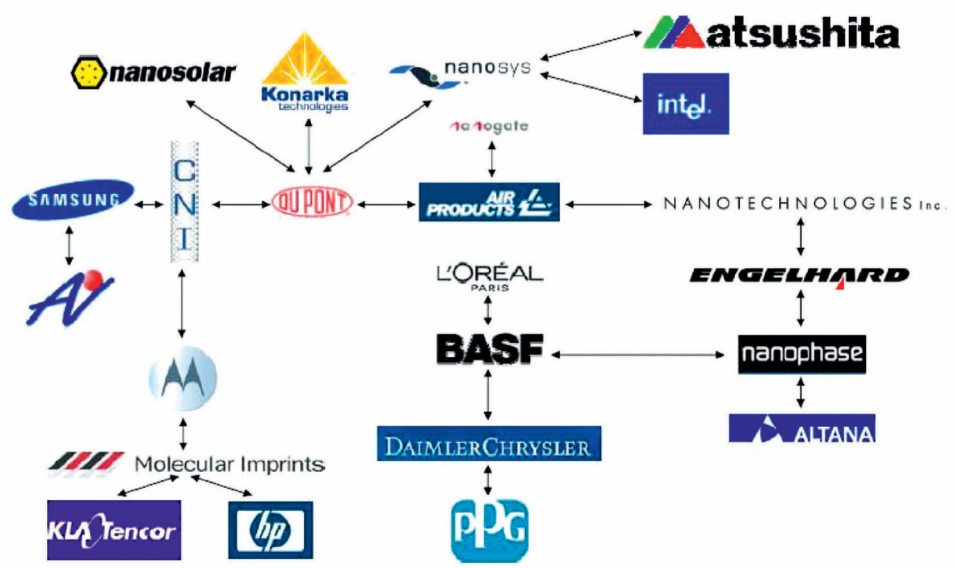
References



Picture of packaged digital micromirror devices (DMD™) from Texas Instruments (TI). These are pixel devices; shown are the VGA (640 × 480), the SVGA (800 × 600), and the XGA (1024 × 768). Digital mirror devices constitute the largest microelectromechanical system (MEMS) product after read-write heads and ink-jet heads.



Another major microelectromechanical system (MEMS) application: an ink-jet printer head. These represent the second largest MEMS market, after read-write heads.



A nanonet of companies, small and large, involved in nanotechnology. Notice that only start-ups use "nano" in their names.

Prediction is extremely difficult. Especially about the future.

Niels Bohr

Cars will cost as little as $200. People will have two-month vacations. They will care little for possessions. The happiest people live in one-factory villages.

Predictions for 1960 by General Motors in a "Futurama"exhibit at the 1939–1940 New York World's Fair

To the electron—may it never be of any use to anybody.

Favorite toast of hardheaded Cavendish scientists in the early 1900s
Michael Riordan and Lillian Hoddeson, Crystal Fire (W. W. Norton, 1998)

People will work every bit as hard to fool themselves as they will to fool others—which makes it very difficult to tell just where the line between foolishness and fraud is located.

Robert Park
Voodoo Science (Oxford University Press, 2000)

Introduction

The field of microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS) is currently undergoing a total makeover. Mechanical and optical MEMS (MOEMS) products have made the transition out of the laboratory and into the marketplace with a plethora of new products, including a strong entry into the consumer electronics field (mobile phones, cameras, laptop computers, games, etc.). Biological MEMS (bio-MEMS), on the other hand, is in the doldrums, with little commercial acceptance, while NEMS, especially BIONEMS, is stealing all the research dollars. In commercial terms, if MEMS is entering adolescence, the word *nano* still seems to derive from a verb that means "to seek and get venture capital funding" instead of from the Greek noun for dwarf. Today the market dynamics for MEMS and NEMS commercial products are very different, and we treat MEMS and NEMS applications separately.

We start with a preamble about today's MEMS market character and the current and future position of the United States in MEMS and NEMS in general. We are interested in both lithography-based and non-lithography-based MEMS. Compared with lithography-based MEMS, thenon-lithography MEMS approach is underreported, and it is especially in

the latter area that progress in the United States is lacking.

To understand where revenues in MEMS are derived from, and to appreciate where further growth can be expected from, we provide definitions of MEMS, bio-MEMS, microsystems technology (MST), micrototal analysis system (μ-TAS), lab-on-a-chip (LOC), micromechatronics, miniature sensors and actuators, microstructures or microcomponents, and first-level packaged MEMS/MST and the smallest commercialized unit. These terms are familiar to micromachinists but are sometimes confused in marketing reports. An evaluation of MEMS applications better describing the MEMS contribution to the final product and the manufacturing method used to produce the MEMS [e.g., non-lithography-based, such as computer numerical control (CNC) machined, or fabricated using a lithography-based technique, such as LIGA] might be of help in interpreting market studies that are often vague about what the term *microfabricated products* comprises. Companies and research organizations interested in entering the field might then at least be able to judge, based on the underlying study, whether they have the manufacturing expertise or know-how to make them possible contenders.

After an analysis of the total world market in MEMS, we consider some selected market categories, i.e., MEMS in the automotive market; MEMS in medical and biotechnology markets (bio-MEMS); MEMS in industrial automation, including environmental monitoring markets; and MEMS in information technology (IT) peripherals and telecommunications markets.

The market expectations for nanotechnology today are much greater than they ever were for MEMS. An important objective in the section on nanotechnology markets is to try to help the reader distinguish between hype and factual science, perhaps helping to avoid yet another lurking technology bubble.

As highlighted in Chapter 1, which compares traditional manufacturing and MEMS and NEMS, we hope that a better understanding of how to match different manufacturing options with a given application will guide the identification of additional killer MEMS and NEMS applications and encourage more companies and research organizations to innovate faster based on their in-house manufacturing tools and know-how.

Preamble on the MEMS Market

Lithography-Based MEMS

Lithography-based MEMS defines all micromanufacturing involving a lithography step. It was not too long ago that lithography-based MEMS was surrounded with a similar hype that surrounds NEMS today. It was said that MEMS, by using highly automated, parallel-manufacturing techniques, would enjoy the same economy of scale as integrated circuits (ICs) and ride similar cost curves. Using "micromachining" technology (much of it borrowed from the IC industry), pumps, sensors, relays, motors—just about any manufactured gizmo in our daily world—would be shrunk down to the micrometer scale, so that it might be hidden, worn, or implanted and in some cases be inexpensive enough for a disposable application. In those early years of MEMS (the early 1980s), when MEMS was almost 100% silicon-based, it was even projected that long before 2008, its market would be larger than that for ICs. This notion was based on the expectation of many more applications for MEMS and combinations of MEMS and ICs than for ICs alone. This market prediction has not been fulfilled yet: the 2006 worldwide IC market was $211.0 billion (see also Volume I, Appendix 1B); in contrast, the worldwide MEMS market for the same year was estimated between $6.3 and $16.5 billion (depending on which criteria one uses to define MEMS). In other words, after more than twenty years, one could claim a MEMS market of at most 8% of the total IC market, taking into account that this was achieved only by changing the criteria about what qualified as MEMS [e.g., including nonsilicon devices, such as thin-film read-write (RW) heads]. The MEMS market in 2007 consisted principally of RW heads, ink-jet printer heads, digital light processing chips, pressure sensors, accelerometers, flow sensors, gyroscopes, microphones, oscillators, and radiofrequency (RF) switches, which, except for digital light processing (DLP) chips, all are becoming commodity products very swiftly and

are now more and more manufactured outside the United States (profit margins have become very thin on most of these products). Early market studies did overstate dramatically what impact Si-based MEMS had on the world economy and how fast that would happen. A realistic number for the market of Si-MEMS products today is less than 3% of the IC market (even the $6.3 billion MEMS sales in 2006, i.e., 3% of the $211 billion IC market for 2006, included nonsilicon devices). The main reason that MEMS sales continue to lag behind IC sales has been the relatively stable price-to-performance (p/p) ratio of sensors and actuators ever since 1960, which is in marked contrast to the p/p ratio of ICs, which has fallen enormously between 1960 and 2007 and is now significantly below that for sensors and actuators. As a consequence of these changes, the cost of a measurement system is, in general, dominated first by the cost of the microactuator and second by the cost of the microsensor.

When including nonsilicon MEMS (which many market studies now do), and especially when also counting non-lithography-based MEMS (e.g., precision-engineered products; see below) (which very few market studies do) MEMS might constitute 10% of all current IC sales.

If some investors still cringe when they hear MEMS mentioned, they might be remembering the 2001 technology bubble around "optical MEMS switches" for telecommunications. Optical channel cross-connecting is a key function in most communication systems. In a *hybrid approach*, one converts optical data streams into electronic data using electronic cross-connection technology. Then, the electronic data streams are converted back into optical signals. In electronic cross-connecting of N-by-N channels ($N \times N$), N may be on the order of thousands. Using arrays of MEMS micromirrors, one can obviate the need for back-and-forth conversion of light to electrical signals. Cross-connecting optical channels directly in the photonic domain is known as *all-optical switching*. These all-optical path switches could, in theory, operate at any wavelength, at any speed, and using any kind of protocol. In 2007, the size of typical commercially available all-optical channel switches may switch between 80 or 128 channels (N = 80 or 128) (see Figure 10.1). But in the late 1990s, all-optical channel cross-connecting switches with N = 1000 were already promised by an army of MOEMS start-up companies (150 worldwide). Unfortunately, the more than $10 billion investment by large telecommunication companies such as Corning, JDF Uniphase, and Nortel in 1999–2000 for those very large arrays of micromirror devices became a textbook example of what former Federal Reserve chairman Alan Greenspan called "irrational exuberance." The all-optical switch mania started in 2000 when Nortel bought up the start-up Xros Inc. for $3.25 billion. By the time Nortel killed off the company in 2002, the market had flooded with competitors, most of them with similar MEMS-based ideas for all-optical switches, and suddenly the 1000-port mania was relegated to the oops file. There were not enough customers for that many start-ups, and

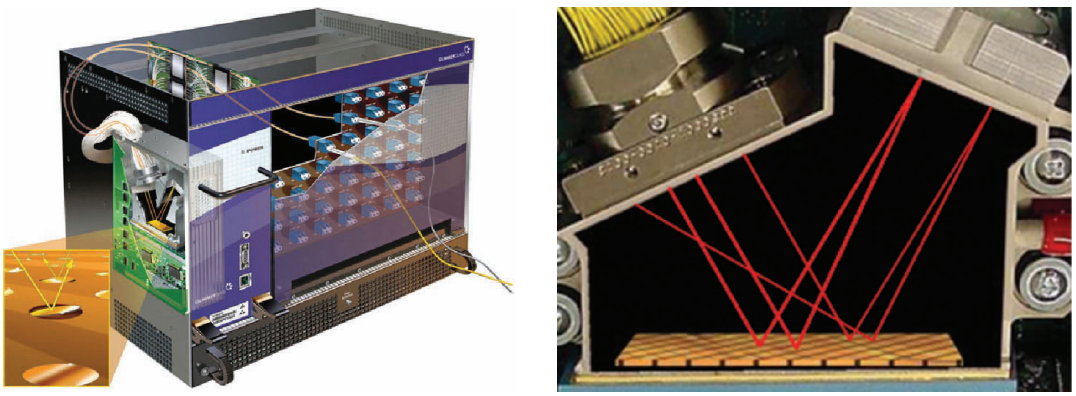


FIGURE 10.1 Glimmerglass Intelligent Optical Switch offers optical to optical switching (http://www.glimmerglass.com/technology.aspx). This switch from Glimmerglass couples light from an input fiber array of up to eighty or more inputs into an output array of the same scale. A microelectromechanical system mirror array directs light from any one input to any of the output fibers. (Images from Glimmerglass.)

the three largest investing companies pulled back, leaving a few individuals very rich, but with neither new jobs nor products to show for their time and with many scientists and engineers out of a job (or working as consultants—often synonymous with jobless in the United States). The rapid growth of the telecommunications industry initially brought great demand for MEMS optical switches. Unfortunately, the expected demand did not meet the actual demand. What was left of the all-optical switching market, in the period between 2002 and 2004, was littered with the broken shells of start-ups that often tried to survive by doing MEMS foundry work. Many of them had never got past the design stage, a few shipped products, and others had nothing to show for but PowerPoint presentations. Surprisingly, even some of the latter managed to cheat investors and large companies out of money as MBAs and scientists alike oversold the technology. The experience embittered many investors, and no wonder.

Today, some surviving MOEMS companies, building all-optical switches, are successfully targeting more modest switch applications; say for apartment buildings or university switchboards and for new emerging applications, such as semiconductor test equipment. Companies such as Glimmerglass claim to have solved the problems previously associated with the MOEMS switches and maintain that their optical switches are more reliable, less complicated, and more cost-effective. The Glimmerglass switch, shown in Figure 10.1, is an 80 × 80 (input × output fibers) switch with promising sales figures. A big challenge for this new generation of optical-switch vendors is overcoming the negative perceptions left over from the boom years.

Instead of switching signals among very large arrays of fiber optic cables for telecommunication switches, Texas Instruments (TI) targeted its digital light processing (DLP) technology to projectors and video displays. Because it focused on the right mass consumer application from the get-go, TI's digital micromirror device (DMD™) turned out to be a big-scale MEMS success (see Example 10.1 below).

A similar case of "irrational exuberance," although on a much smaller scale, played out with the early investments in microfluidics for high-throughput analysis for new drug candidates [called *high-throughput screening* (HTS)] and for high-throughput analysis in diagnostics (Figure 10.2). Early investments for these application came mostly from cash-rich drug companies and huge diagnostic firms, to help them with faster discovery of new drugs and faster diagnostics, respectively, but there was no market beyond some research equipment, and again, too many competitors chased too few customers.

Microfluidic applications in point-of-care (POC) diagnostics—a mass-market application—appear more promising. In diagnostics, one often deals

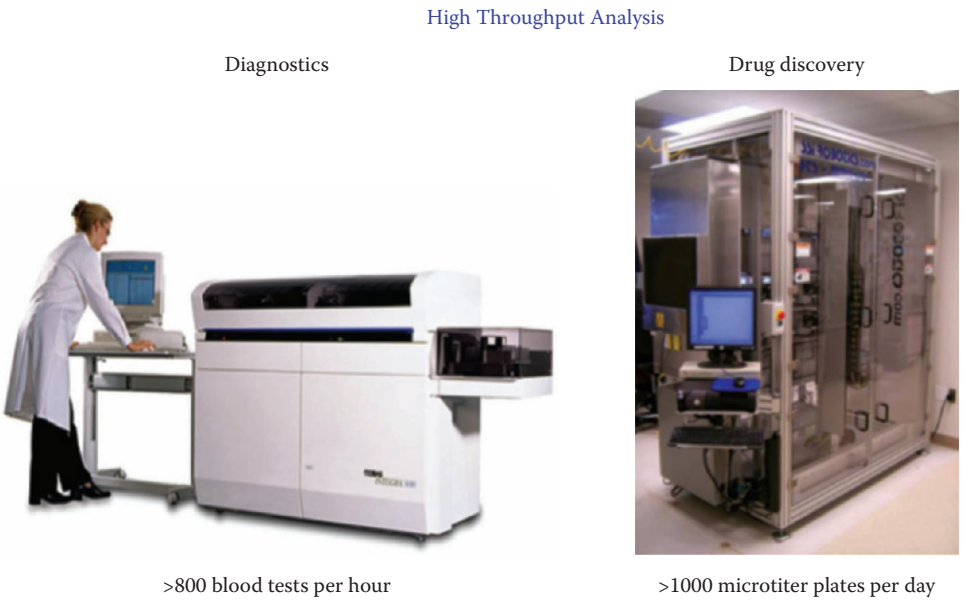


FIGURE 10.2 High-throughput analysis in diagnostics and high-throughput screening in drug discovery.

with disposable products of considerable size compared with a typical IC, and one must compete with traditional products made in huge batches on paper or plastic substrates or even in roll-to-roll (R2R) continuous processes (say, glucose sensor paper strips and aspirin pills). This leaves little room for batch-type Si technology, and as a consequence, most new biotechnology-related MEMS products are nonsilicon-based. Although it failed for HTS applications, nonsilicon microfluidics is poised to succeed in diagnostic applications. This parallels the case of micromirrors that failed for very large communication switches (expensive product/small market) but succeeded for portable projectors (less expensive product/huge market; see next section).

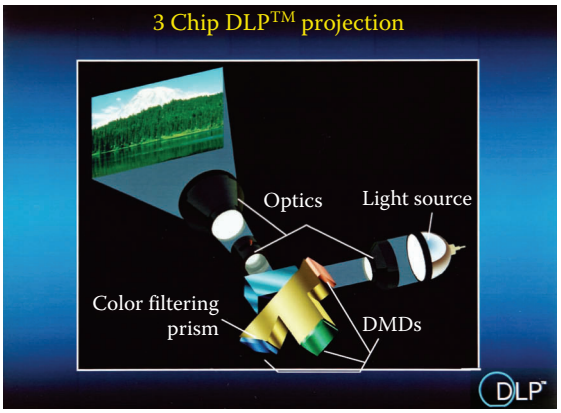


FIGURE 10.3 Texas Instruments DMDs. Digital light processing replaces three conventional LCDs with one to three DMDs, one each for the red, green, and blue components of a video image. Each chip contains approximately 750,000 or more microscopic mirrors or pixels. A single light source (bulb) drives all DMDs, the monochromatic images are converged internally using a prism system, and the resulting full-color image is projected via a lens onto a screen. DLP projectors are lightweight, relatively bright, and simple to set up and operate, and they provide a much better contrast range and higher resolution than LCD projectors. The DLP technology is more expensive at present than LCD.

Example 10.1: Texas Instruments Digital Micromirror Devices

An example of a successful MOEMS application comes from the optical switching arena with the DLP chip from TI. DMDs, invented in 1987 by TI's Larry Hornbeck, consist of an array of hundreds of thousands (from 480,000 to >2,000,000) of electrically actuated micromirrors built as a superstructure over complementary metal-oxide-silicon (CMOS) address circuitry. Two pixels are schematically illustrated in Volume II, Figure 7.98, together with the underlying Si chip and circuitry. Each pixel is made up of a reflective aluminum micromirror supported from a central post. The manufacture of this chip illustrates the integration of a nonsilicon MEMS with a CMOS SRAM 8-inch wafer process line. The surface micromachining process to fabricate DMDs on wafers incorporating SRAM CMOS electronic address and control circuitry is illustrated in Volume II, Figure 7.100. The aluminum alloy DMDs are used in consumer products such as projectors and digital television (Figure 10.3). For a picture of DMD chips, see the illustration in the heading of this chapter.

Millions of those DMDs have now been built into projectors (the first one came out in 1996 from InFocus) and big-screen televisions (the first came out in 2002 from Samsung) (Figure 10.3). It was projected that by 2009, MEMS displays would become the second largest MEMS market, after thin-film read-write (RW) heads (Nexus). A major advantage for TI is that DLP chips are an exception to the high-volume/low-price rule, typically for all ICs and most MEMS devices: the 2004 asking price for a DLP chip was $400; it was estimated at $190 in 2009 (Nexus).[1] There is little question that DMD is the most complex MEMS in the world, and some at TI are bragging that TI has more moving parts on one wafer than Analog Devices Inc. (ADI) builds accelerometers in a year (ADI ships 2 million units a week).

Mechanical and MOEMS devices have been around since the 1980s, and it appears that this market finally has achieved mass-market acceptance in IT peripherals. A primary reason is the demand for read-write heads, ink-jet heads, and microdisplays. Since about 2004–2005, these MEMS devices have been incorporated in consumer products. Read-write heads used to be mostly used in personal computers (PCs) (a market that has slowed down) but can now be found in mp3 players, smart phones, digital video cameras, set-top players, and DVD recorders. Ink-jet heads are migrating from regular printers to printers of digital photos—especially photos from cell phone cameras—and microdisplays found their way

from projectors and high-definition, large-screen televisions to microdisplays for mobile handsets [e.g., the iMoD (interferometric modulator display), developed at Qualcomm] (see also "MEMS in IT Peripherals and Telecommunications Markets").

The automotive field, also, continues to be a strong market driver for MEMS, especially for security applications, with the demand for pressure sensors, accelerometers, and gyros. For example, the electronic stability program (ESP) and the tire pressure monitoring systems (TPMS) are currently promoting worldwide yet more use of MEMS in the car industry.

Recently, the market for accelerometers in consumer electronics (CE) (computers, games, phones, etc.) has also started to explode. Consumer electronics have the potential to dwarf the automotive application, not only for accelerometers but also for pressure sensors and gyros. Consequently MEMS growth is expected by some to follow a hockey stick curve, with the value of products rising to $40 billion in 2015 and $200 billion in 2025![2] (see also an estimate by Dr. Marinis from Draper Laboratory: http://www.sem.org/PDF/s01p01.pdf). The problem now is one of profitability rather than the absence of real product opportunities: how to earn money on, say, $0.32 microphones or $1.35 triaxis accelerometers for mobile handsets? Only by becoming more similar to the IC industry can MEMS devices achieve this goal, i.e., through more integration on larger (8-inch) Si wafers. In a sense, as far as mechanical and MOEMS are concerned, the IC industry has now swallowed the MEMS field.

Despite two important "bubble busts" in the MEMS market, we do expect that besides a booming MEMS business in consumer electronics, many new MEMS applications are yet to be realized and that MEMS will be essential in facilitating the handshake between the macro and nano worlds. Furthermore, this author speculates that it is in nanotechnology, especially when considering bottom-up manufacturing (nanochemistry), that a paradigm shift away from IC-type manufacturing will take shape, and that nanotechnology holds the potential for having a much larger impact on society than IC technology ever did. The IC industry will then become subservient to this larger nano industry, just as MEMS is subservient to the IC industry today.

Non-Lithography-Based MEMS

In this book, we also cover non-lithography-based miniaturization techniques and their applications. Compared with lithography-based MEMS, this technology area is underreported, as results often are evolutionary rather than revolutionary. Less widely advertised than lithography-based MEMS advances, which are oversold in public relation hungry universities and government institutes, non-lithography micromachining, practiced mostly in highly competitive private companies of the likes of Sankyo Seiki, Samsung, Philips, Olympus, etc., actually leads to more practical products faster than lithography-based MEMS. Applications include computer hard discs, photocopier drums, mold inserts for compact disc reader heads, high-definition TV projection lenses, and VCR scanning heads. Some examples of non-lithography MEMS products are shown in Figure 10.4. The United States is in a rather difficult position when it comes to non-lithography-based MEMS [such as electrical discharge machining (EDM) and CNC]. Below, we discuss some reasons why the United States is losing out in this type of miniaturization.

Over the last twenty years, the United States' MEMS efforts focused on exploiting silicon planar lithography as the core technology for microstructure fabrication, whereas in Japan and Europe, a much wider variety of miniaturization technologies were pursued, including several non-lithography-based MEMS methods. A prominent example of this, reviewed in Volume II, Chapter 6, is the ten-year Micromachine Technology Project [also called the *Desk Top Factory* (DTF)], initiated in Japan in 1991 and focused on both a wide range of fabrication technologies and a wide range of materials, including extensions to more traditional machining processes (Figure 10.5). More recently, Korea and the European Community have started similar efforts. There are no such concentrated efforts going on in the United States[3] at the moment.

In Japan, where MEMS was driven by mechanical engineering, there was also more focus on the fabrication and assembly of discrete products, whereas MEMS in the United States, being driven by the IC industry, viewed such work as

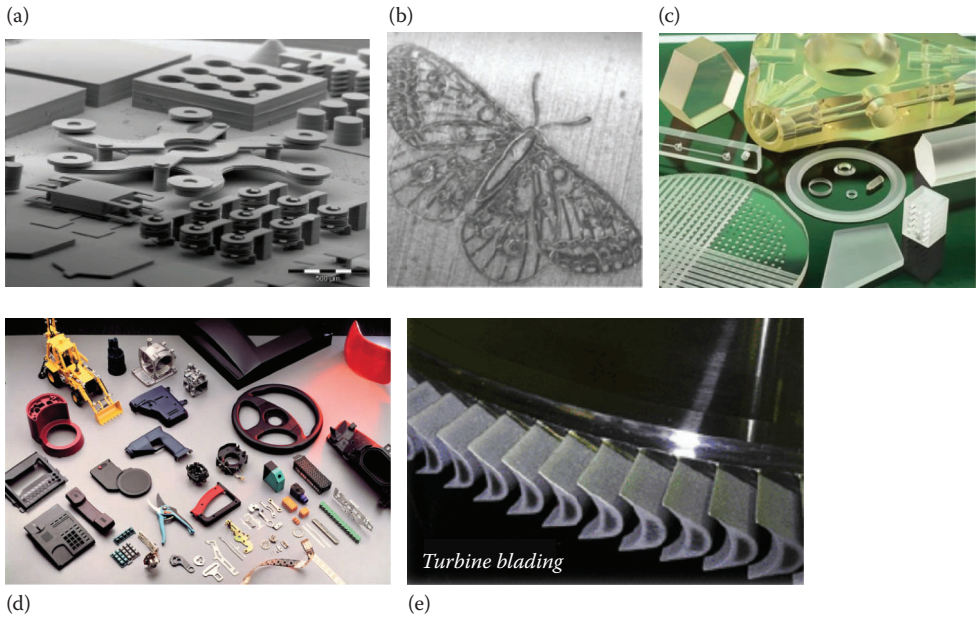


FIGURE 10.4 Examples of nonlithography microelectromechanical system products: products made with precision machining. (a) Electrochemical fabrication (EFAB) is a solid, free-form fabrication technology that can create complex, miniature three-dimensional metal structures that are impossible or impractical to make using other technologies, such as electrical discharge machining (EDM), laser machining, or silicon micromachining. (b) Laser marking and engraving. (c) Precision glass components produced through computer numerical control (CNC) machining, grounding, lapping, and polishing. (d) Electrical discharge-machined products. (e) Turbine blades made with electrochemical machining.

fundamentally violating the paradigm of batch fabrication.[3] Assembly projects are only now gaining vigor in the United States as part of efforts in nanotechnology, embodied, for example, by the Zyvex Nanomanipulator, shown in Figure 10.6 (see also Chapter 4, this volume). The Zyvex Nanomanipulator is able to grab and position objects as small as 10 nm in diameter.

It is difficult to convince U.S. funding agencies to invest in evolutionary-type technology, such as developing better CNC machining tool bits (see Volume II, Chapter 6), but the miniaturization of devices associated with a large number of application fields is today demanding the production of components with manufactured features in the range of a few microns to a few hundred microns, requiring those new tool bits. These fields include optics, electronics, medicine, biotechnology, communications, sensors,

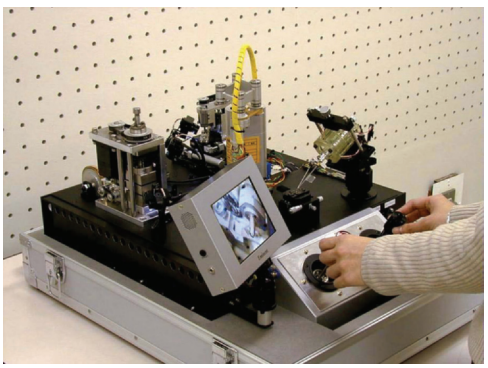


FIGURE 10.5 Example Desk Top Factory (DTF) at AIST (see http://www.dtf.ne.jp/en/achievements.html for other examples).

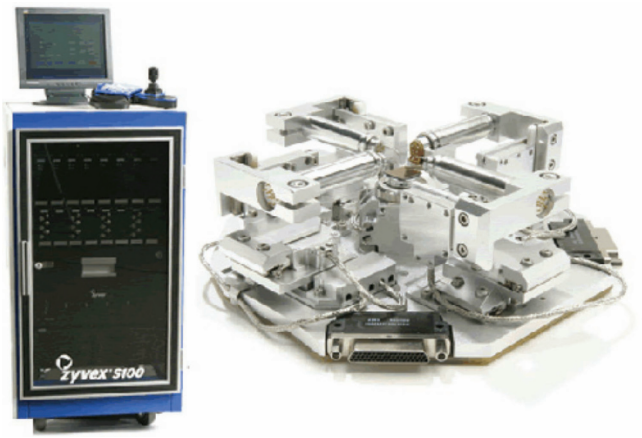


FIGURE 10.6 The S100 Nanomanipulator System from Zyvex is a manipulation and testing tool used with a scanning electron microscope for micro- and nanoscale research, development, and production applications (http://www.zyvex.com).

and avionics, to name just a few. Specific applications are medical implants, diagnostic and remediation devices, microscale batteries and fuel cells, fluidic microchemical reactors requiring microscale pumps, valves and mixing devices, microfluidic systems in general, microholes for fiber optics, micronozzles for high-temperature jets, micromolds and deep x-ray lithography masks, optical lenses, etc. Functional requirements for these applications demand tight tolerances and the use of a wide variety of engineering materials, e.g., steels, titanium, brass, aluminum, platinum, iridium, ceramics, polymers, and composites.

The United States has lost dominance of the tooling industry and the manufacturing skills in many of the precision machining areas needed for the identified applications. The U.S. machine tool industry enjoyed a period of global dominance between the early 1900s and the late 1970s. In the final quarter of the last century, however, almost every major U.S. producer lost domestic market share as a result of import competition. By the early 1990s, several segments of the U.S. machine tool industry appeared poised on the edge of market exit. In a 2007 World Technology Evaluation Center (WTEC) publication, we detailed our observation that countries not manufacturing high-technology goods today are increasingly at a loss, as they do not learn about the newest manufacturing needs from developing the next-generation products.[3]

In the short term, free market forces in a global economy will push mass-product manufacturing of highly accurate miniaturized devices, requiring quality control and multiple assembly steps, toward low-cost, low-wage countries, and the manufacture of small-lot, high-added-value products and items that require intense service assistance will remain in the more developed countries. But the real question is this one: can an advanced country, in the long run, survive without continued input on practical manufacturing, processes, and new materials know-how? The often quoted excuse for exporting manufacturing jobs abroad, usually coming from higher management and MBAs (immune, at least for now, from losing their own jobs), is that the United States and other developed countries will still design, market, and sell those next-generation products. That argument will be rendered wishful thinking, though, when the new materials and processes that go into next-generation products are foreign owned.

MEMS lithography-based and non-lithography-based micromachining will continue to enrich the nations that invest in it most heavily. To stem the hollowing out of the manufacturing base in developed nations, the governments of many countries have made considerable investments in the miniaturization of new products (MEMS and NEMS) and the miniaturization of manufacturing tools (for example, the DTF, mentioned above and detailed in Volume II, Chapter 6). Based on the state-of-the-art and current investment levels (private and government), Europe (especially Germany and Switzerland), and Asia (particularly Japan and Korea) will gain the most from developments in non-lithography-based machining, as they have a long tradition in it and have invested more heavily in this field. As pointed out above, the United States, over the last twenty years, has emphasized lithography-based MEMS with outstanding research results and a dominant market position, but as many MEMS products have become commodity products, Asian countries stand to reap more benefits from them. In more traditional micromanufacturing, the United States is now lagging behind Japan, Germany, Switzerland, and Korea, and as pointed out above, such a lag tends to get worse as time goes on as every new generation of products (from television sets to mobile phones) incorporates some new manufacturing and materials know-how.

If the U.S. government continues to invest in nanotechnology, it will be the United States that will again reap the initial benefits, first in new materials and then in new devices. But to be able to manufacture and gain from those investments in the longer term, it might be wise to start something on the scale of the Manhattan Project in advanced manufacturing techniques; otherwise, the United States will not reap the benefits from its research and development investments itself, as it would no longer be able to implement what it invents.

MEMS Market Studies

As pointed out above, almost all MEMS market projections from before the year 2000 were too optimistic (e.g., $6.3 billion realized vs. a projected more

Table 10.1 From Concept to Full Commercialization

Product	Discovery	Product Development	Cost Reduction	Full Commercialization	Duration (Years)
Pressure sensors	1954–1961	1961–1975 (1961 first prototype at Kulite)	1975–1984	1984	23
Ink-jet printer nozzles	1977 (IBM)	1984–1990	1990–1996	1996	19
Accelerometers	1970 (Kulite piezoresistive)–1979 (Stanford-capacitive)	1985–1990	1990–1995	1995	25 16
Microrelays	1977–1982	1982–1998	1998–2005	2005	28
Photonics/displays	1979 IBM	1986–1998	1998–2005	2001	22
Bio-/chemical sensors	1980–1994	1994–2000	2000–2005	2005	25
Gyros/rate sensors	1982–1990	1990–1996	1996–2004	2004	22
Radiofrequency	1994–1998	1998–2002	2002–2007	2007	13

Source: Grace Associates.

than $200 billion in 2006!). MEMS market studies today have become more judicious, but this author believes that many are still misleading for one or more of the following reasons:

- Many MEMS market studies do not specify whether their market numbers pertain to the MEMS structure alone or whether they include one or more higher system levels. A refreshing change in this regard came from Nexus (http://www.nexus-mems.com) and Yole Développement (http://www.yole.fr), which distinguish in their latest market studies between the first-level packaged MEMS and the smallest commercialized unit. This is illustrated for the case of an ink-jet cartridge in Figure 10.13 (see under Definitions). One cannot buy ink-jet heads alone, and from the point of view of the market (products sold and bought), the smallest commercialized unit is the cartridge or printer head. This decreases the total volume of the market (number of MEMS pieces) but provides a more realistic vision of component-based systems.
- A few market projections still include both MEMS optical switches for telecommunications and MEMS-based microfluidics for high-throughput screening (HTS), despite the fact that both applications have almost disappeared from the commercial scene.
- Most market studies do not specify the machining method used to make the MEMS part. The product could be precision machined, silicon- or non-silicon-based, and involve lithography or not.
- Looking back at the many market reports from the past, it is striking how the development time for a new MEMS product was underestimated over and over again. Table 10.1, based on a report by Grace Associates (http://www.rgrace.com), presents a sobering reality. On average, it still takes about 21 years to go from a new idea to a fully commercialized product. As MEMS products become more integrated and more IC-like, this development time will shrink dramatically.

The lack of more commercially successful applications for micromachines conflicts with the fact that in the evolution of electronic products, miniaturization only follows as a very natural development initiated by a continuing market pull. With radio transceivers, optical pickup systems for compact disc players, hard disks, thin-film read-write (RW) heads, camcorders, etc., features or functions that take up too much space must be compressed; otherwise, they do not survive. Outside the electronics area (such as in automotive sensors and biomedical devices), the market pull is similarly toward miniaturization. So most probably, the real MEMS market has not yet materialized, and product offerings are not meeting the specifications of the market.

There are several reasons why the MEMS market still lags behind the IC market by a factor of 10 or more:

- Micromachining is too often equated with microelectronics, specifically with application-specific integrated circuit (ASIC) manufacture. In reality, MEMS lacks the high degree of generality characteristic of microelectronics. ICs can be grouped in a limited number of classes, within which design and production follow well-defined and common steps. As a consequence, the price/performance (p/p) ratio allows industry to make profits on complex ICs, and it makes ICs suited for mass production. Silicon-MEMS sensors and actuators, on the other hand, are very specific.*
- Sensor applications also call for complicated, novel packaging schemes and MEMS packaging experience is relatively new. MEMS often have to be in direct contact with their surroundings, and each environment imposes its own constraints. This makes packaging for MEMS relatively more important, in terms of cost and (often) size, than ICs, further illustrating our claim in the Packaging section in Chapter 4, this volume that MEMS design should start from the packaging constraints and work its way to the miniature sensor inside rather than the other way around. Packaging and testing costs as much as 70% of total value MEMS product.
- Successive "bubble busts" [illustrated above in terms of microfluidics for high-throughput screening (HTS) and large optical switch arrays for telecommunications] have scared investors and product planners.
- The market is very fragmented.
- Multidisciplinary knowledge is required.
- There is a lack of capital formation opportunities.
- There is a lack of well-defined direction from roadmaps, industry standards, and industry associations.
- Perhaps equally important has been the lack of marketing skills in the early MEMS companies (and, as Regis McKenna says, "marketing is everything"). A survey suggests that 65% of MEMS gurus entered the field because the technology itself seemed compelling ("technology push"). So, obviously, there was a lack of technology/application matching. Micromachinists initially were too far removed from a detailed understanding of the applications. They were not alert enough to market demands and pushed micromachining technologies they controlled rather than solving industrial problems with the most suitable micromachining technology. In reality a MEMS solution based on a more mature precision engineering technology is as good a solution as one based on, say, surface micromachining.

Lessons Learned

One important MEMS market lesson from the last decade is that only MEMS applications geared toward mass consumption applications are succeeding commercially. As an important consequence, silicon-based MEMS devices (e.g., TI's DMDs, ADI's surface-micromachined accelerometers, and all types of RF-MEMS) must be compatible with existing IC lines to reduce costs. The consumer market should be a strong driving force for the MEMS industry to develop more integrated devices and is pushing the MEMS industry to 8-inch wafer size manufacturing. The need to adhere closer to IC manufacturing rules has become even more urgent, as few U.S. MEMS foundries have survived. As we saw above, the emergence of mass consumer MEMS applications has made MEMS very hot again, and the companies and MEMS foundries that did survive the MOEMS bubble can expect to thrive if they can scale up to compete with traditional IC fabs fast enough. In the author's own experience, MEMS foundries are hard to make profitable in the United States, as they need to keep on accepting a wide variety of jobs to keep the expensive clean-room operation going. In the process, they do not stay long enough on the learning curve of one new product to become world leaders for that particular product. Consequently, these laboratories are moving more and more to Asia, where large companies seem to have more patience in keeping the doors open till the MEMS foundry turns profitable.

* To be fair, for micromachines, the economics dictating profitable IC manufacture do not always apply; a microinstrument—say, a micro-gas chromatograph—may cost considerably more than a typical IC. Moreover, sensors and actuators, often using quite esoteric non-IC type materials, are produced in relatively low volumes and are so very application specific that, in principle, they may command a much higher price tag.

Olympus, in Japan, for example, runs the largest MEMS foundry in the world.[3] Similarly Bosch, in Germany, opened a MEMS foundry in 2005 (Bosch Sensortec). More traditional IC foundries have better survival chances in the United States, but it is worth pointing out that China will control 5% of the IC market by 2010 and the largest number of new 8-inch fabs are in Taiwan and Korea.[3] Although Korea and Taiwan are leading in terms of added capacity in 2007, Japan still holds most of the worldwide capacity, with about 24% followed by Taiwan, the Americas, and South Korea.

In case new non-IC materials and new processes need to be developed for a mass consumer MEMS product, as was the case for TI's digital mirror device, long time commitments (5–10 years) and huge sums of money are required (even with initial government funding) to bring the product to market. Few companies can afford such a large investment of time and money.

Definitions

Introduction

Micromachining marketing studies do not always clarify whether the quoted revenues apply to the micromachine alone or include the micromachine with some ancillary electronics and whether they comprise a higher-level subsystem or describe the whole system. Moreover, the precision machining technology employed often remains unspecified. A good market study should start from a clear definition of the products covered and should specify the manufacturing method.

MEMS

The terms micromachining, microfabrication, micromanufacturing, and MEMS describe the fabrication of devices with at least one of their dimensions in the micrometer range. The acronym MEMS originated in the United States[4] (Figure 10.7) and originally applied exclusively to silicon-based mechanical applications. But just as the twentieth century eventually became the electronics century, the twenty-first century promises to become the age of biotechnology and information/communication technology. This shift has manifested itself so rapidly in miniaturization science that we now see bio-MEMS problems eagerly tackled by electrical and mechanical engineers. Twenty years ago, the MEMS field was dominated by mechanical applications; today, most new applications are either information/communication-related or chemical and biological in nature. As a consequence, today MEMS refers to all subminiaturized systems, including silicon-based mechanical devices, chemical and biological sensors and actuators, and miniature nonsilicon structures (e.g., devices made from plastics or ceramics). The application of MEMS to medical and biotechnology problems is called *bio-MEMS* (which includes microfluidic structures, drug delivery devices, immunosensors, and DNA arrays).

More specific terms, such as MOEMS (e.g., micromirror arrays, fiber optic connectors), RF-MEMS (e.g. inductors, capacitors, antennas), commercial-off-

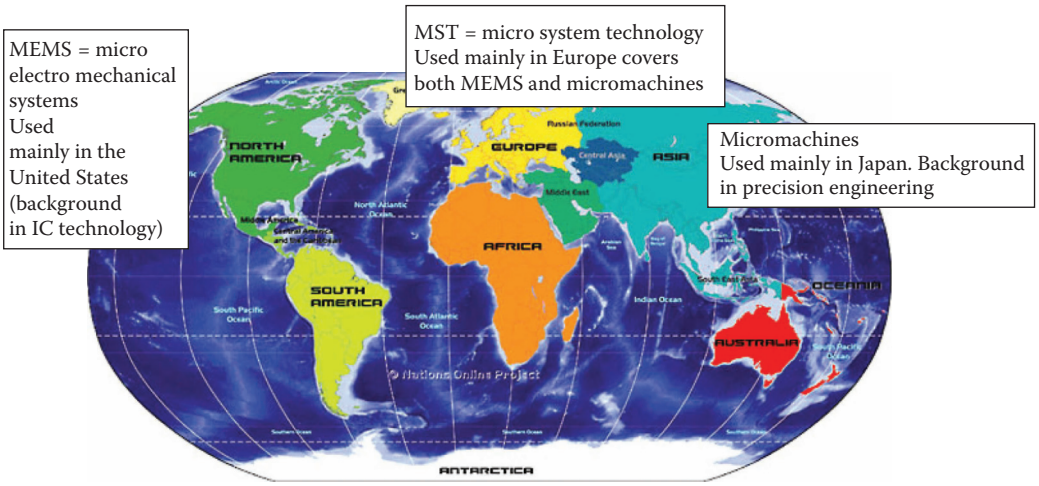


FIGURE 10.7 Micromachining terms around the world.

the-shelf microelectromechanical systems (COTS-MEMS), and high-aspect-ratio MEMS (HARMEMS) have also been introduced, and many more can be expected as many untenured professors want to have at least named something in the MEMS field.

Microsystems Technology

In Europe, the term *microsystems technology* (MST) is more prevalent than MEMS (Figure 10.7). One European definition of MST reads: "A microsystem is an intelligent miniaturized system comprising sensing, processing and/or actuating functions. Normally these would combine two or more of the following: electrical, mechanical, optical, chemical, biological, magnetic or other properties, integrated onto a single multichip hybrid." Alternatively, "A microsystem or microinstrument integrates sensors, actuators, and electronic components on a small footprint, collects and interprets data, makes decisions, and enforces actions upon its environment."[5]

As microsystems are combinations of sensors, actuators, and processing units, they are very application specific. They should be distinguished though from application-specific ICs (ASICs), which can be grouped into a much more limited number of classes within which design and production follow well-defined and common steps.[6]

Micrototal Analysis Systems/Lab-on-a-Chip

Micrototal analysis (μ-TAS) constitutes MST with an analytical function. An alternative term often used is lab-on-a-chip (LOC). In most of the literature today, little distinction is made among MST, μ-TAS, and LOC. Examples of commercial products in this general category are Nanogen's NanoChip™ molecular biology workstation (http://www.nanogen.com) (see Volume II, Example 8.2), TI's Spreeta and handheld reader (http://www.sensata.com/sensors/spreeta-analytical-sensor-highlights.htm) (see Figure 10.47), Caliper's microfluidic LabChip systems (http://www.calipertech.com), RotaPrep's sample preparation stage (Figure 10.51), and Agilent's MicroGC (Figure 10.8). It is important to note that most μ-TAS today still involve sizable instruments. Only some of the components have been miniaturized. Microsystems of this type still carry a price tag similar to that of large instruments ($2,000 to $120,000). The miniaturized components enable some improved

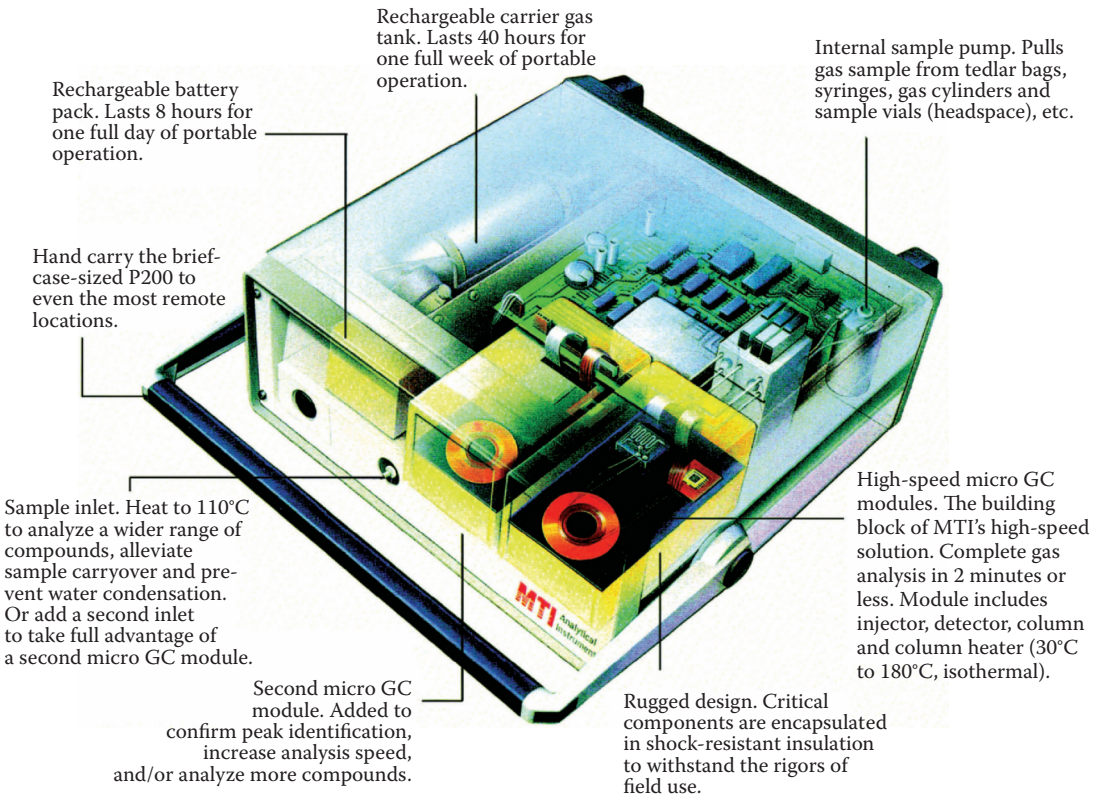


FIGURE 10.8 Agilent's MicroGC (formerly MTI).

functionality, but not yet a lower cost or a smaller, perhaps handheld, instrument. Some of the notable exceptions are the portable surface plasmon resonance sensor (Spreeta) shown in Figure 10.47 (with an instrument of $250 and disposable component of $30) and Johnson & Johnson i-STAT handheld automated blood analyzer, shown in Figure 10.42. The same instrument shown in Figure 10.8, but in a handheld format, would command a price tag somewhere between that of large instruments ($2,000 to $120,000) and sensors ($0.50 to $100), i.e., $200 to $500. To develop more handheld analytical equipment like the Spreeta is one of the μ-TAS challenges that lie ahead. Through the widening of MEMS applications, the word *MEMS* in the United States is more broadly interpreted now, and MST and MEMS have become more or less synonyms.

Micromechatronics

A mechatronics product is defined as a smart device that involves mechanical and electrical principles in design and performs hybrid mechanical and electrical functions. In Japan, the focus has been on micromechatronics or micromachines, defined as follows: "Micromechatronics is the technology that involves the design and manufacture of mechatronics products with dimensions in the range of 1 μm to 1 mm and are capable of performing complex microscopic tasks."[5] (For an example, see Figure 10.9.) In addition to the difference in sizes between mechatronics and micromechanotronics, major differences are in design methodologies, fabrication techniques, and packaging and testing.

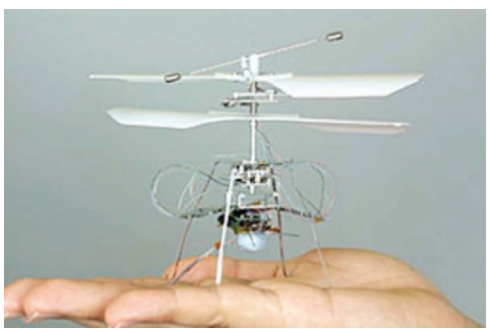


FIGURE 10.9 The world's smallest and lightest micro-flying robot. Epson pioneered the micro-flying robot, based on micromechatronics technology—the world's lightest and most advanced microrobot, which also features Bluetooth wireless control and independent flight.

In micromechatronics, structural components are designed on the base of scaling laws and electromagnetic actuators often are replaced by thermal, piezoelectric, and electrostatic microactuators. Components more often are fabricated by chemical-physical processes than by machine tools. Assembly, packaging, and testing are also different from those in macroscale mechatronics. Whereas MEMS and MST definitions have grown closer together, this Japanese term retains its separate meaning.

Microsensors

A sensor element is a device that converts one form of energy into another (e.g., ZnO, a piezoelectric material, which converts mechanical energy into electricity) and provides the user with a usable energy output in response to a specific measurable input. Measurands may belong to the radiation, thermal, electrical, chemical, mechanical, or magnetic field domains. The sensor element may be built from plastics, semiconductors, metals, ceramics, etc. To qualify as a microsensor, it must have at least one dimension in the micrometer range. Since the early 1980s, the term *sensor element* often invoked the notion of a silicon-built device, but there are many non-silicon-based microsensors.

A sensor includes a sensor element or an array of sensor elements with physical packaging and external electrical or optical connections. Synonyms for "sensor" are *transducer* and *detector*. A sensor system includes the sensor and its assorted signal processing hardware (analog or digital). *Transducer* sometimes refers to a sensor system, especially in the process control industry.

In the case of silicon-based sensors, some additional jargon has developed. A Si sensor element is called a *sensor die*, which refers to a micromachined Si chip. It typically sells for $0.10 to $2 as a commodity product, although the price tag can rise to $50 or more for a high-performance structure sold in smaller quantities. A *silicon sensor* alludes to a first-level, packaged Si die or sensor element with or without basic electronic circuitry. An *integrated silicon sensor* is a monolithic device comprising the sensor and one or more electronic components, which amplify and condition (standardize) the sensor output signal. The standardization of the sensor signal

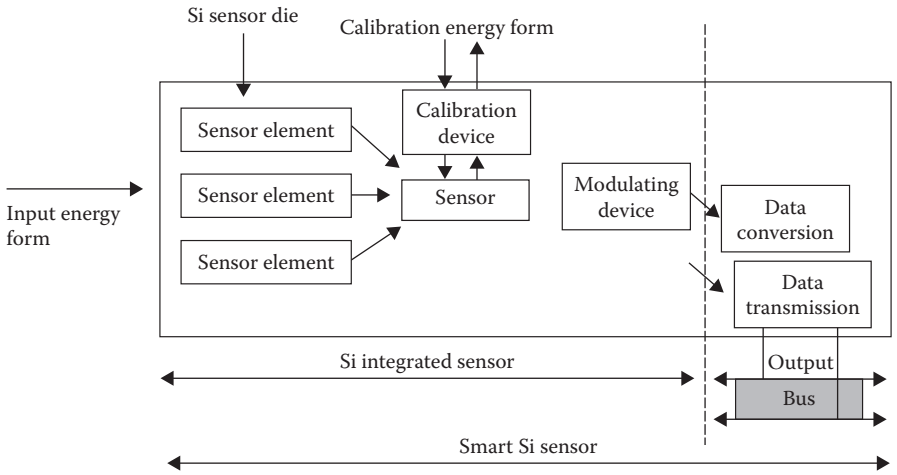


FIGURE 10.10 Definition of a silicon sensor die, a silicon sensor, a silicon integrated sensor, and a smart silicon sensor. (After National Research Council, 1995.)

makes the device bus compatible, enabling efficient communication between the central processor and the sensor. Typical selling prices range from $2.50 for quantities of a million units per year to over $100 for complex sensors in smaller unit quantities. At yet a higher level is a *smart silicon sensor,* which is a packaged integrated sensor containing some part of the signal-processing unit to provide performance enhancement for the user. Signal processing might include autocalibration, interference reduction, compensation for parasitic effects, offset correction, and self-test. The various silicon-based parts defined above are represented schematically in Figure 10.10.

Microactuators

The definition of a sensor element as a device that converts energy from one form into another also applies to an actuator. In the case of an actuator, one is interested in the ensuing action, whereas in the case of a sensor, our interest is in the information garnered. In other words, if the intent is to measure a change, one will refer to a sensor; if the intent is change or action itself, one defines an actuator. The appropriate use of *sensor* or *actuator* is not based on the physics involved but on the intent of the application.[7]

Actuators are components that convert energy into an appropriate action, often dictated by a sensor control unit. They facilitate a function such as opening a valve, positioning a mirror, moving a plug of liquid, etc. Microactuators are actuators with at least some dimension in the micrometer range. Since an actuator "acts," some power is usually needed. Miniaturization of actuators is consequently less obvious than that of sensors, as power does not scale advantageously in the microdomain (power is proportional to the volume, i.e., l^3; see Chapter 8, this volume). To induce micron-scale motion, actuators do not need to be micron-scale themselves. Busch-Vishniac suggests we call the latter *microactuators* in contrast with *microfabricated actuators*, which are micromachined actuators.[8]

The selling price for Si-based actuators in large quantities may range from $5 to $200. A few Si micromachined actuators have reached the commercial market. Examples include Redwood Microsystems' (company now defunct) thermopneumatic valve (Figure 10.11a),[9] TiNi's shape memory alloy microvalve (http://www.sma-mems.com) (Figure 10.11b), and ADI's surface-micromachined accelerometer qualifies (Volume II, Figure 7.92). The latter involves a sensor/actuator combination with a self-test involving electrostatically actuated interdigitated polysilicon fingers (Volume II, Figure 7.93).

Microstructures or Microcomponents

A microstructure or microcomponent refers to a precision machined part that has at least one of its dimensions in the micrometer range but is not a sensor, actuator, or microsystem. Rather, it concerns an item such as a microlens, micromirror, micronozzle, microneedle, etc., that acquires a useful function only when combined with other components. Selling prices of such Si microstructures in large

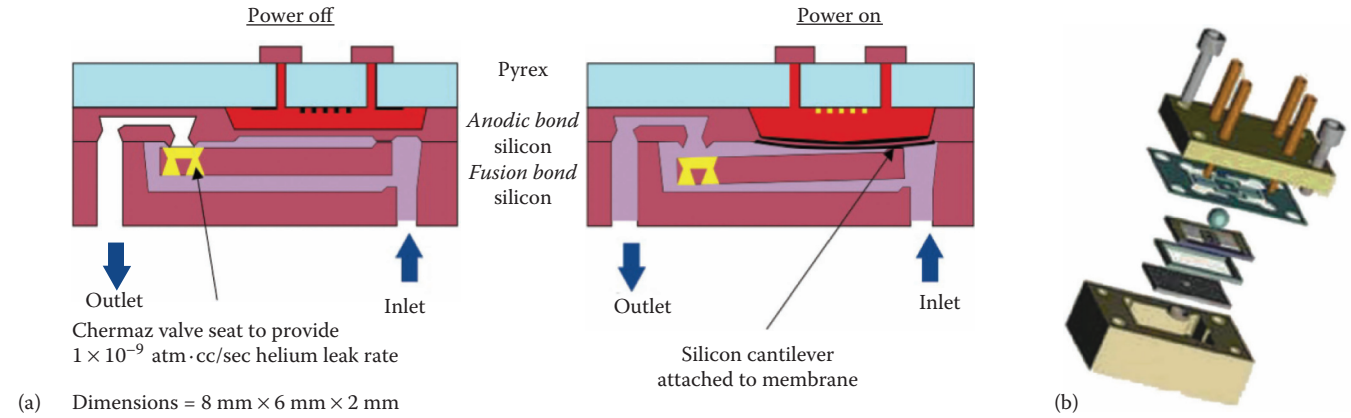


FIGURE 10.11 Examples of commercially available actuators. (a) Thermopneumatic valves by Redwood Microsystems (Fluistor™). Normally closed shut-off microvalve featuring a liquid-filled cavity, which flexes a silicon diaphragm when heated, forcing the valve cover to lift off the valve seat. (After Zdeblick, M. J., R. Anderson, J. Jankowski, B. Kline-Schoder, L. Christel, R. Miles, and W. Weber. 1994. In *Technical Digest: 1994 Solid State Sensor and Actuator Workshop,* 251–55. Hilton Head Island, SC.) (b) TiNi Alloy Company's shape memory alloy valve ($190/valve in small quantities). This is a normally closed microvalve. A current through the NiTi die lifts it from the orifice die. (Courtesy of Dr. D. Johnson, TiNi Alloy Company.)

volumes may range from $0.25 to $100.[10] The latter numbers are hard to confirm, since most microstructures are delivered to clients as part of large development contracts.

An example of a Si microstructure built by the author and his team is the scanning tunneling microscope (STM) cantilever with integrated sharp Si tip shown in Figure 10.12.[11,12] The silicon tip is sharpened by consecutive oxidation sharpening, a process developed by Ravi and Marcus[11] (see also Volume I, Chapter 4, "Orientation and Dopant Dependence of Oxidation Kinetics"). While we delivered the first STM tips to a client company under a best effort-type research and development contract, eventually, these components were sold at a fixed price. Since an STM tip determines the quality of an STM picture, this product is an example of how a small micromachined component might provide a competitive edge to an STM instrument (~$20K to $60K) manufacturer. In this case, it is important that the instrument manufacturer controls the micromachining technology of the crucial microcomponent or, better yet, has it in-house.

First-Level Packaged MEMS and Smallest Commercialized Units

Nexus, a consulting firm on MEMS, makes an important distinction between the first-level packaged MEMS/MST and the smallest commercialized unit.[1] This is illustrated in the case of an ink-jet cartridge

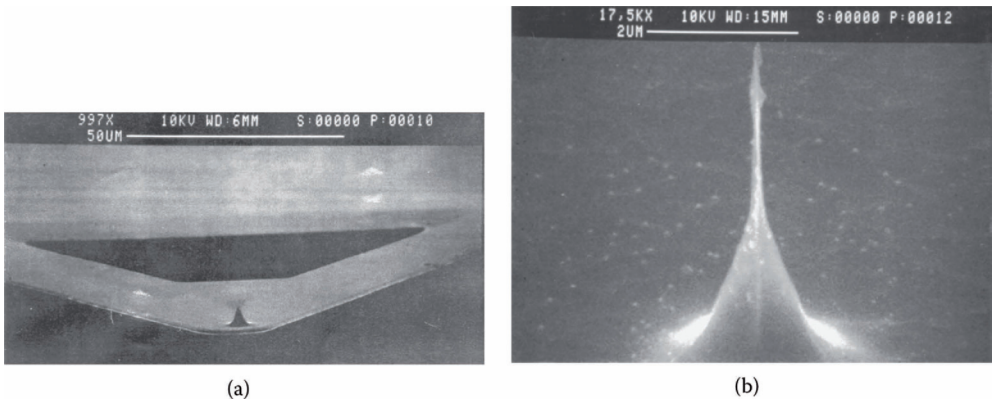


FIGURE 10.12 Example of microcomponent made with silicon micromachining: single-crystal scanning tunneling microscope tip on single-crystal cantilever beam. (a) High-aspect-ratio silicon tip. Photograph taken from a 45° angle. (b) Tip is actually 40% taller than it appears. (From Ravi, T. S., and R. B. Marcus. 1991. Oxidation sharpening of silicon tips. *J Vac Sci Technol B* 9:2733–37; Editorial. 1991. Menlo Park, CA: TSDC [Teknekron]. With permission.)

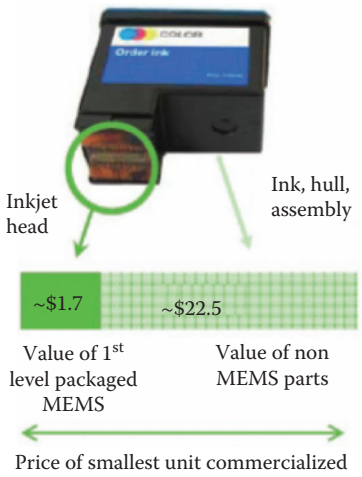


FIGURE 10.13 Example of an ink-jet printer head. First-level packaged MEMS and smallest commercialized unit. (From Nexus. 2006. *Nexus market analysis for MEMS and microsystems III, 2005–2009.*)

in Figure 10.13. One cannot buy ink-jet heads alone, and from the point of view of the market (products sold and bought), the smallest commercialized unit is the cartridge or printer head. The MEMS/MST in the cartridge/printer (~$1.7) is worth much less than the total value of the cartridge/printer (~$22.5).

This simple clarification is very relevant for Si foundries interested in manufacturing the Si die only.

Overall Market for Micromachines

Introduction

Richard Feynman presented two lectures on micromachining: "There's Plenty of Room at the Bottom," at Caltech on December 16, 1959, and "Infinitesimal Machinery," at the Jet Propulsion Labs on February 23, 1983.[13,14] In these now famous lectures, he speculated on the use of micromachining for some esoteric devices such as microfabricated light shutters, probing devices for electronic circuits, and adjustable masks, the first two of which have been realized. He was also convinced of the future of truly three-dimensional circuitry in computers. In general, he doubted the wide applicability of the technology, though. Feynman could not suggest many uses for micromachinery, but he was not worried. He said, "I'm fascinated but I don't know why."

In this section, we investigate applications and overall markets for MEMS and show that the field has matured, with MEMS products now entering consumer electronics so that the market is poised for a major take-off. We conclude with a very positive outlook for MEMS, with the understanding that micromachining be defined as broadly as we have in this book—beyond silicon-based technology and perhaps even beyond batch methods; that is, including continuous manufacturing methods.

Overall MEMS Market Size

The numbers for the overall market size and market projections for miniaturization technology presented here principally are culled from marketing studies and insights by Nexus (http://www.nexus-mems.com), Yole Développement (http://www.yole.fr), Bosch Sensortec (http://www.bosch-sensortec.com), SEMI (http://www.semi.org), Frost & Sullivan (http://www.frost.com), Draper Labs (http://www.draper.com), Wicht Technologie Consulting (WTC) (http://www.wtc-consult.de), Janusz Bryzek (janusz@bryzek.com), and Roger Grace Associates (http://www.rgrace.com).

According to the Nexus Market Analysis for MEMS and Microsystems III 2005–2009,[1] the MEMS/MST market volume was worth $12 billion in 2004 and was expected to reach $25 billion in 2009. The growth was projected to be rapid, with a 16% compound annual growth rate (CAGR) in value in the period 2004–2009 (Figure 10.14). We like to point out that the 1998 Nexus Task Force[5] projected a MEMS/MST market for 2002 of $38 billion already. Obviously, a more detailed analysis about the real MEMS/MST contribution has made for a more careful forecast in this latest report. Considering the traditional Nexus definition used in earlier reports—the market for the smallest unit, incorporating MST component(s), that is commercially available—the market was expected to increase from $36 billion in 2004 to $52 billion in 2009 (Figure 10.14).

In 2005, according to Yole (http://www.yole.fr/), more than half of Si-based MEMS by value, which this consulting firm estimates at $2.6 billion, were produced in North America. Japan was the next largest MEMS producing region, accounting for one-fifth of the total 2005 MEMS revenues ($1.1 billion). About 16% of MEMS sales, over $815 million, originated from European production. Singapore was also a significant manufacturing location for MEMS

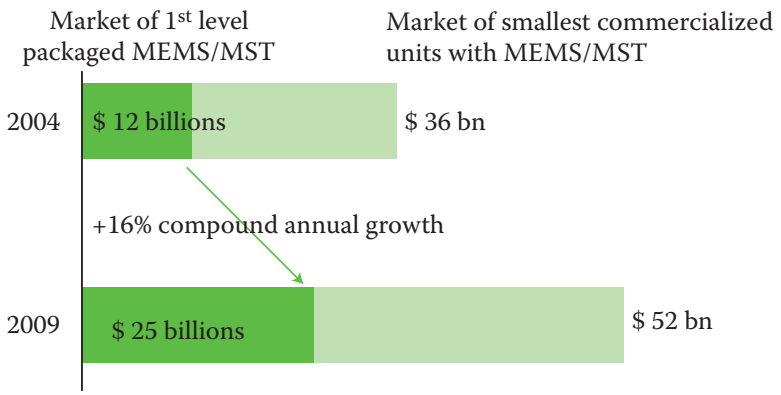


FIGURE 10.14 Total market for microelectromechanical system/microsystems technology in 2004 and projections for 2009. (From Nexus. 2006. *Nexus market analysis for MEMS and microsystems III, 2005–2009.*)

manufacturers (e.g., STMicroelectronics with ink-jet heads), with 11% of MEMS device revenues attributed to Singapore fabs. The MEMS industry is highly fragmented with a large number of device manufacturers worldwide in a wide variety of application fields. Many start-ups and small and medium-sized companies participate in the industry. In 2004, there were 230 manufacturers of MEMS products. Most of these were developing new products, some of which are just starting to significantly affect the market. The top thirty MEMS device manufacturers accounted for about 84% of MEMS device sales in 2005. No single device manufacturer has reached the billion-dollar sales mark as of yet: the average revenue for a top thirty manufacturer is about $147 million (http://www.yole.fr/pagesAn/products/semi.asp).

In Table 10.2 we list 2005 MEMS sales of some important MEMS companies. The top three MEMS companies today are TI, Hewlett-Packard, and Bosch.

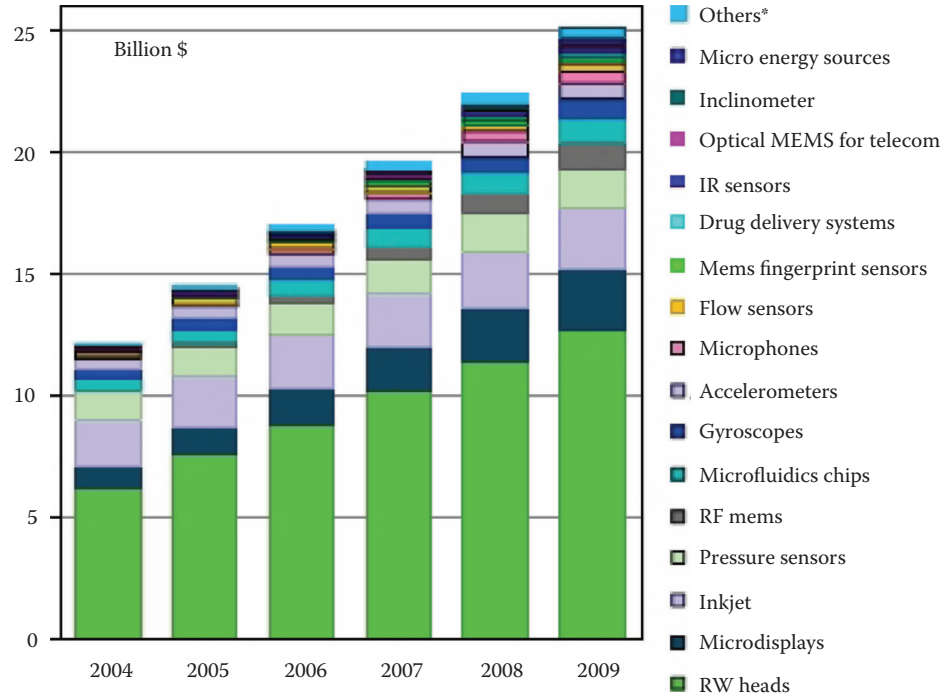
In Figure 10.15 we show the Nexus market breakout for sixteen types of first-level packaged MEMS/MST (a seventeenth category, "other," includes microreactors, chip cooler, inclinometers, memories, fingerprints, liquid lenses, microspectrometer, wafer probes, micromirrors for optical processing,* micropumps, micromotors, and chemical analysis systems).[1] According to this breakout, the MEMS/MST market in 2009 was still 70% made up of only three products: read-write (RW) heads, ink-jet heads, and microdisplays. RW heads still represented around 51% of the market in 2009: while traditional application of RW heads in PCs would grow only moderately, the RW market was experiencing a renaissance in consumer electronics, as hard discs were entering music players (e.g., in many iPod models) and smart phones [Samsung introduced the first cell phone with a hard disc drive (HDD) in 2004], as well as digital video cameras, set-top players, and DVD recorders. It is important to note that MEMS will lose a very important market when HDDs are replaced by solid state disc (SSD) memory (a flash drive), as both RW heads and accelerometers will fall away. There are a lot of arguments to suggest that SSDs will replace HDDs in many applications. The SSD market is growing fast, approximately doubling in revenue every year for the foreseeable future. In the long term (5–10 years) the trusted HDD will eventually retire, to be replaced by the flash drive.

Ink-jet heads continued to be one of the most profitable markets for MEMS. MEMS microdisplays are the new MEMS blockbuster, and their revenues were expected to overtake ink-jet heads in 2009. TI was forging ahead with the digital light processing (DLP) chip for projectors and for front projection

TABLE 10.2 Sales (in 2005) of Microelectromechanical System Products by Some of the Largest Players

Company	Sales ($, in Millions)
Texas Instruments	788
Hewlett-Packard	750
Bosch	325
Lexmark	230
ST Microelectronics	200
Seiko	199
Cannon	184
Freescale	182

* In Figure 10.15, Nexus somehow has digital micromirrors categorized under both *other* and *microdisplays*.



* Others: Microreaction, chip cooler, inclinometers, MEMS memories, MEMS fingerprints, liquid lenses, microspectrometer, wafer probes, micro-mirrors for optical processing, micro-pumps, micromotors, chemical analysis systems

FIGURE 10.15 Nexus market breakout for first-level packaged microelectromechanical system/microsystems technology devices.[1]

and rear projection TVs (home theater). Three other products that, according to Nexus, would make each over $1 billion in sales in 2009 are pressure sensors, RF MEMS, and inertial sensors.[1] Other fast-growing markets were microphones and tire pressure monitors, next to established pressure and motion sensors, which were increasingly being driven by consumer electronics.

The relative contributions to various application fields of these MEMS/MST for the years 2004 and 2009 are illustrated with two pie charts in Figure 10.16. The three dominant application fields were IT

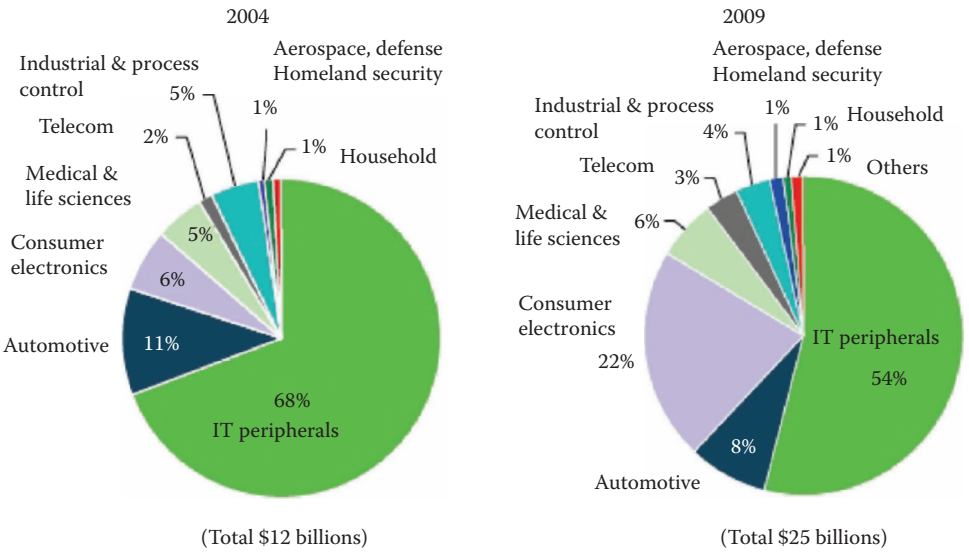


FIGURE 10.16 The relative contributions to various application fields for microelectromechanical system/microsystems technology products for 2004 and projections for 2009. (From Nexus. 2006. *Nexus market analysis for MEMS and microsystems III, 2005–2009.*)

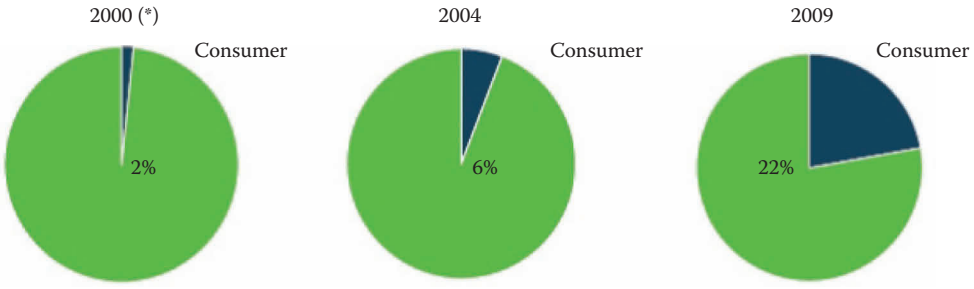


FIGURE 10.17 MEMS in consumer products will have gone from less than 2% of the MEMS/microsystems technology market in 2000 to 22% in 2009. (From Nexus. 2006. *Nexus market analysis for MEMS and microsystems III, 2005–2009*.)

peripherals, automotive, and consumer electronics. IT peripherals were expected to remain by far the first application field of MEMS and MST in 2009, mainly with RW heads and ink-jet heads. However, the share of IT peripherals was expected to decrease from 69% to 54%. The biggest predicted change was in the area of MEMS in consumer electronics (CE) that was predicted to grow from less than 2% of the MEMS/MST market in 2000 to 22% in 2009 (Figure 10.17). Consumer electronics were expected to drive the MEMS market growth for the remainder of this decade. The three drivers for MEMS in consumer electronics were 1) large-screen high-definition television for everybody (rear projection and front projection home theater TVs), 2) more storage in digital equipment (with HDDs that enter digital video cameras, music players, smart phones, set-top players, and DVD recorders), and 3) mobile handsets (with accelerometers today, and tomorrow gyros, microdisplays, micro fuel cells, fingerprint sensors, zoom sensors, gas sensors, weather stations, projection displays, etc.).

Automotive applications were expected to remain a major MEMS field, with several killer application fields, such as airbags and tire pressure monitoring system (TPMS). The numbers of MEMS for this application would still grow at a rapid pace; however, the growth in revenue was expected to be moderate due to a continuous pressure on prices (e.g., gyroscopes increasing 15% per year in units but only 8% in revenue).

Yole analyzes the overall MEMS market structure a bit differently. It introduces a MEMS supply chain, as shown in Figure 10.18, and distinguishes among MEMS-based systems, MEMS alone, front-end equipment, and materials and chemicals.[15] Also, the Yole MEMS numbers refer to Si technology only, so they do not include, for example, RW heads, which are not silicon-based.

According to Yole, the worldwide MEMS systems market reached $40 billion in 2006 and was expected

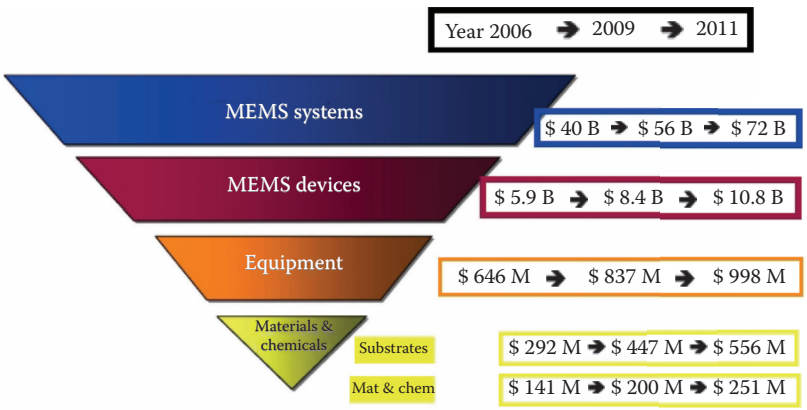


FIGURE 10.18 Yole's microelectromechanical system (MEMS supply chain from 2006 to 2009 to 2011. (From Yole. 2007. Evolution of the MEMS markets and evolution of the MEMS foundry business. *Transducers*.)

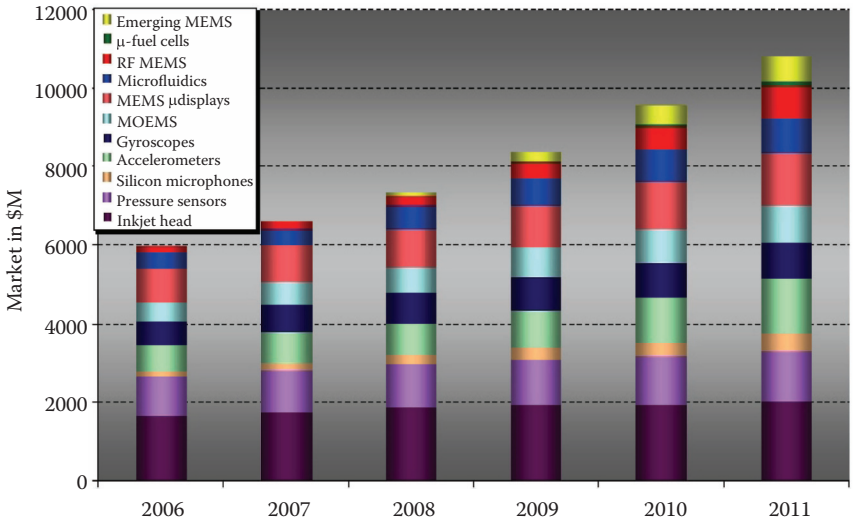


FIGURE 10.19 Global microelectromechanical system market for 2006–2011, according to Yole.

to rise to $56 billion by 2009 and $72 billion by 2011, for a compound annual growth rate (CAGR) of about 15% over the five-year period.[15] Globally, the MEMS market was $5.9 billion in 2006 and was forecast to grow to $8.4 billion in 2009 and to $10.8 billion in 2011, with a 13% CAGR. The MEMS equipment market was $646 million worldwide in 2006 and was expected to expand to $837 million in 2009 and to $998 million in 2011. The five-year CAGR forecast for MEMS equipment was 6%. MEMS materials can be divided into substrates and chemicals and other materials. Together these markets totaled $433 million in 2006 and were forecast to be $807 million in 2011. MEMS materials are expected to grow at a five-year CAGR of 15% through 2011.[15]

The global MEMS market for ten different products and one category called *emerging MEMS*, over the period 2006–2011, according to Yole, is illustrated in Figure 10.19.[15] These numbers are only 1/3 of the Nexus numbers, mostly because Yole does not include RW heads (which are not based on Si technology; Yole's market study is Si-centric) and because under its Emerging MEMS category, Yole is considering some other MEMS products than Nexus lists in its Others category.

Yole puts the ink-jet head market at $1.8 billion in 2006 and expects it to grow to $2 billion in 2011. Although digital imaging is a strong driver for this relatively mature market, Yole expects that the overall trend toward nondisposable ink-jet heads will dampen volume growth for this application.

As in the Nexus report, Yole sees a big change in the MEMS market today in that an important part is now driven by the growth of consumer electronics, which is new in MEMS industry. The period from 2007 to 2011, Yole claims, will certainly see the dominant growth of consumer applications compared with automotive, industrial, and/or medical business. This move toward the consumer applications also is pushing the MEMS industry to 8-inch wafer size manufacturing, and this means more involvement of IC foundries in MEMS business. This switch to 8-inch wafer has a lot to do with the increase of ink-jet head die size and the transition to 8-inch by TI for digital mirror production.

Fast-Changing MEMS Market Characteristics

The very large projected markets for MEMS in market studies published before 2000 are realistic only if nonsilicon approaches are included (e.g., thin-film RW heads) or a first-level system in which the MEMS technology is embedded is counted as well. After the 2001 optical and fluidic MEMS bubbles, major changes in the markets for mechanical and MOEMS devices (RW heads, ink-jet heads, pressure sensors, accelerometers, gyros, digital micromirror arrays, microphones, etc.), are causing upward revisions of MEMS market projections again. These types of MEMS are now experiencing a growth spurt because of their incorporation in consumer products (cars, mobile phones, games, PCs, cameras, etc.). Mechanical and MOEMS largely are made

with IC-based techniques, and the cost of ICs can be lowered only when the quantity of parts numbers in the millions per year. Consequently for silicon-based sensors to succeed on the same scale as ICs, the current manufacturing environment dictates that one must concentrate on mass consumption products such as cars, PCs, air conditioners, TVs, cell phones, vacuum cleaners, toys, printers, and dishwashers. This transition to consumer products has finally happened for optical and mechanical MEMS, and for these products the market dynamics have now become similar to those of the IC market. Consequently MEMS growth is expected to follow a hockey stick growth curve, with the total value of MEMS rising, according to one market report, to $40 billion in 2015 and $200 billion in 2025 (Dr. Marinis from Draper Laboratory: http://www.sem.org/PDF/s01p01.pdf). The problem now is more one of profitability rather than the absence of real product opportunities, i.e., sensors will become smaller and smaller, more and more integrated and manufactured on larger (8-inch) Si wafers. In a sense, one might say that the IC industry has swallowed the mechanical and MOEMS field. The more established MEMS such as accelerometers (ADI), ink-jet heads (TI), and RW heads (IBM, Hitachi) are exploited now mostly by large companies. The involvement of MEMS and IC foundries is on the rise as well, and the tremendous cost pressure is driving more and more MEMS manufacturing overseas.

However, one must still make a distinction between ICs and MEMS. There are about 50,000 types of sensors to measure 100 different physical and chemical parameters,[16] so often only 1,000 to 10,000 are needed a year. We can conclude that fragmentation remains the outstanding characteristic in the sensor market. This often excludes integrated semiconductor solutions, even though they might be technically superior. Cost/performance improvement does not always justify the application of IC technology, and hybrid approaches remain an important option especially in microinstrumentation and for chemical sensors. The dynamics of the market for MEMS-based microinstrumentation is, furthermore, different from that of ICs in that MEMS instrumentation can command considerably higher price tags (say, hundreds of dollars rather than $10 or less). In the latter case, smaller numbers of parts may be offset by a higher profit margin. The situation has been compared to that of application-specific integrated circuit (ASIC) manufacture, leading to the acronym ASIS (application-specific integrated system). Hybrid technology will be employed in this case not solely out of cost considerations but because one monolithic manufacturing technique is unlikely to deliver the bandwidth in feature sizes and material types necessary to cover the various needs common in small instrumentation.

The situation for disposable biosensors (e.g., glucose sensors) applied to diagnostics is very different again. In this nascent field one competes with industrial giants and with very efficient established manufacturing processes. In the bio-MEMS field, to succeed similarly to the mechanical and MOEMS

Table 10.3 New MEMS Companies and Products

Company	Product	Where
Silmach	Silicon-based micromotors for the watch industry	France
Perpetuum	Energy harvesting systems using MEMS as building blocks	United Kingdom
Rhevision Varioptic Siimpel (United States)	Liquid lenses for consumer applications	United States France United States
Lilliputian Systems	MEMS-based micro-fuel cells	United States
Qualcomm Liquavista	Innovative microdisplays	United States Netherlands
SiTime Discera Silicon Clocks VTI Technologies	MEMS-based oscillator	United States United States United States Finland
Microvision Philips	New optical microelectromechanical system applications	United States Netherlands

area, one will need to develop mass manufacturing techniques based on very large batches or roll-to-roll manufacturing on paper on plastic substrates.

Most of the risk taking in innovative MEMS concepts is still by U.S. entrepreneurs, although activity in Europe and Asia has picked up, as is clear from the list of new MEMS companies in Table 10.3. There remains about a three- to five-year delay in time between start-up activity in the United States and Europe on a specific subject (e.g., U.S.-based Affymetrix, created in 1992, versus the European OMM in 1996). So far, even successful MEMS start-up companies remain two orders of magnitude smaller than companies in the mainstream IC market.

MEMS Market Segments

Introduction

We consider four MEMS market categories: MEMS in the automotive market; MEMS in medical and biotechnology markets (bio-MEMS); MEMS in industrial automation, including environmental monitoring markets; and MEMS in IT peripherals and telecommunications markets.

MEMS in the Automotive Market

Introduction

Nexus did forecast the number of passenger vehicles, sport-utility vehicles, and light trucks produced globally to reach an estimated 68 million per year by 2009 (add another 15 million for other vehicles, such as buses, trucks, the Popemobile (smile, please), etc.).[1] In a complex electronic/electromechanical system such as the modern car, the need for effective, accurate, reliable, and low-cost electronics and sensors is pressing. While in 1980 electronics accounted for merely 2% of the total cost of a car, today it has reached almost 20%, and it was estimated that by 2010, electronics and software would account for up to 40% of the value of a car (see Figure 10.20). MEMS are making for a larger and larger contribution to this electronic content.

New vehicles feature increasingly sophisticated engine management and safety systems, while a larger number of sensor systems are making their way into lower-priced cars, fueling the overall growth

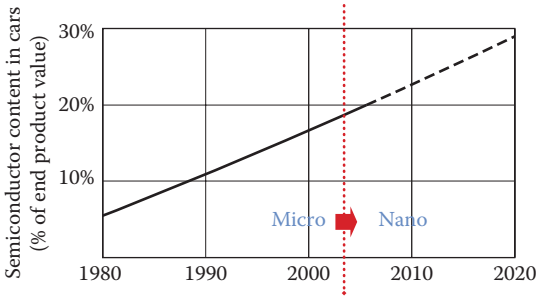


FIGURE 10.20 Semiconductors underpin all business sectors. It was estimated that by 2010, electronics and software would account for up to 40% of the value of a car.

of MEMS in the automotive sensor market. Today, typical mid-sized vehicles contain more than 50 sensors and luxury-class cars more than 100. About one-third of these sensors are MEMS-based, including frontal impact airbags, tire pressure monitoring systems (TPMS), and engine monitors to improve fuel efficiency and lower toxic emissions. This trend will continue as governmental regulations and the needs of the consumer drive auto manufacturers to provide improved safety, quality, and comfort in vehicles. The total market for silicon-based MEMS in safety systems, vehicle navigation, engine management, and cabin comfort, according to a Nexus market report, was $1.29 billion in 2004, and was expected to grow at 9% CAGR to reach nearly $2 billion in 2009, as illustrated in Figure 10.21.[1] Between 2000 and 2004, the largest growth in MEMS applications occurred in the automotive market, with applications expanding to the point that a typical automobile contains twenty-five silicon MEMS devices. Pressure, acceleration, and gyros dominate the present Si MEMS automotive market. The technologies involved are mostly Si bulk micromachining and polysilicon surface micromachining.

Even though the quantity of MEMS devices sold into the automotive market is expected to continue climbing steadily, the automotive percentage of the total MEMS market is actually expected to decline, as the consumer electronics category expands.

When considering all automotive sensors (not only MEMS), the worldwide market was more than $10 billion in 2005, and it was expected to continue to increase at an average annual growth rate of 6% to $14.2 billion in 2010. From automotive sensors, speed and position (38%) and oxygen (20%) sensors would

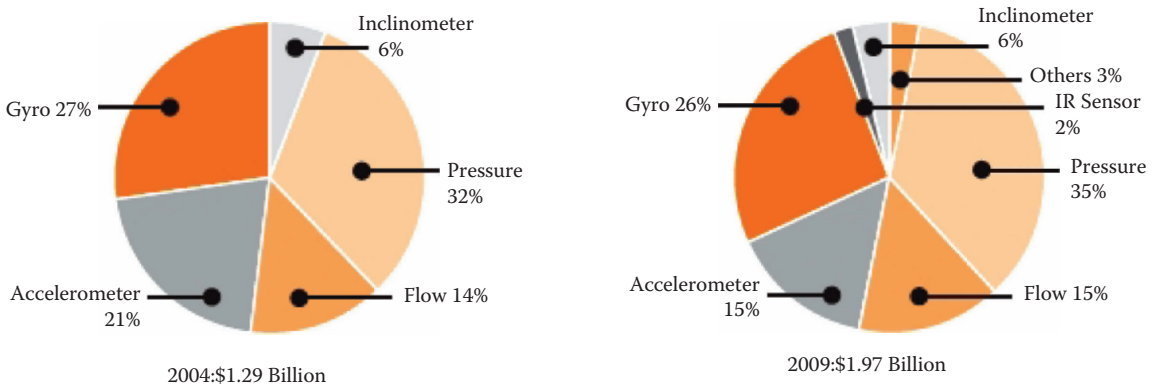


FIGURE 10.21 Various silicon microelectromechanical system sensors in passenger vehicles. (From Nexus. 2006. *Nexus market analysis for MEMS and microsystems III, 2005–2009*.)

remain the largest market segment, rising from $6.2 billion in 2005 to $8.6 billion in 2010 (58%) [*Automotive Sensors*, November 2005 (http://www.the-infoshop.com/study/bc34281-automotive-sensors.html), and Automotive Sensors Market Emerging & Current Sensors Update Now and Future, September 2005 (http://www.researchandmarkets.com/reportinfo.asp?report_id=307920)].

A most important automotive chemical sensor is the so-called *lambda probe*, a solid state oxygen sensor (all new European cars are fitted with a minimum of two lambda sensors). The lambda probe involves traditional ceramics manufacturing, but planar, micromachined oxygen sensors are gaining ground (see Volume II, Figure 9.24).

Many other opportunities for micromachining in automotive sensing exist; a list of sensor needs (met and unmet) in the automotive and transportation industry in general is presented in Table 10.4. There is a continuing migration in the automotive industry from larger electromechanical automotive components to MEMS. In Table 10.5 we summarize this trend. The newest automotive applications for MEMS include better exhaust monitoring devices, wireless sensors for engine management systems, and energy scavengers that harness the energy generated from rotational or vibratory movements within the vehicle as a future power source for TPMS.

Figure 10.22 shows where some of these sensors are positioned in a car. Component cost is a significant factor in the selection criteria of automotive systems designers. Getting your MEMS sensor into a car might take anywhere from seven to ten years. The cost of a MEMS device itself constitutes one-third or less of the total delivered component cost (often $10 or less!). Over 50% of the cost of a typical sensor is in calibration and testing, highlighting the need for innovation. Therefore, a significant challenge to achieve highly efficient design for manufacturability and testing is imposed on all suppliers who wish to be successful in the automotive sector. As a consequence, the use of IC-compatible process lines for MEMS production is growing. There clearly are important benefits in using CMOS or BiCMOS-compatible manufacturing lines to target high-volume MEMS markets, such as the automotive market, and toward adoption of more integrated solutions.

Pressure

Introduction and Market Perspectives

Pressure measurements are absolute, relative, or differential. For most pressure sensors, the common design element is made from a thin, flexible diaphragm that deforms elastically in response to an applied force. Diaphragm thicknesses in these sensors typically range from 20 μm to several millimeters and employ steel, ceramic, or silicon. The ceramic capacitive pressure sensor has a diaphragm cut from relatively thick "cast" sheets of ceramic (e.g., from Kavlico). According to an analysis by Frost & Sullivan (http://www.sensors.frost.com), pressure sensors represent one of the most broad, mature, but very fragmented sensor markets (similar to temperature sensors). In the study "Pressure Sensors and Transmitters Market," the company reveals that the market for all pressure sensors earned revenue of $4 billion in 2004 and projects that this market

Table 10.4 Examples of Sensors and Other Micromachined Components in Use or Needed in Transportation Context

Micromachine	Application	Status
Accelerometers	Airbag release, vehicle dynamics control, antilock braking systems, antitheft systems, active suspension control, headlight leveling	Micromachined products are on the market replacing the old "ball and tube" approach
Air flow sensor	Air/fuel ratio control	On the market but expensive
Alcohol sensor	Detection of alcohol level in blood	No reliable products available
Angle and position sensor	Throttle valve angle, pedal angle or position, and steering angle	All are available
Angular rate sensor	Vehicle dynamics control, navigation, and rollover systems	Available
Azide sensor	Airbag propellant	Does not exist
Battery sensor charge/ density	Charge/density of electrolyte	Research
Contraband detector	Customs	No good solutions
Driver identification	Theft prevention	Some solutions available
Head-up display	Information on window	Not affordable yet
Humidity sensor	Cabin climate	Unsatisfying products available
Level sensor	Oil and gas level	Available
Light sensor	Turn on the lights	Available
Micronozzle system	Fuel injection	Available
NOX sensor	Pollution control	Not available
Oil quality monitor	Engine protection	Not available
Oxygen sensor (two are required in all new cars in Europe)	Air/gas ratio	Available, but new technology is needed for more demanding applications
Pressure sensors	Air intake control (manifold absolute pressure), turbocharger pressure, oil pressure, atmospheric pressure, fuel-tank pressure, brake pressure, climate control, fuel pressure, tire pressure (by September 2007, 100% of the vehicles sold on the United States markets were required to have a tire pressure monitoring system)	Micromachined sensors exist and are on the market and replacing LVDT technology
Parking sensors	Collision avoidance	Acoustic technology available
Rain sensors	Automatic	Technology available
Road/bridge condition sensors	Preventive road maintenance	Not available
Road sensors	Roughness, ice, rain	Not available at cost
Rotational speed and phase sensors	Wheel speed for antilock braking system, engine speed, camshaft, and crankshaft phase for motor control, gear shaft speed for transmission control	Available
Smart windows	Defog, clear view	Not available at low cost
Solid state cameras	Looking behind truck or car	Available
Spill detector	Clearing of truck accidents	No good solutions available
Temperature sensors	Inside and outside vehicle	Available
Torque sensors	Work	Under development

will reach $5.6 billion in 2011. The major industry sectors for pressure sensors are automotive, medical, industrial and process control, aerospace, and military. Household items and consumer products represent upcoming market opportunities.

According to Nexus, narrowing the survey down to silicon-MEMS pressure sensors, sales for first level packaged devices in 2004 were $1.084 billion, and they were expected to grow at a 7% CAGR, resulting in a $1.696 billion market in 2009.[1] Narrowing the scope of the survey even further, the automotive Si pressure sensor market was $408 million in 2004, and revenues for 2009 were expected to amount to $677 million (a 10% growth rate). Silicon

Table 10.5 Migration Status of Automotive Sensors from Large Electromechanical to MEMS

Application	Previous Approach	MEMS Approach	Status
Coolant pressure	Ceramic capacitive	Bonded silicon strain gage	Available (e.g., Keller, SSI Technologies, Fasco)
Exhaust gas recirculation pressure	Ceramic capacitive	Bulk-micromachined silicon	Available (e.g., Ford, Chrysler)
MAP	LVDT	Bulk-micromachined silicon	Mature (e.g., Delco, Nippondenso, Bosch) and more and more replaced by MAF
Airbag accelerometer	Breed "ball and tube" TRW Teknar "rollamite"	Surface and bulk-micromachined silicon	Mature (Analog Devices, Motorola, SensoNor, Delco, and Nippondenso)
Rate sensor	Piezoelectric	Surface and bulk-micromachined silicon	Available (Daimler Benz, Delco, Bosch)
MAF sensor	Hot wire anemometer	Surface-micromachined silicon	Available (Bosch)
Tire pressure	Not applicable	Bulk-micromachined silicon	Available (Sensonor, Motorola, TRW NovaSensor)
Oxygen sensor	Ceramic casting of YSZ	Green tape laminated planar zirconia sensors	Available but expensive (e.g., NGK's UEGO sensor)

Source: Expanded and updated from http://www.rgrace.com/Papers/auto2.htm.

MEMS costs for pressure sensors have come down from $1000 in the 1960s to $10 in the mid-1980s and, in some cases, less than $0.50 today (varies with application-mostly determined by the required packaging and testing). Pressure sensor applications in the automotive industry are listed in Table 10.4 as manifold absolute pressure (MAP), turbocharger pressure, oil pressure, barometric air pressure (BAP) sensor, fuel-tank vapor pressure, brake pressure, climate control, and tire pressure. According to the pie

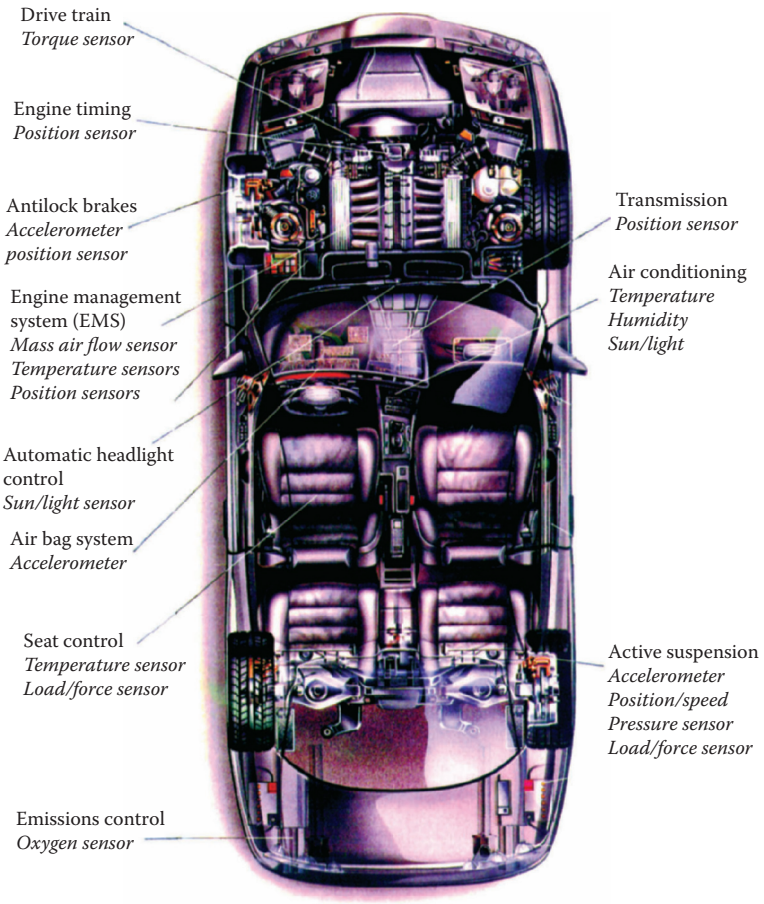


FIGURE 10.22 Automotive sensors.

Table 10.6 Comparison of Pressure Sensor Technology

Parameter	Piezoelectric	Piezoresistive	Capacitive
Self-generating	Yes	No	No
Impedance	High	Low	High
Signal level	High	Low	Moderate
High temperature	Yes	No	No
Accuracy (full scale)	±1%	±1%	±0.2%
Static calibration	No	Yes	Yes
Rugged, high sensitivity	Yes	No	Yes
Damped designs	No	Yes	Yes
Cost	High	Low	High
Electronics required	Yes	No	Yes

(For resonant techniques, see Table 10.7.)

chart in Figure 10.21, pressure sensors will make for 35% of the automotive silicon sensors in 2009 (http://www.wtc-consult.com/cms/cms/upload/PDF/Nexus_Sample_PressureSensors.pdf).

Two main varieties of Si pressure sensors used in cars are piezoresistive and capacitive. In a capacitive design, a thin diaphragm forms one plate of a two-plate device that detects capacitance changes when the diaphragm deflects. In a piezoresistive approach, deflection is measured by the change in the resistance of the silicon strain elements on the diaphragm, or directly in the diaphragm with resistors implanted into the silicon diaphragm. These pressure sensor approaches are compared with a more traditional piezoelectric sensor approach in Table 10.6. A fourth approach is based on resonant structures, as discussed below (see "Resonant Beam Sensors") and compared with piezoresistive and capacitive in Table 10.7. Also, two major manufacturing options are employed for capacitive and piezoresistive pressure sensors: bulk micromachining in single-crystal Si (SCS) and surface micromachining in polysilicon. Recently, Bosch has developed a third pressure sensor manufacturing technology: advanced porous silicon membrane (APSM). This technology makes use of porous silicon to form a cavity underneath the surface of the monocrystalline silicon, which helps to make the wafer manufacturing process fully CMOS compatible and turns the traditional bulk-micromachining process into a surface technology (see Volume II, Figure 4.78). Silicon pressure sensors cover a range spanning vacuum, low pressure (0.02 to 0.1 bar), medium pressure (0.25 to 10 bar), and high pressure (60 to more than 1000 bar).

A capacitive pressure sensor is more sensitive, is less temperature dependent, and holds up better in harsh environments than a piezoresistive one. If one relies on bulk micromachining for the manufacture of a capacitive pressure sensor, it must be carved out of a larger piece of Si to attain the same sensitivity as a piezoresistive sensor, making it more expensive. The electronics for a capacitive pressure sensor are more complicated and expensive. Most of all, in the case of a capacitive sensor, the small signals necessitate integration of sensor and electronics on the same chip. Thus, piezoresistive devices provide higher output signals and are smaller than capacitive sensors, but they require temperature compensation that may be integrated within the same silicon substrate.

The major suppliers of pressure sensors include Honeywell (United States), Bosch (Germany), Infineon (Germany), Freescale (United States), Denso (Japan), Delphi-Delco (United States), All Sensors (United States), Melexis (United States), Integrated

Table 10.7 Performance Features of Resonant, Piezoresistive, and Capacitive Sensing

Feature	Resonant	Piezoresistive	Capacitive
Output form	Frequency	Voltage	Voltage
Resolution	1 part in 10^8	1 part in 10^5	1 part in 10^4–10^5
Accuracy	100–1,000 ppm	500–10,000 ppm	100–100,000 ppm
Power consumption	0.1–10 mW	≈10 mW	<0.1 mW
Temperature cross-sensitivity	-30×10^{-6}/°C	$-1{,}600 \times 10^{-6}$/°C	4×10^{-6}/°C

Source: Greenwood, J. C. 1988. Silicon in mechanical sensors. *J Phys E Sci Instrum* 21:1114–28.

Sensing Systems, Inc. (United States), Silicon Microstructures, Inc. (United States), Kavlico (United States), Dalsa (United States), Merit Sensors (United States), Measurement Specialties, Inc. (United States), GE Nova Sensor (United States), Thales Avionics (France), MEMSCAP AS (Norway), Endevco (United States), and GE Druck (United Kingdom).

Automotive Pressure Sensor Examples

Introduction

Historically, pressure sensors in the automotive industry were used predominantly in engine management, e.g., for the measurement of manifold air pressure (MAP) at up to 5 bar. Silicon MEMS pressure sensors have been used in the automotive industry for many years; Bosch, for example, introduced its first Si pressure sensor for the automotive market in 1993. A MAP sensor determines air intake pressure to provide feedback on the correct air-fuel mixture for pollution control and fuel efficiency. Barometric atmospheric pressure (BAP) sensors compensate the air-fuel mix according to altitude. Fuel vapor pressure monitoring is a mandatory requirement on new U.S. vehicles (since 2006). Oil pressure sensors can be either ceramic capacitive, silicon strain gauges (bonded and fused to membranes), or piezoresistive, with silicon chips in silicone oil-filled reservoirs and capped by stainless steel diaphragms. Also, pressure sensors are used in high-pressure fuel injection systems, coolants, braking, and so forth.

A new market opportunity for Si pressure sensors is the direct tire pressure monitoring system (TPMS), a mandatory requirement on new U.S. vehicles (see below). Side airbags and passenger occupation detection are two additional required applications. The pressure sensor competes against accelerometers in the first case and with an infrared low-resolution imaging array in the latter. Other solutions for passenger detection include iBOLT from Bosch, which uses Hall-effect sensors. Pressure sensors react faster in the event of a side impact and are available on the BMW 5 series cars.

Automotive pressure sensor sales are projected to grow at 10% as a result of increased deployment in engine management applications such as BAP and TPMS. Side airbags and occupancy detection systems, where several different solutions are expected, will also help spur sales of automotive sensors. TPMS is driven by legislation and was expected to exceed $200 million in 2009.

Manifold Absolute Pressure Sensor

A prime example of a Si pressure sensor in the automotive application is the MAP sensor, used to control the air-fuel ratio in the engine's management system. Prices, in 2004, varied from about $3 to $5 in volume. Although MAP feeds into a major market, the application is on the decline. Because of a desire for a more direct and accurate measurement technique, MAP sensors are being replaced by mass air flow (MAF) sensors, which provide a more direct control of the engine. The major drawback to MAF sensors at this point is their high cost (e.g., about $30 for the hot wire anemometer from Hitachi), a significantly higher price than for MAP sensors (<$5). Several micromachining options are being considered in an attempt to reduce the cost of an MAF sensor, but none has yet led to a promising approach. Bosch introduced a thin-film equivalent to a hot wire anemometer in 1995. Several organizations are searching for a MEMS version of such a thin-film anemometer. In MAF-based systems, the MAF sensor replaces the MAP sensor in conjunction with a BAP sensor needed to correct for altitude. These BAP sensors are often identical to MAP sensors. Obviously, losing the MAP sensor application does not mean losing all the pressure sensor applications in engine control. In Figure 10.23, we illustrate the Bosch BAP sensor mounted in a premold housing.

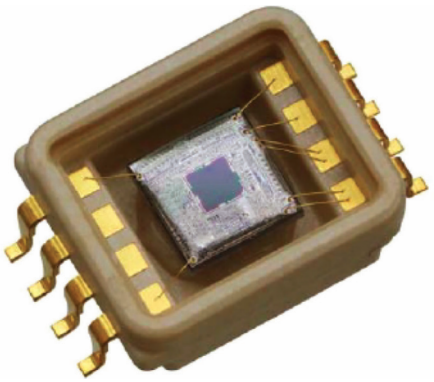


FIGURE 10.23 A barometric air pressure sensor in a premold housing by Bosch Sensortec.

Tire Pressure Monitoring Systems

One of the fastest growing new applications of MEMS pressure sensors is in Tire Pressure Monitoring Systems (TPMS). The U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA) has mandated a new safety standard to warn the driver when a tire is significantly underinflated. The mandate required manufacturers to install a four-tire TPMS that is capable of detecting when a tire is more than 25% underinflated and warning the driver (mandated compliance of 100% for passenger cars was in September 2007). This mandate had its origin in 2001, with a series of tire problems experienced by Firestone/Ford (100 accidents). Two TPMS types are available. The first is based on an indirect measurement: slight changes in wheel diameter in response to pressure loss are detected and computed by the changes based on wheel speed measurements provided by the antilock braking system (ABS). The second type of system is based on direct measurement with a pressure sensor, motion switch, and wireless transmitter installed in the tire's valve stem or wheel to make the measurement (Figure 10.24). The sensor sends information on inflation pressure levels by radio signals to a receiving unit inside the vehicle and a dashboard indicator alerts the driver. The direct method is considered to be more accurate than the indirect approach, but the advantage of greater accuracy is offset by a cost of $65 to $80 per vehicle, since a market-friendly price for a system consisting of a pressure sensor, electronics, and a wireless transmitter is $12–$15 per wheel. High-volume manufacturing should reduced the price to below $10. Sixty million TPMS units were expected in 2008 for the U.S. market alone.

Whereas the TPMS technology today involves a unit mounted in the wheel powered by a battery, in the near future we will see the arrival of the intelligent tire. In this case the TPMS unit will be mounted in the tire itself instead of on the wheel. This will represent a tremendous market opportunity: from 2010 onward, 200 million tires per year were expected on new cars and a 1 billion per year replacement tire market (two or three sets in lifetime). By that time, it was anticipated tires would feature an embedded battery-less pressure sensor/wire antenna for continuous monitoring, and additional data on wear and traction [feed data for the electronic stability program (ESP)] will be available. Interestingly, in this scenario the responsibility for safety of the passenger will shift from car to tire manufacturer.

WTC estimated that the market for TPMS systems would grow at 50% in 2008 due to the U.S. mandate described above before leveling off in 2009–2010 to a growth rate of under 10%. Technology options are resistive and capacitive pressure sensors and SAW devices. Infineon/Sensonor and GE Novasensor are the traditional leading MEMS sensor manufacturers for this application. Whereas two companies were competing in this market in 2005 (Infineon/Sensonor and GE Novasensor), as of 2008, fifteen companies had products available. Other players are

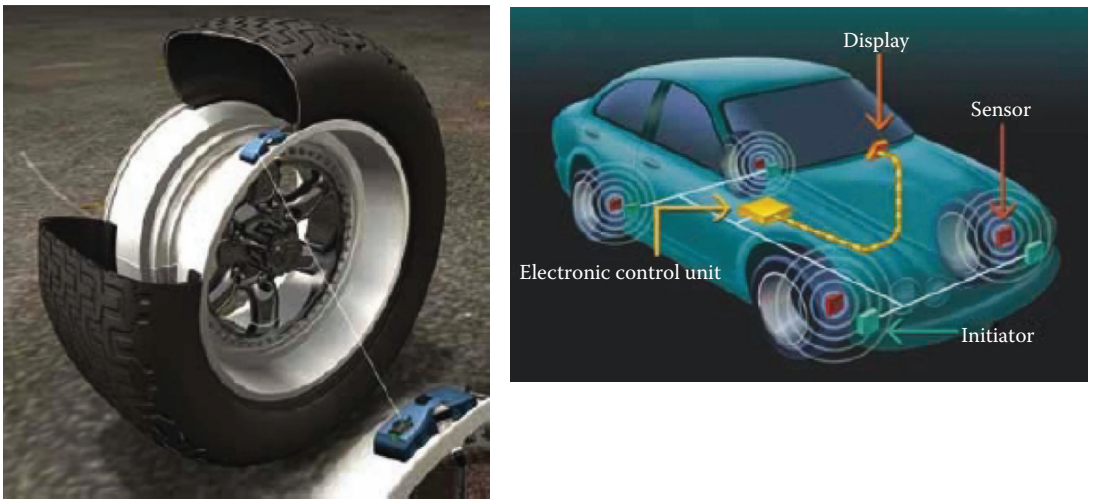


FIGURE 10.24 Microelectromechanical system pressure sensor with wireless data transfer in rim of tire.

TI, Freescale, Bosch, Melexis. Kavlico, Continental, Schraeder, Siemens VDO, and Philips (not the sensor chip). Many other companies have products under development (e.g., at LVsensors, http://www.lvsensors.com).

Important future automotive pressure sensor developments include high-pressure sensors, such as those required for cylinder pressure sensors (one per cylinder). The latter will be required for direct pressure determination inside the combustion chamber, starting with diesel engines. Lower emissions and improved fuel economy are expected to justify the higher cost. The extreme pressure rules out silicon, and in-cylinder sensors may alleviate the need for MAP and air intake MAF devices altogether. Honeywell has licensed fiber-optic pressure sensing technology for this demanding application.

Comparison of Manufacturing Approaches for Pressure Sensing

The majority of Si MEMS pressure sensor processes today are CMOS-compatible. About 55% of production is on 6-inch wafers, and this is expected to increase to 80% in 2010.[15] Manufacturing on 8-inch wafers is becoming more and more popular. It is expected that by 2011, 29% of MEMS manufacturing will be on 8-inch wafers. In 2007, Omron, Freescale, Silex, and DNP all announced new investments in 8-inch manufacturing. The pressure sensors market in the automotive world is under tremendous pressure to reduce cost. Several manufacturers have set up their production centers in Asia Pacific regions because of the lower production and labor costs.

The manufacturing approach for pressure sensing and, to some extent, for all Si micromachined sensors for large industrial market applications where cost is the main issue must incorporate as many of the following attributes as possible:

- CMOS compatibility because of the increasing demand for digital-compatible (see also resonant beam approach, described below), highly integrated Si smart sensors
- Six-inch compatible because of IC trend/availability (and 8-inch silicon MEMS is coming up fast)
- As little fixturing as possible
- Al- (in the future, this might become Cu instead) and IC-line-compatible etchants
- Generic, making it easy to implement a different sensing range or different package or even a different sensor type
- Highest accuracy potential (favors capacitive and resonant devices; see below)
- Wafer-level isolation of electrical connections, making it easier to use in harsh media (removing sensor metallization from the media themselves)
- Dry etching whenever possible
- Single-sided processing if feasible

Polysilicon now holds the cost advantage over a silicon-on-insulator (SOI) approach, but polysilicon will not readily lead to a high accuracy method. Like SOI, bulk micromachining constitutes a more secure option to meet the accuracy demands of future pressure sensor generations but often involves too many glass and/or Si wafers stacked up in one device to allow for an economic solution. Moreover, bulk micromachining is a more expensive technology and up to recently less CMOS compatible than polysilicon. We say up to recently, since Bosch succeeded in making bulk micromachining more CMOS compatible [see the Bosch advanced porous silicon membrane (APSM) process in Volume II, Chapter 4, (Figure 4.77)]. SOI might be able to compete better with polysilicon, as clever designs could obviate the complicated multiple wafer solutions that characterize today's bulk micromachining stacks for pressure, acceleration, and gyros (a three- to five-wafer stack design is not uncommon). SOI technology will also become more and more available and less expensive as it becomes more commonly used in the IC industry. We believe that the SOI approach will become a competitive technology for low-cost, high-accuracy sensors without the reproducibility and long-term stability issues of the polysilicon material.

SOI wafers are finding increasing applications for pressure sensor devices, especially for industrial applications. Most pressure sensors still are bulk micromachined, so wet etching remains a key process. Packaging is a large part of the component cost, and wafer bonding for wafer-level packaging (WLP)

is a growing trend (see Chapter 4, this volume, on packaging, assembly, and self-assembly).

We can summarize that, for now, an inexpensive solution for a pressure sensor with moderate accuracy is based on polysilicon. In the long run, a single-crystal approach based on the Bosch APSM process or an SOI approach, perhaps with a resonant design (see next), is more suited for a generic strategy to produce higher accuracy devices.

Piezoresistive, Capacitive, Resonant Beam, and Other Sensor Approaches

Introduction

The arguments we present in this section about what sensor mechanism to choose for a given pressure sensor application pertain equally well to accelerometers and angular rate sensors. The impetus for obtaining higher accuracy and sensitivity in automotive sensors usually hinges on energy efficiency and/or regulatory mandates. This may require a review of the transducing methods, currently still dominated by piezoresistivity. Examples of alternatives include optical techniques such as Honeywell's in-cylinder fiber-optic pressure sensor, the significantly higher sensitivity of both capacitive and resonant approaches, piezoelectric sensing (mainly for high frequency applications), and tunneling current sensing (very sensitive but increased complexity). The inherent digital compatibility of resonant beams may propel them to the foreground for high accuracy and very low-signal sensor applications. Because of the promise held by the resonant sensing application, we detail this technology a bit further here.

Resonant Beam Sensors

Electrical readout in a pressure sensor, an accelerometer, or angular rate sensor may be performed using piezoresistors located in a flexure area, by measuring the change in capacitance of the moving members with respect to the fixed member(s), or by the measurement of frequency change of a small resonant structure. A resonator is a mechanical structure fabricated to vibrate at a particular resonant frequency. The resonant frequencies of such microresonators can be made extremely stable, so they are used as a time base (the quartz tuning fork in watches, for example) or as the sensing element of a resonant sensor. The performance benefits of a well-designed resonant sensor are compared with piezoresistive and capacitive techniques in Table 10.7.[17] A common advantage of resonant beam and piezoresistive is that they do not depend on measurements relative to "fixed" members, as the capacitive approach does. The small signal magnitude (small capacitance change ΔC) of a capacitive sensor means that on-chip electronics are mandatory to avoid errors due to stray capacitance variations. Piezoresistive and capacitive sensing elements must be used with high-precision amplifiers and trimmed analog circuits to accurately sense and condition small changes in the value of the electronic components. These sensitive circuits and small signals can be highly susceptible to fluctuations in temperature, electrical noise, mechanical and electrical drift, temperature and pressure cycling hysteresis, ionizing radiation, and magnetic fields. These factors and others result in measurement inaccuracies, effectively degrading overall precision and resolution. As a result, best-case production performance of solid-state sensors generally has been limited to a long-term accuracy of about 0.1%, or 1000 ppm.

Resonating devices, on the other hand, are affected primarily by the mechanical characteristics of the structure, not by its electrical behavior. In addition, since the system itself measures only frequency, the interface to a digital system vastly is simplified. In practice, resonant sensors can attain measurement accuracies at least 10 times higher than conventional sensing methods. The fabrication of resonant structures, however, is more complex, and the requirement for packaging such devices more demanding than for piezoresistive or capacitive sensors.

The basic building blocks for a resonant beam structure embedded in a pressure-sensitive membrane are shown in Figure 10.25.[18]

The specific structure shown here is one by the Yokogawa Corporation. To operate as a differential pressure sensing device, this chip incorporates two H-shaped bridges, located in the Si substrate and operating at the same identical frequency (90 kHz). Both bridges (resonators) are located at the same strain-free plane (depth) from the chip's surface, but in different locations. A resonant sensor is designed

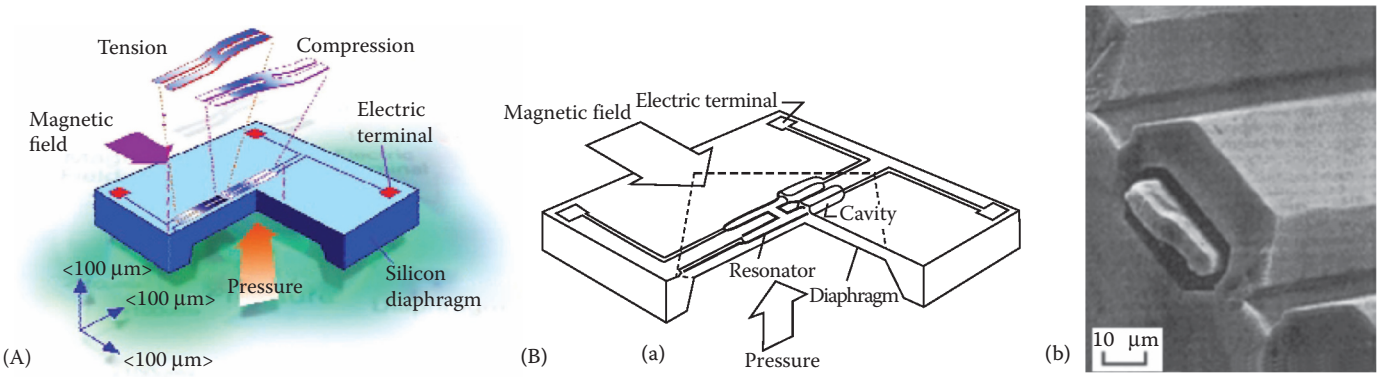


FIGURE 10.25 (A) Digital sensor technology by Yokogawa Corporation. (B) Construction of the sensor (a) and cross-sectional scanning electron microscope photograph of the resonator (b).[18]

such that the resonator's natural frequency is a function of the measurand (pressure in this case), and typically the measurand alters the stiffness, mass, or shape of the resonator, hence causing a change in its resonant frequency. In the current case, when pressure is applied, because of their position and exact depth in the substrate, one bridge goes into tension and the other into compression. The other important features of a resonant sensor are the vibration drive and detection mechanisms. The Yokogawa chip incorporates an electromagnetically driven resonant beam sealed in a vacuum shell, integrated onto a standard silicon pressure sensor diaphragm (an SOI approach is used). The drive mechanism excites the vibrations in the structure while the detection mechanism senses these vibrations. The frequency of the detected vibration forms the output of the resonant device. As pressure is applied the compression bridge changes from 90 to 110 kHz and the tension bridge from 90 to 70 kHz. To obtain a differential output the microprocessor simply counts the change in bridge frequencies (Δ40 kHz). In the case of gauge pressure measurement, both bridges are used. One bridge measures the reference pressure (atmospheric or head pressure), and the other one measures the applied pressure (gauge).

With the digital pressure sensor pictured in Figure 10.26, it was demonstrated that an accuracy of 0.01%, or 100 ppm, is attainable. The device also offers wide dynamic range, high sensitivity, and high stability, along with a frequency output for easy interface to digital compensation circuitry. The chip fabrication is challenging though, comprising four sequential selective epitaxial growth and critical selective etching steps, and the sensor package is extremely complex and therefore expensive to produce. It is, nonetheless, an excellent example of the power of silicon microfabrication technology

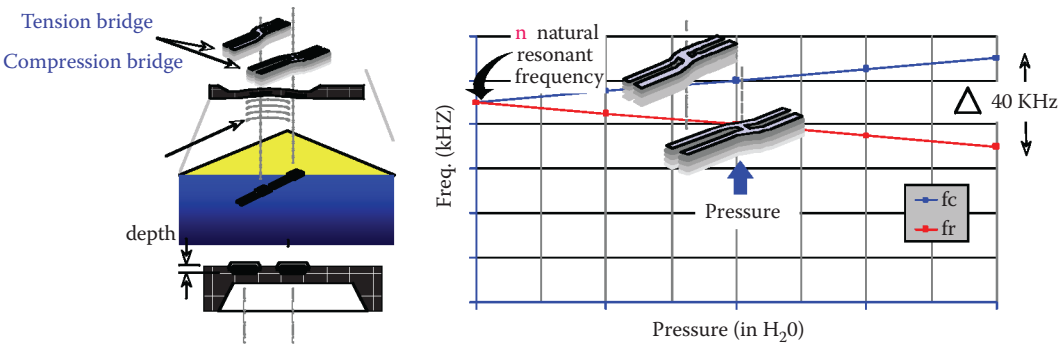


FIGURE 10.26 Operating principle of the digital pressure sensor. The first bridge is located directly in the center of the chip and the other nearer the edge. As pressure is applied, the center bridge will be in a state of "compression" and the outer bridge in a state of "tension." As a result, the natural frequency of the two bridges changes. The compression bridge frequency increases and the tension bridge frequency decreases, thereby effectively creating a differential frequency. The two frequencies are easily managed by a CPU as a time-based calculation directly proportional to the applied differential pressure. (From Yokogawa Corporation of America.)

and represents a major technical achievement. Morover, Yokogawa's application represents the industrial sensing and control business, which is aimed at high-performance sensors that can sell for higher prices in modest volumes.

GE Druck also developed and incorporated a resonant silicon sensor in its pressure transducers (http://www.gesensing.com/druckproducts), as well as GE NovaSensor.[19] Zook et al.[20] fabricated clamped-clamped beams in a polysilicon surface micromachining process and, by locating the polysilicon beam in a high-vacuum polysilicon shell, avoided damping. Surface micromaching simplified the manufacturing process over previous attempts in SOI by Ikeda et al. shown in Figures 10.25 and 10.26[21] and may be able to produce cost-effective resonators as replacements for piezoresistors. The sensitivity to applied axial loads of the beam exceeds the gage factor of single-crystal piezoresistors easily by a factor of 10–100. Devices with sensitivities as high as 700 Hz/g have been reported.[22]

Today, the resonant beam approach remains the least-developed sensing mode, despite the fact that the resonant beam method showed very good potential in the past (although expensive and difficult to implement). So it is interesting that there is not more commercial activity to validate this sensing approach further and making these resonant sensing devices less expensive. It could be that one of the main reasons why direct frequency sensing is not investigated more is the fabulous resolution ADI achieved with its capacitive sensing scheme. Because the signal conditioning is so tightly coupled to the sensing element, the integrated ADI sensors can measure extremely small changes in differential capacitance. As an example, the capacitance resolution of their inexpensive ADXL202E, a surface-micromachined 50-mm^3 accelerometer, is as low as 20 zeptofarads (20×10^{-21} F) (see next section).

Accelerometers

Introduction and Market Perspectives: Victim of Success

WTC estimated that in 2006, the silicon-MEMS market in the automotive sector was $1.6 billion (this firm only counted flow, pressure, acceleration, and rate),

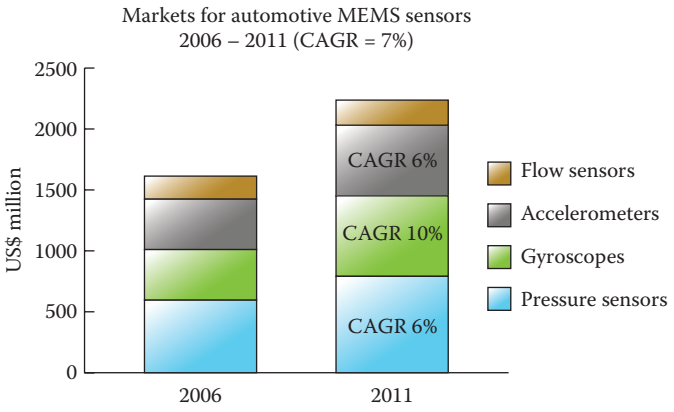


FIGURE 10.27 Markets for automotive microelectromechanical system sensors, according to WTC.

making this the second biggest MEMS opportunity after RW-heads and ink-jet print heads. By 2011, the same market was pegged at $2.2 billion, with a CAGR of around 7%. The main applications in revenue terms were, in order, pressure sensors, gyroscopes, accelerometers, and flow sensors, as shown in Figure 10.27.

The total number of Si sensors WTC claimed would grow from over 430 million units in 2006 to 780 million in 2011, an annual growth of 13% (for a reference point, consider that by 2006, Bosch had produced 600 million sensors, 130 million units in 2006 alone). This outstrips the dollar growth because of price erosion, running at 4–5% per year.

Yole estimated the 2005 worldwide accelerometer market at $393 million, representing about 8% of the overall silicon-MEMS market. By 2010, they expected the accelerometer market to be $869 million, for a five-year CAGR of 17%.

Major players in this space include Bosch, Infineon, ADI, Denso, and Delphi. Since the early 2000s, manufacturers have been putting accelerometers in consumer electronics such as laptops [it started when IBM (now Lenovo) put a dual-axis accelerometer in its notebook in 2004], cameras, and cell phones. WTC fears that in coming years the accelerometer market will suffer from price pressures reminiscent of that in consumer electronics (10% per year). Accelerometers might become victims of their success!

Accelerometers signified the next big Si micromachining entry in the commercial world after the Si pressure sensor. Today, polysilicon surface-micromachined capacitive and single-crystal Si bulk-micromachined accelerometers have taking over a

large share of the automotive acceleration market. Manufacturing volumes of tens of millions a year have brought down the price to less than $2 per sensor. A company such as ADI makes 2 million surface-micromachined accelerometers a week.

A major difference between accelerometers and pressure sensors in cars is that less diverse types of accelerometers are needed. Pressure sensors require various ranges and many different packages to accommodate the sensors in their unique environments. Accelerometers, on the other hand, require little difference in packaging; two main types (i.e., high *g* and low *g*) cover all needs. This explains why the first surface-micromachined commercial Si sensor was an accelerometer rather than a pressure sensor.

The first demonstration of a capacitive Si micromachined accelerometer took place in 1979 at Stanford University.[4] Over the years, a variety of accelerometers based on different physical detection schemes, such as piezoelectric, piezoresistive (thin- and thick-film), capacitive, tunneling, and resonant members, have been attempted in a wide variety of manufacturing schemes. As listed in Table 10.4, in cars, accelerometers are used as crash detection sensors both in the electronic control unit (ECU) and in peripheral sensors such as upfront sensors (UFS) or pedestrian contact sensors close to the bumper, peripheral acceleration sensors (PAS) in the doors, and rear crash sensors.

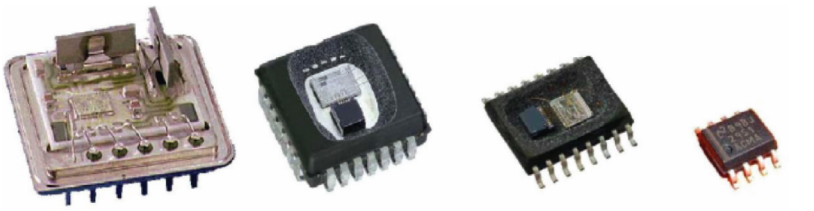
Automotive Accelerometer Examples

Airbag Accelerometer

The principal mass market for Si acceleration sensors in the automotive industry is in crash detection and airbag release systems. A MEMS-based accelerometer detects with great precision the sudden slowing of a vehicle and fires the airbag when needed. Now produced in the millions per year, these devices have gotten smaller, cheaper, and more capable, riding the learning curve originally promised by MEMS proponents. This learning experience by Bosch-Sensortec is illustrated in Figure 10.28 as a sequence of product generations for automobile airbag systems introduced over time. The MEMS elements are not only getting much smaller than the formerly used fine-mechanical piezoelectric device (shown in Fig. 10.28a) but also much more cost efficient, and yet they include an increasing number of functions (such as self-tests and accuracy check routines).

In 1991, the Intermodal Surface Transportation Efficiency Act (ISTEA) was signed into law. ISTEA airbag requirements ensured that by 1998, 100% of the new U.S. vehicle fleet be equipped with airbags. System integrators like Autoliv expect airbag accelerometer penetration to increase from 40 to 60 million vehicles over the five-year period between 2006 and 2011 (i.e., 80% of cars worldwide). WTC estimated that this penetration equals an insertion of 200 million accelerometers in 2006 going up to 350 million in five years, with as many as eight accelerometers per electric control unit (ECU).

An earlier generation of airbag sensors used in the United States consisted of electromechanical types supplied by Breed and TRW (see Table 10.5). Three electromechanical accelerometers discriminated between zones according to the severity of a vehicle crash. Two sensors were used in the crush zone, usually on frame rails behind the front bumpers or on the lower portion of radiator supports. These sensors determined the severity of impact and activation of the airbag.



1984 – 1997	1997 –	2002 –	Coming soon
Hybrid piezoelectric	Silicon micromachined capactive	Silicon micromachined capactive	
(a)	(b)	(c)	(d)

FIGURE 10.28 Generations of high-*g* acceleration sensors used in automotive airbag systems at Bosch.

Another sensor was located outside the crush zone, either in or near the passenger compartment. This sensor, often known as the *safing sensor*, prevented deployment of the airbag in the event someone inadvertently impacted (or tampered with) the front sensors.

Micromachined accelerometers displaced these electromechanical devices, as they made single-point sensing possible. Single-point sensing requires only one accelerometer, located in the passenger compartment, reducing the number of sensors and associated wiring needed in the system. In addition to reducing the number of sensors required in each system, surface-micromachined accelerometers from ADI were costing approximately $4 per unit in volume in 1998 (and below $2 today), significantly less than electromechanical accelerometers at approximately $15 per unit.

The early bulk-micromachined Si accelerometers lacked damping and overload protection until TRW NovaSensor introduced a piezoresistive accelerometer combining a microactuator for reliable self-testing and overload protection on the same chip. This Si bulk-micromachined ±50-*g* accelerometer based on hybrid technology, with signal conditioning electronics on a separate chip, bonded into a TO-5 header, is shown in Chapter 4, this volume, Figure 4.9.[23] In 1991, ADI introduced the first commercial product to use polysilicon surface micromachining, the so-called *ADXL-50*, a capacitive accelerometer with an electrostatic actuator built in for self-testing. The construction of this sensor was studied in Volume II, Example 7.9 (see also Volume II, Figures 7.92–7.96).

More and more vehicles are being equipped with side airbags, ensuring continued growth of the accelerometer market. Primarily ADI, GM/Delco, and Motorola supply silicon accelerometers in the 50-*g* range for U.S. vehicles. Sensonor (Norway) has historically provided a large portion of the European market, while Nippondenso is the major provider in Japan. The ADI accelerometer was the first integrated sensor; most other accelerometers on the market still use a multichip solution to sense and provide the appropriate signal condition.

Suspension Accelerometers

ADI, between 1990 and 2000, branched out from their high-*g* range of airbag (50 *g*) accelerometers, to medium and low-*g* sensors for suspension and stability control devices (say, 0–2 *g*).

Silicon accelerometers in suspension systems face more hurdles than the Si accelerometers in airbag systems. Low *g* is more difficult (signal-to-noise limit), signal conditioning is required with each remotely mounted package (one suspension sensor per wheel), and the sensor is exposed to the harsh automotive environment, i.e., more complex packaging. Many companies, including ADI and Motorola were and still are aggressively pursuing this application. Low-*g* accelerometers are also of utility in antilock braking systems (ABS) featuring traction control. No federal law mandates ABS/traction control, and because of the cost constraints, only high-end and midsize cars are expected to utilize the system.

In 1995, using its patented iMEMS® (integrated MEMS) approach, ADI introduced a 5-*g* accelerometer, pushing the polysilicon surface micromachining technology it pioneered with their 50-*g* device (the ADXL50) a bit further. Also out of ADI, in 1998, came the breakthrough 2-*g* dual-axis ADXL202E accelerometer, an inexpensive, surface-micromachined accelerometer capable of measuring both static (gravitational force) and dynamic (vibration and shock) forces, pushing the surface micromachining sensor technology even further. This sensor detects movements on two axes and outputs digital signals to microcontrollers as simple as 4-bit microcontroller units (MCUs). A single moving "beam" in the ADXL202E has "fingers" that are used to detect beam movement (through capacitance change) in two directions, and this goes through an on-chip algorithm and is sent along as digital data. In Figure 10.29a we show how silicon is used to make the spring and proof mass and how a finger is added to make a variable differential capacitor. The beam is made up of many interdigitated fingers, and each finger can be visualized as shown in this figure. The electronics measure the change in displacement by measuring the change in differential capacitance against a set of fixed fingers (plates). The critical dimensions of fingers and the proof mass (0.7 μg) of this sensor are clarified in Figure 10.29b. The finger height (2 μm for the example product) is fixed by process technology, while the overlap (125 μm here) is adjustable to

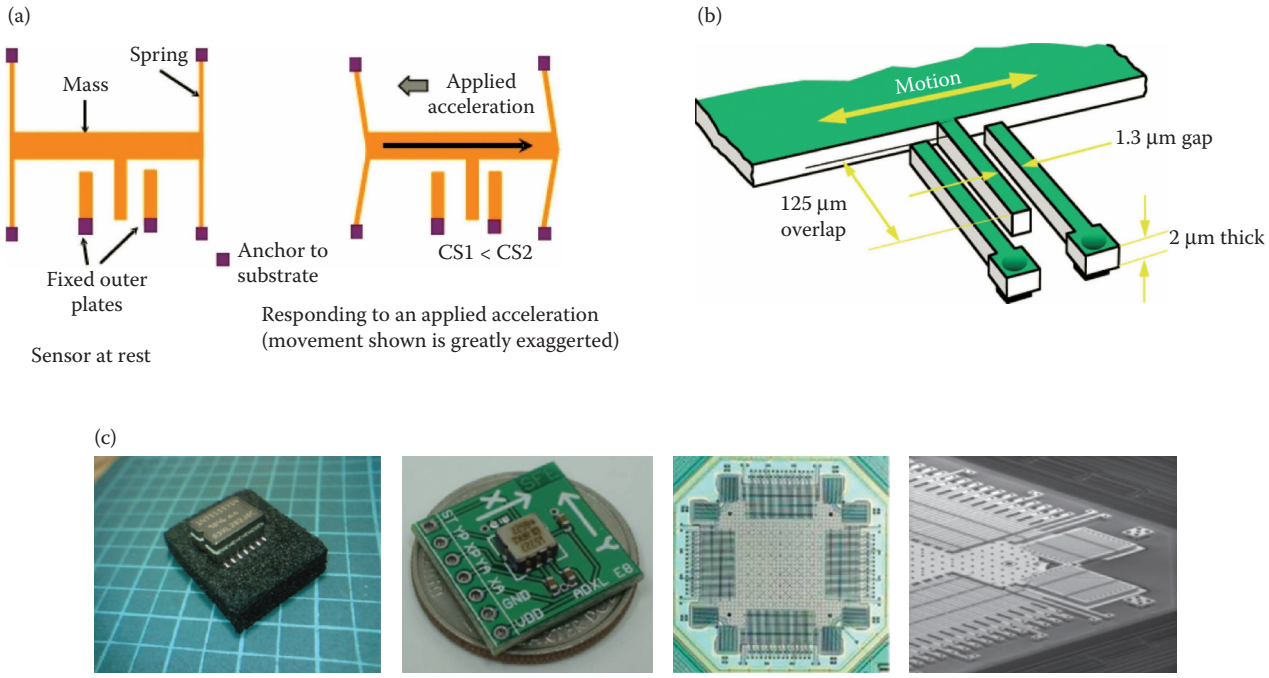


FIGURE 10.29 Analog Devices' ± 2 *g* dual-axis ADXL202E accelerometer is an inexpensive, surface-micromachined accelerometer capable of measuring both static (gravitational force) and dynamic (vibration and shock). (a) Silicon is used to make the spring and proof mass, and fingers are added to make a variable differential capacitor. (b) Critical dimensions and proof mass weight and dimensions for a single finger. The proof mass is 0.7 µg. The accelerometer has some impressive numbers: 0.1 pF per side for the differential capacitor, 20 zF (10^{-21} F) smallest detectable capacitance change, and 2.5 pm minimum detectable beam deflection (one tenth of an atomic diameter). (c) In the ADXL202 beam structure, note the four pairs of orthogonal serpentine springs at the corners of the proof mass. The entire mechanical structure is suspended from the four points where the springs meet.

some extent. Longer fingers are not desirable, because they are harder to manufacture and result in larger beam areas, which in turn translates to more expensive sensors. The fingers are released when the oxide layer between the underlying bipolar IC and the beam is etched away, leaving a 1.6-µm gap with fingers spaced every 1.3 µm. Rather than using two discrete beams placed orthogonally, the fingers that constitute the *x*- and *y*-axes, variable capacitors are integrated along the sides of a single square beam. This results in a reduction of the overall sensor area, yet the larger common beam mass enhances the resolution of the ADXL202E. The differential capacitance of each finger is proportional to the overlapping area between the fixed outer plates and the displacement of the moving finger. Because these are very small capacitors, one needs the largest practical differential capacitance to reduce noise and increase resolution. The fixed outer plates are driven with square waves that are 180 degrees out of phase. When the movable fingers (and hence the beam) are centered between the fixed outer plates, both sides of the differential capacitor have equal capacitance and the AC voltage on the beam is zero. However, if the beam is displaced due to an applied acceleration, the differential capacitance will be unbalanced, and an AC voltage of amplitude proportional to the displacement of the beam results. In this sensor movements as minute as 0.02 Å are detectable. Because of the combination of low cost and high performance acceleration/tilt sensing possible with this device, other applications have become possible. The ADXL202E is currently used in car alarms (for jacking or towing sensing), pen-based handwriting recognition systems, portable electronic games, and wearable sports equipment.

The ADXL202E was an early demonstration that there is a path for MEMS from automotive applications to consumer electronics. The average price of MEMS accelerometers across all applications is decreasing, from an average of $2.50 in 2004 to less than $1.90 in 2009, with consumer applications driving the price erosion. The average selling price (ASP) for a three-axis MEMS accelerometer was still

approximately $2.00 in consumer products in 2006, while lowest cost products were already shipping at under $1.50 in 2006. It was expected that by 2009, the average selling price would drop to $1.35, and the least expensive (first-level packaged) three-axis accelerometers would ship for under $1.00.

Comparing Manufacturing Options for Accelerometers

In the automotive market 80% of accelerometers are manufactured using deep reactive ion etch (DRIE) equipment. Growth in the accelerometer market overall actually is driving growth of DRIE equipment sales. The primary wafer size for accelerometers is currently 6-inch (accounting for more than 80% of device production), but 8-inch wafers are expected to make inroads and were projected to account for nearly 10% of production by 2010. Devices are typically processed on silicon, with SOI wafers in use or in development by the major manufacturers.

Obtaining sufficient sensitivity was originally thought to be a potential problem when using polysilicon to manufacture low-*g* accelerometers. To make a large proof mass is especially challenging for polysilicon accelerometers. Process control and handling considerations limit the minimum flexure thickness for suspending the inertial mass to about 2 µm and, given the relatively low density of Si, obtaining deflection at 1 or 2 *g* without making the proof mass large is difficult. Realizing a large proof mass directly affects chip size and thus cost, presenting the developer with an unattractive option. The proof mass of a bulk-micromachined accelerometer may be 0.1 mg, whereas a typical value for a surface-micromachined device is 0.5 µg, 200 times smaller. With a source capacitance in the neighborhood of 150 femtofarads (i.e., 0.15 pF), and orders of magnitude smaller differential capacitance changes, say 100 attofarads (10^{-18} F) for a 1-*g* acceleration, a resolution of 23 zeptofarads (10^{-21} F) is required, and the signal output of these devices easily could be lost in the presence of parasitic capacitance or noise. Despite these early misgivings expressed here, the surface-micromachined ±2-*g* capacitive accelerometer announced in 1995 by ADI demonstrates that low-*g* polysilicon sensors and cost-effective integration of all the electronics is quite possible. It is actually within ADI's abilities to provide on-chip signal conditioning that allows the minute signals to be read with minimal interference. A synchronous demodulation technique ("AC bridge") is used to increase noise immunity and improve the signal to noise ratio. A square-wave carrier drives the sensor, and the modulated signal produced by the change in differential capacitance is amplified and synchronously demodulated, minimizing noise and boosting the signal-to-noise ratio.

In silicon bulk-micromachined accelerometers, the formation of a mesa structure (proof mass) from a single crystal Si is more complicated because of rapid and irregular etching of convex corners by orientation-dependent etchants, such as ethylene diamine pyrocatechol (EDP) and KOH. Corner compensation schemes must be invoked to avoid irregular corners (see Volume II, Chapter 4).

Critics of the capacitive polysilicon accelerometer still maintain that the integration of electronics is too complex and that a bulk-micromachined sensor requires only twelve mask levels, whereas the polysilicon device requires twenty-eight mask levels. Along this line, major Japanese companies have chosen single-crystal solutions for building accelerometers to avoid the difficulties associated with reproducing polysilicon material properties and the possible unknown long-term instabilities.

The pros and cons of surface micromachining (see Volume II, Chapter 7) can then be summarized as follows:

Pros

- A surface-micromachined device has a built-in support, resulting in cost effectiveness.
- Surface micromachining only requires one-sided processing compatible with conventional IC processes.
- Devices made by surface micromachining are smaller in size, have higher yields, and accommodate more on-chip circuitry.

Cons

- The out-of-plane dimensions in surface micromachining in general is limited to ~2 µm, hence the light mass of the proof mass, resulting in a low sensor sensitivity.

- The suspended Si structure is more compliant (due to the high ratio of length to thickness of the device structure) and hence has a tendency to stick to the support substrate, especially in overdrive situations.

As we made clear in the section on pressure sensing, we believe that SOI surface micromachining combines the best features of surface micromachining with the best features of single-crystalline Si. The same benefit befalls the Bosch IMAPS process, which enables electronic integrations just as in the case of ADI (see above). Using dry etching to etch the proof mass from an epimaterial avoids corner compensation issues; the epimaterial ensures larger mass and single-crystal performance. This approach might overtake both surface and bulk micromachining as SOI wafers become less expensive and more available. A generic design of a resonant beam structure in an SOI wafer may be possible. This design could constitute the basis of a wide family of pressure, acceleration, and angular rate sensors.

MEMS Angular Rate Sensors

Introduction

First, some clarifications are in order about the definitions of an angular rate sensor, gyros, and accelerometers. An angular rate sensor is not a gyro, but it is common jargon to refer to an angular rate sensor as an *electronic gyro* or even simply a *gyro*. In reality they are quite different, although one can be made to act like the other. Angular rate sensors measure angular rate (surprise!), i.e., how fast (degrees per second or radians per second) something is rotating. A gyroscope, on the other hand, is used to measure angular position (not rate), using the principle of "rigidity in space" of the gyroscope (see Figure 10.30). For example, this is how the attitude indicator (artificial horizon) of an aircraft works. A gyroscope can be made to measure angular rate (using the principle of precession), but that requires a finely calibrated measuring spring. Conversely, an angular rate sensor can be made to measure angular position by integrating the signal over time. So, an angular rate sensor can be made to act like a gyro, and a gyro can be made to act like an angular rate sensor, but they do not act the same by themselves. Lastly, an accelerometer is nothing like a gyro and cannot be made to act like one. An accelerometer measures linear acceleration. Using calculus, this can be integrated into linear velocity and position. The rotation typically is measured in reference to one of three axes: yaw, pitch, or roll (see Figure 10.30). The difference between an accelerometer and an angular rate sensor is that the gyro measures the angular acceleration (how fast the rate is at which an object rotates changes) and the accelerometer measures how much something is accelerating (not angular).

As an example of how an angular rate sensor could be used, a yaw-axis angular rate sensor that is mounted on a turntable rotating at 33 1/3 rpm (revolutions per minute) would read a constant rotation of 360° times 33 1/3 rpm divided by 60 seconds, or 200°/s. In the case of an electronic angular rate sensor, a voltage proportional to the angular rate, as determined by its sensitivity, measured in millivolts per degree per s (mV/°/s), would be the output. The full-scale voltage range determines how much angular rate can

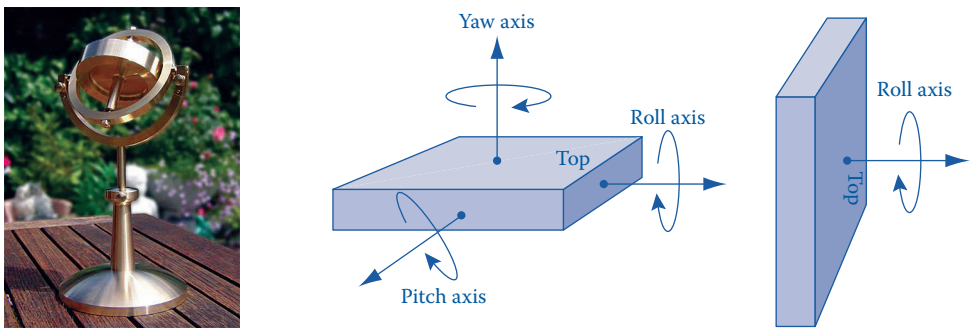


FIGURE 10.30 Illustration of a conventional mechanical gyroscope and the three rotational degrees of freedom it can measure (http://www.brassgoggles.co.uk/images/brassgyroscope.jpg). Shown are gyro axes of rotational sensitivity. Depending on how a gyro normally sits, its primary axis of sensitivity can be one of the three axes of motion: yaw, pitch, or roll (see also Figure 10.31).

be measured. In the turntable example, an angular rate sensor would need to accommodate a full-scale voltage corresponding to at least 200°/s. The full-scale range is limited by the available voltage swing divided by the sensitivity. The ADI MEMS angular rate sensor we will analyze below, the ADXRS300, comes with 1.5 V full-scale and a sensitivity of 5 mV/°/s, so it handles a full-scale of 300°/s. One practical application of an angular rate sensor is to measure how quickly a car is turning. If the angular rate sensor, mounted in the car, senses that the car is spinning out of control, differential braking engages to bring it back into control. As we saw earlier, the angular rate signal can also be integrated over time to determine angular position particularly useful for maintaining continuity of GPS-based navigation when the satellite signal is lost for short periods of time, for instance, when the car is in a tunnel.

Before satellite-based global positioning systems (GPSs), there was the mechanical gyroscope: a reliable rotation sensor, based on the inertial properties of a rapidly spinning rotor. This mechanical contraption, which has been around since the late 1700s, is used for maintaining a fixed orientation with great accuracy regardless of the earth's rotation, as shown in Figure 10.30. Traditional spinning gyroscopes work on the basis that a spinning object that is tilted perpendicularly to the direction of the spin will have a precession. The precession keeps the device oriented in a vertical direction so that the angle relative to the reference surface can be measured. A rotation in the frame thus imparts a torque (rotation) on the spinning disk, which precesses (rotates) as a result (conservation of angular momentum). In other words, the flywheel's large angular momentum counteracts externally applied torques and keeps the orientation of the spin-axle fixed. Practical applications limit the movement to measure only one axis of rotation (roll, pitch, or yaw; see Figure 10.30). Precision-machined mechanical gyros may cost anywhere between $10,000 and $100,000.

The ring-laser gyroscope, first discussed by Georges Sagnac (1869–1926) in 1913 and demonstrated in 1962 by Warren Macek of Sperry Corporation, quickly displaced mechanical versions for high-end and high-precision applications such as aviation and is now even used in high-end cars. Sagnac discovered that light sent around a closed loop, in two different directions, would show a phase difference between the two beams when the loop is rotated (refer here to the discussion of the Michelson-Morley experiment in Volume I, Chapter 5, "The Lorentz Transformation"). The Sagnac effect is illustrated in Figure 10.31. The ring-laser gyroscope consists of two laser beams traveling in opposite directions (i.e., counter-propagating) around a closed-loop path. The pulse traveling in the same direction as the rotation of the loop must travel a slightly greater distance than the pulse traveling in the opposite direction. As a result, the counter-rotating pulse arrives at

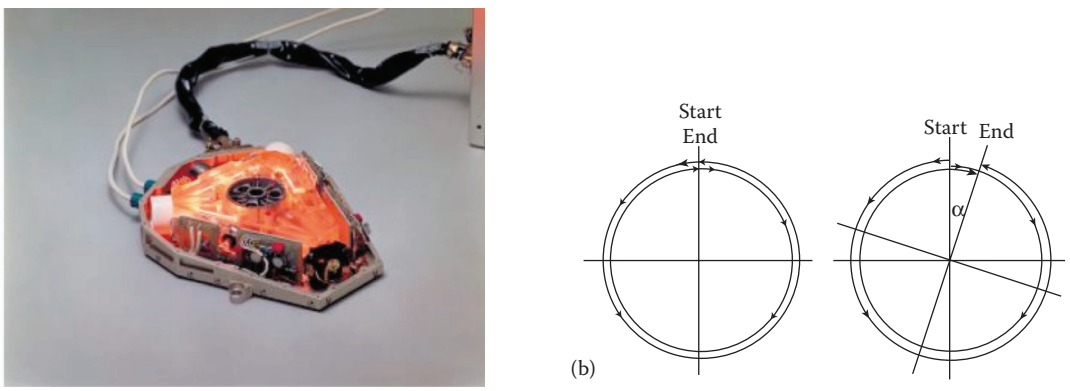


FIGURE 10.31 (a) The RLG. (b) The Sagnac effect: if two pulses of light are sent in opposite directions around a stationary circular loop of radius *R*, they will travel the same inertial distance at the same speed, so they will arrive at the end point simultaneously. This is illustrated in the left-hand drawing. The drawing on the right indicates what happens if the loop itself is rotating during this procedure. The symbol α denotes the angular displacement of the loop during the time required for the pulses to travel once around the loop. For any positive value of α, the pulse traveling in the same direction as the rotation of the loop must travel a slightly greater distance than the pulse traveling in the opposite direction. As a result, the counter-rotating pulse arrives at the "end" point slightly earlier than the corotating pulse.

the "end" point slightly earlier than the co-rotating pulse. The constructive and destructive interference patterns formed by splitting off and mixing parts of the two beams can be used to determine the rate and direction of rotation of the device itself.

The fiber optic gyroscope (FOG) is based on the Sagnac effect and consists of a loop of single-mode optical fiber (often a polarization-maintaining fiber) and related coupler components, a semiconductor laser or superluminescent LED, and signal-processing electronics. The coupler components generally are fabricated in proton-exchanged $LiNbO_3$ integrated-optic circuits. This material is chosen for its ability to modulate the light beam for improved detection. The FOG, being a simpler device, currently is receiving more attention due to its potential to achieve the required performance at a lower cost than with ring laser gyroscopes (RLG) or mechanical gyroscope technology.

These instruments still cost about $1000. For example, the fiber optic gyro from Andrew Corp. (now KVH Industries) costs $950. It is a cylindrical device that measures 77 mm in diameter and is 88 mm long. The higher-end Honeywell Ring Laser Gyro (RLG) has revolutionized the way the world navigates on land, sea, air, and space. It is highly reliable because it has virtually no moving parts, replacing spinning mass with spinning light.

Interferometric FOGs and MEMS angular rate sensors are changing the commercial landscape of gyros once again. We cover only MEMS gyroscopes (angular rate sensors really) here; they are constructed out of silicon (often SOI) and quartz, or they may be ceramic-based.

Market Perspectives

The requirements for inertial sensors, i.e., accelerometers and angular rate sensors, vary drastically depending on the applications they are intended for. Figures 10.32a and 10.32b give an overview of the main performance requirements for accelerometers and angular rate sensors, respectively, in various applications.

In the late 1990s, inexpensive, solid state, micromachined angular rate sensors became the Holy Grail for many micromachinists around the globe. The same companies that made a lot of headway in micromachined accelerometers, such as ADI, are also at the forefront of Si angular rate sensor development today.

In 2005, according to Yole, global MEMS-based angular rate sensors sales were $558 million, and they were expected to grow by a CAGR of 11% through 2010 to reach $930 million. MEMS gyroscopes are included in three automotive applications: Antirollover systems, global positioning system (GPS) navigation, and electronic stability control (ESC). These automotive applications represent more than 70% of the MEMS gyroscope market. ESC is a rapidly growing market as these systems could potentially reach 100% adoption in new cars shortly after 2010. Current navigation system designs use a combination of global positioning (GPS) and CD-ROM maps in addition to wheel rotation sensors and angular rate sensors, or magnetic compasses. Angular rate sensors enhance the accuracy of navigation instruments in situations where satellite reception is lost, as for example in tunnels, between tall buildings, and in mountainous terrain. The cost of current navigation systems is in the $200 to $800 range, and MEMS solutions are expected to reduce this cost further.

As in the case of accelerometers, once gyros make it big in the automotive markets, they also succeed in consumer electronics (CE). The sensor specifications in consumer electronics are much more relaxed than in the automotive application, but price erosion and product cycles are much faster than in the automotive field.

Five consumer applications integrate MEMS gyroscopes: camcorder stabilization, camera stabilization, cell phone stabilization, video games, and navigation for small electronics. In 2006, Nintendo Wii's motion sensing remote control gave, for the first time, public visibility to inertial MEMS. The new attractive human-machine interface in game controls such as the Wii is expected to penetrate the cell phone market as well. Five sensors are used to monitor orientation of the Segway scooter, sampled at 100 times/s. Sensors include MEMS angular rate sensors and liquid-filled tilt sensors.

According to Yole, quartz and ceramic substrates were estimated to account for 37% by volume of total gyroscope production in 2005. But most new gyroscope development programs were silicon-based, so this figure was expected to decrease to 28% in

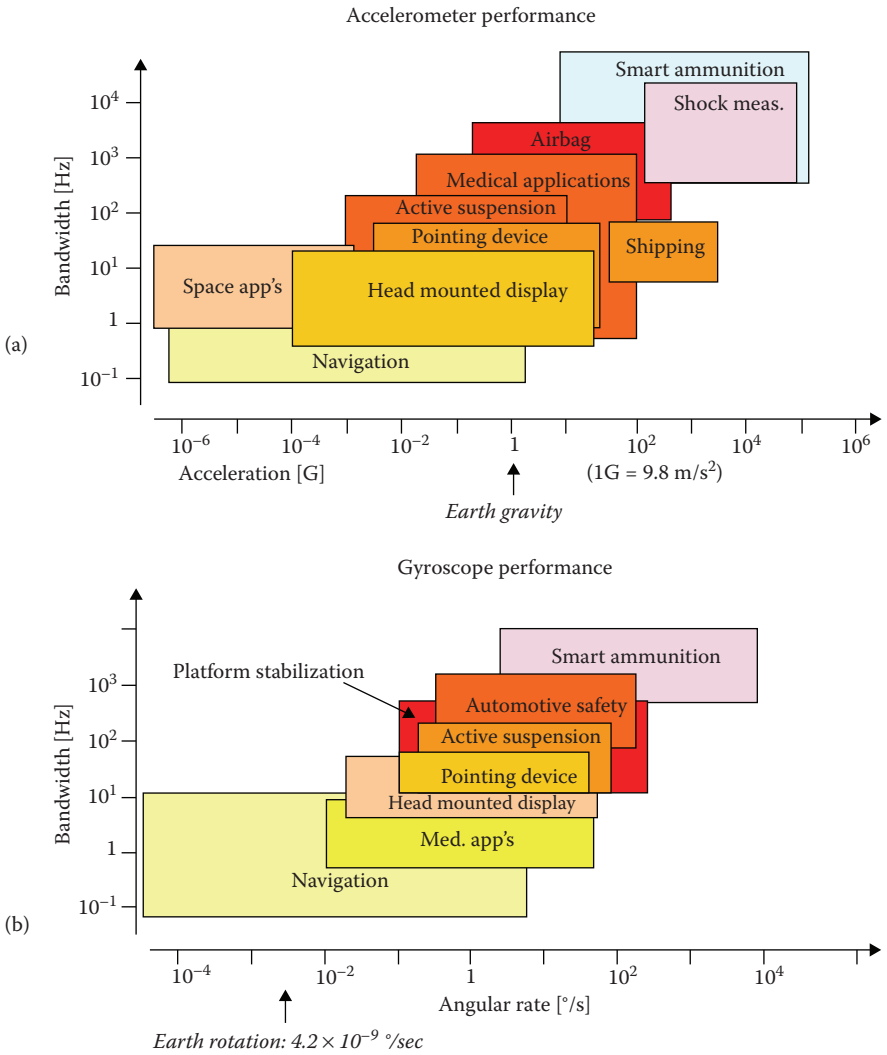


FIGURE 10.32 Overview of applications and required performances for micromachined inertial sensors. (a) Accelerometers; (b) angular rate sensors.

2010. For gyroscopes today, 4-inch wafers are still the dominant technology, but the share of 6-inch wafers is expected to grow to 60% over the next five years. DRIE and SOI increasingly are being adopted in MEMS gyroscope manufacturing.

The average price of a MEMS gyro in defense applications is 4–5 times higher than in automotive applications and more that 15 times the average sales price of consumer applications. Defense and aeronautic applications include missiles guidance, general aviation instruments, platform stabilization, land navigation, and smart munitions in the future. Gyros are made today at ADI, BEI Technologies, Inc., Honeywell, Kionix Inc., Matsushita Electric Industrial Co. Ltd., Motorola Automotive, Murata, Robert Bosch GmbH, SensoNor ASA, Silicon Sensing Systems Ltd, STMicroelectronics, and TRONIC'S Microsystems SA.

MEMS Angular Rate Sensors

Being very small, MEMS gyros obviously cannot operate on the flywheel principle, as that scales with the mass of the flywheel. Even the best MEMS motors still quickly slow down and stop if not externally actuated. The inertial mass of the small MEMS wheels is very, very small (mass moment of inertia, $I = \int r^2 dm$; with a fifth order expression like this, it becomes insignificant very fast when r is made very small). Furthermore, MEMS rotors cannot (thus far) be made so that they precess freely in three dimensions. Instead, MEMS often take advantage of the Coriolis force to detect angular rotation. The Coriolis force, named after Gaspard G. de Coriolis (1792–1843), a French mathematician, is an inertial force that must be applied to maintain a moving body's

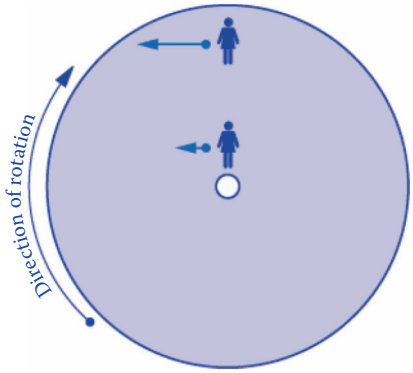


FIGURE 10.33 Illustration of the Coriolis force. A person moving northward toward the outer edge of a rotating platform must increase the westward speed component (blue arrows) to maintain a northbound course. The acceleration required is the Coriolis acceleration. (From 2003. *Analogue Dialogue* 433–37.) New iMEMS angular-rate-sensing gyroscope [John Geen (john.geen@analog.com) and David Krakauer (david.krakauer@analog.com), Analog Devices, Micromachined Products Division.]

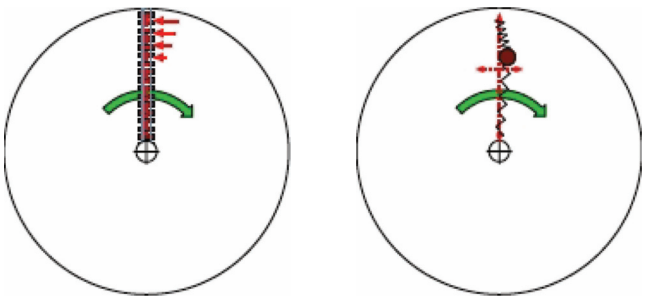


FIGURE 10.34 How to use the Coriolis force in a micro-electromechanical system device.

course on a rotating surface. Motion in a rotating reference frame leads to "sideways" movement. Coriolis showed that if the ordinary Newtonian laws of motion of bodies are to be used in a rotating frame of reference, an inertial force, acting to the right of the direction of body motion for counter-clockwise rotation of the reference frame or to the left for clockwise rotation, must be included in the motion equations (see also Figure 10.33 and Volume I, Chapter 6, "Centrifugal Fluidic Platform—CD Fluidics").

If the path of the moving body on the rotating surface is constrained, the constraining device will feel the Coriolis force (Figure 10.34). By measuring the imparted force (or its effect on an oscillator), we can measure the angular velocity. Almost all MEMS rate sensors use this feature.

The double-ended tuning fork (DETF), shown in Figure 10.35, is one of several structures that is used as the basis of a MEMS rate sensor based on the Coriolis force. One pair of tines of the DETF is driven into primary flexural vibration by a drive-oscillator, thus providing the moving body for the Coriolis force to work upon, as shown in Figure 10.35. The tines of the fork are made to vibrate in opposite directions but in the plane of the fork, which represents the primary flexural vibration mode. The Coriolis force tends to displace the tips of the fork out of this plane, forcing the tips to describe an ellipitical path as the rotation excites a secondary mode of resonance, a torsional mode coupling with the first flexural mode. A quartz DETF from Systron Donner (Concord, CA; http://systron.com) exploits the piezoelectric properties of

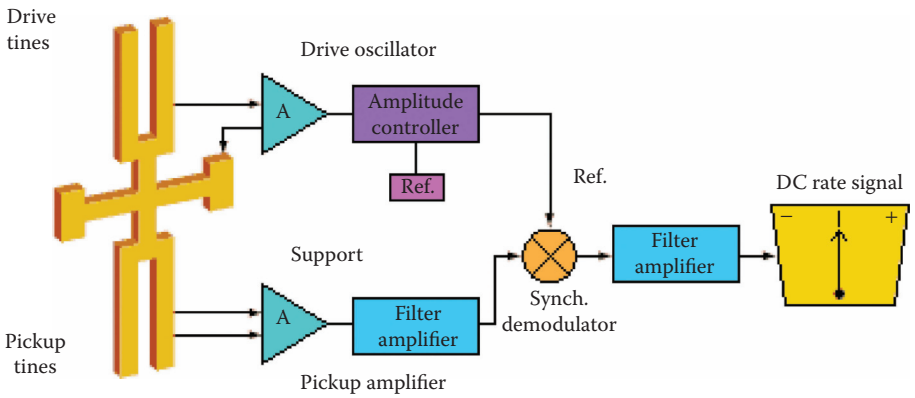


FIGURE 10.35 DETF for a rate sensor. The DETF structure may be etched from quartz or dry etched in the epilayer of a silicon-on-insulator wafer and released by wet etching the oxide layer underneath (from an older version of the http://www.systron.com website).[24]

the material to both excite and sense flexural and torsional vibration modes. The same company made a Si DETF; in this case, the DETF is dry etched out of the episilicon layer of an SOI wafer and under-etched by wet etching. To drive a Si DETF, electrostatic drive electrodes need to be implemented; otherwise, the device works just like the quartz DETF device. Although it gives up some accuracy, the Si device is considerably less expensive and has more than enough accuracy for many applications.

Daimler Benz AG, Stuttgart, Germany, uses a single tuning fork based on SOI micromachining technology for angular rate sensing. The tines of the silicon tuning fork vibrate out of the plane of the die, driven by a thin-film piezoelectric aluminum nitride actuator on top of one of the tines. The Coriolis forces on the tines produce a torque moment about the stem of the tuning fork, giving rise to shear stresses that can be sensed with diffused piezoresistive elements. The shear stress is maximal on the center-line of the stem: this forms the optimal location for the piezoresistive sense elements. The selling price of angular rate gyros such as Bosch's surface-micromachined gyro and Systron Donner's or Matshushita's tuning forks was approximately $25 in 1996, and they were relegated to the top of the line model vehicles.

Other types of angular rate Si sensors, taking advantage of the Coriolis effect, are under development based on comb or ring structures, and both surface and bulk-micromachined sensors have been demonstrated. A comb has the advantage of a larger capacitive area; however, the ring has the advantage of better vibration immunity.

One common comb design, as illustrated in Figure 10.36a, for ADI's ADXRS angular rate sensors, is a proof mass, which can move along the two in-plane

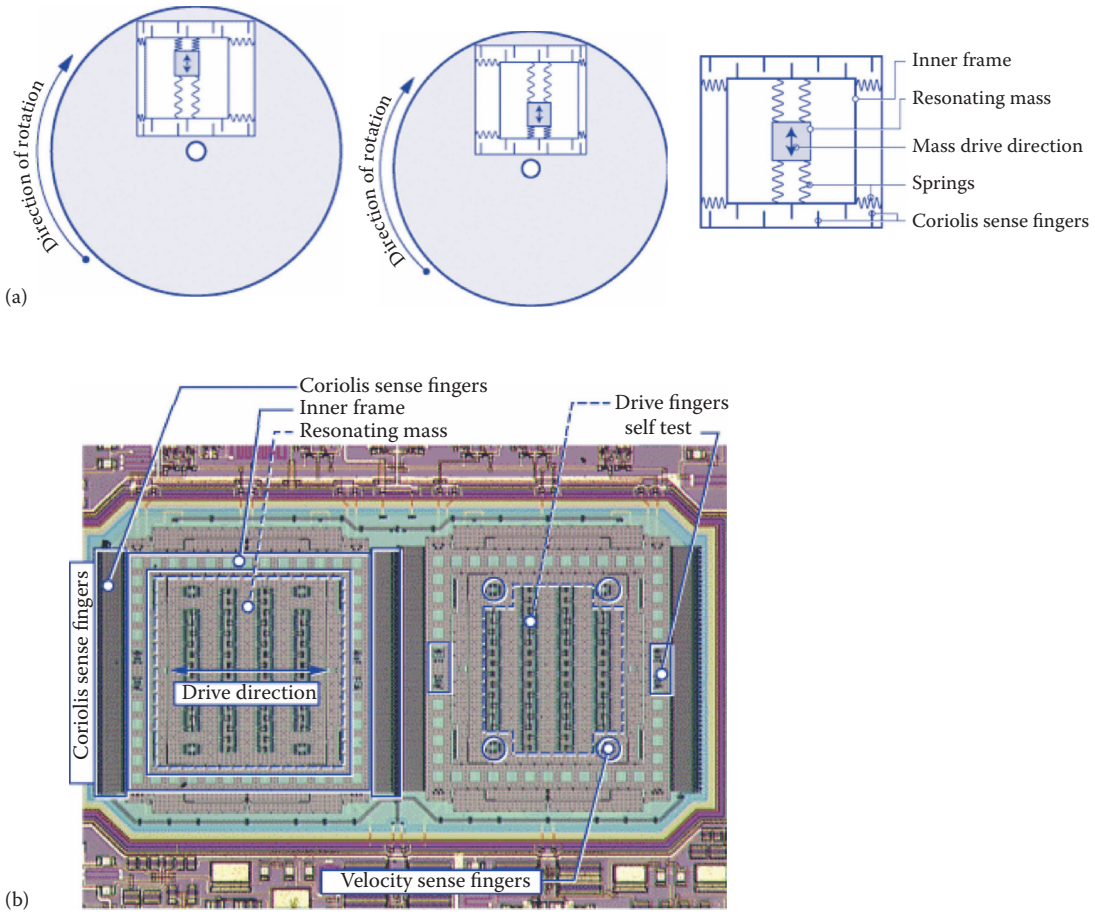


FIGURE 10.36 (a) Design principle behind the ADXRS150 and ADXRS300 from Analog Devices. These yaw-axis gyros have full-scale ranges of 150 and 300°/s, but they can measure rotation about other axes by appropriate mounting orientation. (b) The ADXRS sensors include two structures to enable differential sensing in order to reject environmental shock and vibration. (From 2003. *Analogue Dialogue* 433–37.) New iMEMS Angular-Rate-Sensing Gyroscope. [John Geen (john.geen@analog.com) and David Krakauer (david.krakauer@analog.com), Analog Devices, Micromachined Products Division.]

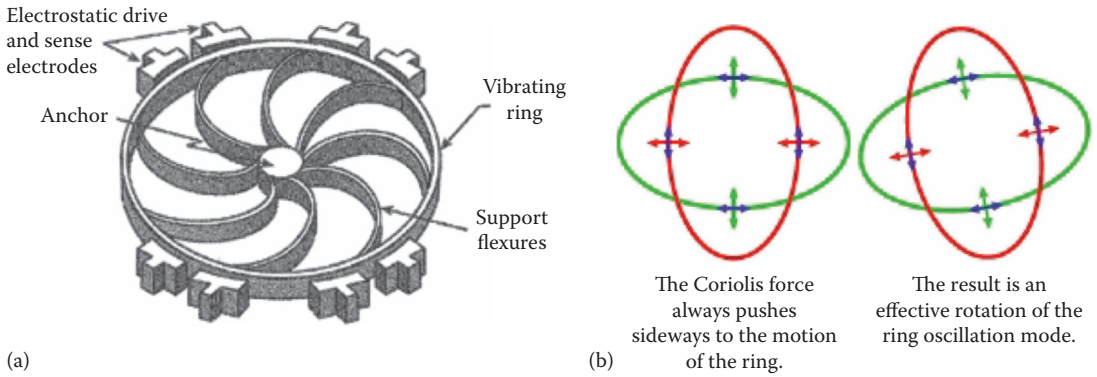


FIGURE 10.37 Operating principle of the rotating ring gyroscope.

axes through a suspension system. The suspended mass is surface-micromachined from polysilicon and tethered to a polysilicon frame so that it can freely resonate only along one of the two in-plane axes. To measure the Coriolis acceleration, the frame containing the resonating mass is in turn tethered to the substrate by springs situated at 90° relative to the resonating motion. With the resonating mass moving, and the surface on which the sensor is mounted rotating, the mass and its frame experience the Coriolis force. ADI's ADXRS sensors measure the displacement of the resonating mass and its frame due to the Coriolis effect through capacitive sensing elements attached to the resonating mass. The comb design in this sensing approach refers to the Coriolis sense fingers that are used to capacitively sense displacement of the frame in response to the force exerted by the mass. These sensing elements are silicon beams interdigitated with two sets of stationary silicon beams attached to the substrate, just as we saw for accelerometers in Figure 10.29b.

The ADXRS gyro electronics can resolve capacitance changes as small as 12×10^{-21} farads (12 zeptofarads) from beam deflections as small as 0.00016 Å (16 femtometers). The only way this can be done in a practical device is by situating the electronics, including amplifiers and filters, on the same die as the mechanical structure. These subatomic displacements are somewhat difficult to comprehend, as the individual atoms on the surface of the moving parts are moving randomly by much more. There are about 10^{12} atoms on the surfaces of the capacitors, so the statistical averaging of their individual motions reduces the uncertainty by a factor of 10^6. So one might wonder whether we can do better by removing the air molecules to prevent their impacting on the sensor. The answer is no. The suspension mass weighs only 4 μg, its flexures are only 1.7 μm wide, and this assembly is suspended over the silicon substrate. It is air that cushions the structure, preventing it from being destroyed by violent shocks and crashes into the substrate. The ADXRS sensors include two identical suspended masses to enable differential sensing in order to reject environmental shock and vibration (see Figure 10.36b).

Delphi[25] and Silicon Sensing Systems (formerly British Aerospace Systems and Equipment), in collaboration with Sumitomo Precision Products Company, are using the vibrating ring gyroscope (VRG) approach to exploit the Coriolis effect.[26] In the vibrating ring approach, a ring is flexured back and forth in resonant mode. The Coriolis effect induces a flexure that is sideways (and out of phase) with the driving flexure (Figure 10.37).

This principle is further detailed in Figure 10.38, which shows a wheel that is driven to vibrate about its axis of symmetry. Rotation about either in-plane axis results in the wheel's tilting, which can be detected with capacitive electrodes under the wheel, as shown in Figure 10.38a. The Delphi-Delco CMOS integrated single-chip surface-micromachined angular rate sensor is shown in Figure 10.38b. The micromachined ring shown here is driven into resonance in the plane of the chip and its vibration forms nodal points. During angular motion, the amplitude of vibration at the defined nodal points increases, and this vibration signal is fed back to reduce the amplitude, thereby maintaining a stationary pattern with nodes at the defined locations. Sacrificial micromachining fabrication techniques are used to allow the ring and

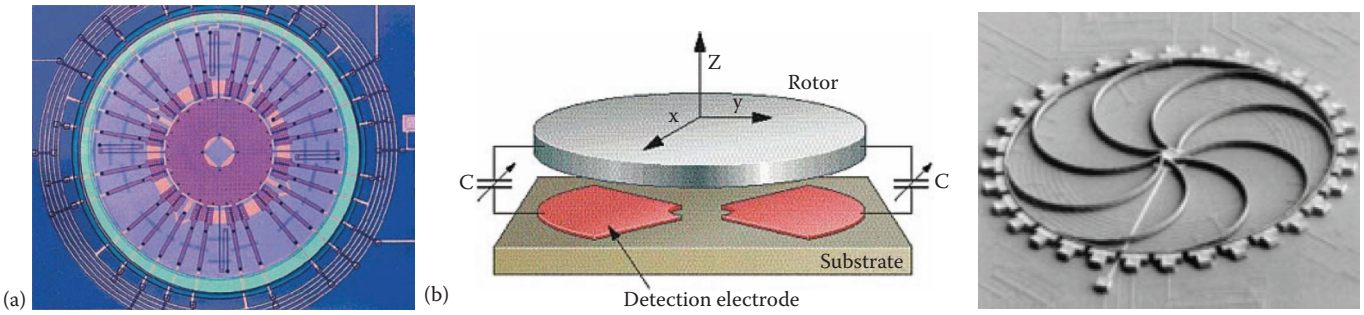


FIGURE 10.38 (a) Operation principle of the vibrating ring gyroscope. This polysilicon surface-micromachined vibrating wheel gyroscope was designed at the Berkeley Sensors and Actuators Center. (b) Delphi-Delco complementary metal-oxide-silicon integrated single-chip surface-micromachined angular rate sensor.

springs to vibrate over the surface of the IC. A scanning electron photomicrograph of the earliest version of the sensor element is shown in Figure 10.38b. The micromachined ring sensor incorporates an electroformed vibrating ring structure fabricated on a silicon IC control chip. Sensors are vacuum-sealed at the chip level using wafer-to-wafer bonding. A vacuum is required since the ring sensor has to be driven into resonance for normal operation with a quality factor >1000. Semicircular springs support the 1-mm-diameter ring and store the vibrational energy. The springs are attached to the substrate with a symmetric post. The post/spring design greatly reduces the effect of packaging stresses on the sensor.

Package-induced stress has been a longstanding problem for many other sensor programs. The Delphi-Delco angular rate sensor is a CMOS integrated single chip and is used for automotive steering assistance, active brake control, and rollover detection.

At the University of California, Irvine, Dr. Shkel's group is developing a new class of micromachined vibratory angular rate sensors, emphasizing structural design concepts to improve robustness, temperature range, and shock survivability. These sensors combine comb and ring mechanical elements. The Shkel group has explored multiple degree-of-freedom (DOF) dynamical systems in order to obtain sensing structures that are inherently insensitive to fluctuations in operation conditions, such as temperature variations, or shock loads.

The Shkel angular rate sensor, shown in Figure 10.39, features a distributed mass architecture utilizing multiple linear oscillators in the drive mode and a single central torsional oscillator in the sense

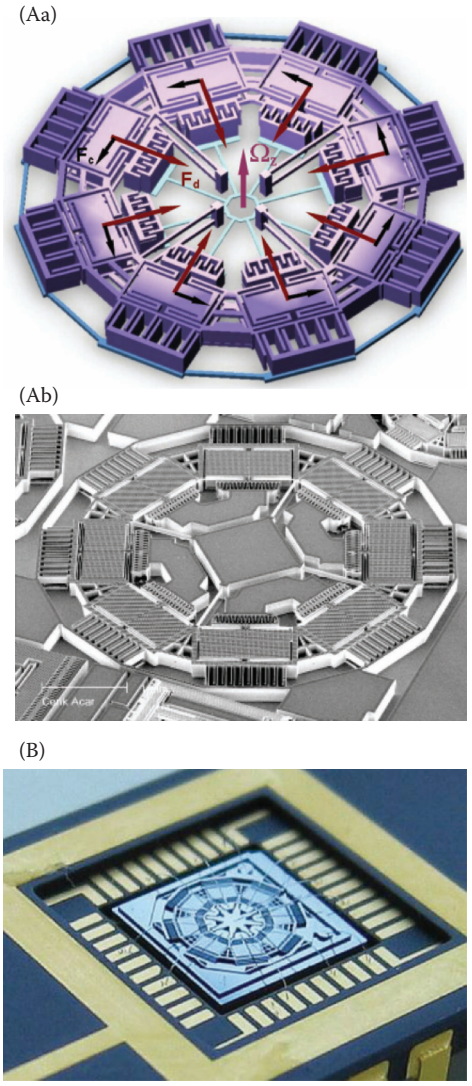


FIGURE 10.39 Microelectromechanical system angular rate sensors developed at University of California, Irvine. (Courtesy of Dr. Andrei Shkel.[27]) (Aa) Illustration of the distributed-mass gyroscope. (Ab) Scanning electron micrograph of a in-house fabricated bulk-micromachined prototype. (B) A prototype is packaged in forty-pin dual in-line package.

mode. This design address a major technical difficulty in the design of MEMS vibratory angular rate sensors, i.e., to keep matched frequencies in the drive and sense directions and eliminate structural cross-axes coupling between the drive and sense modes of the sensor.[27]

The distributed drive-mode oscillators are driven in-phase toward the geometric center of the structure and are constrained in the tangential direction with respect to the supporting frame. In the presence of an angular rotation rate about the *z*-axis, the Coriolis forces are induced on each proof mass orthogonal to each drive-mode oscillation direction. Thus, each of the induced Coriolis force vectors lies in the tangential direction, and they combine to generate a resultant torque on the supporting frame. The net Coriolis torque excites the supporting frame into torsional oscillations about the *z*-axis, which are detected by sensing capacitors.

The multidirectional and axisymmetric nature of the drive-mode oscillators has several benefits over a conventional angular rate sensor design. Firstly, since the drive forces applied to the drive-mode oscillators cancel out in all directions due to the radial symmetry, the net driving force on the structure reduces to zero and eliminates quadrature error in the presence of structural imperfections. Secondly, instability and drift due to mechanical coupling between the drive and sense modes are minimized, since the structure allows complete decoupling of multidirectional linear drive mode and the rotational sense mode. The symmetry of the drive-mode oscillator structure about several axes also cancels the effects of directional residual stresses. The most prominent advantage of the design concept is the capability to provide a wide-bandwidth operation region in the drive-mode frequency response. By designing each drive-mode oscillator to have incrementally spaced resonance frequencies, a constant total Coriolis torque is achieved over a wide range of driving frequency. If the sense-mode resonance frequency is designed to match the center frequency of the drive-frequency range, robustness and reduced sensitivity to structural and thermal parameter fluctuations are achieved. Consequently, the presented design concept results in improved robustness and stability over the operating time of the device, and it is expected to simplify the control architecture and relax tight fabrication and packaging constraints.

Conclusions

The automotive sector has long been a growth market for MEMS sensors with applications in airbag release, vehicle dynamics control, ABSs, antitheft systems, active suspension control, electronic stability programs (ESPs), tire pressure monitoring sytems (TPMSs), headlight leveling, etc. However, the industry continues to change and face new opportunities and challenges—regulations, saturation in some applications, and price erosion issues, to name a few. Today's high-end vehicles feature up to 100 different sensors. About thirty of these are now MEMS based. The market is made up of accelerometers, gyroscopes, and inclinometers, as well as pressure and flow sensors. Emerging applications include infrared (IR) sensors for air quality, microscanners for displays, and, further out, MEMS oscillators and energy scavengers for TPMS.

Consumer electronics applications are growing very rapidly and likely will overtake automotive MEMS applications. There is a convergence of requirements between the automotive and consumer electronics sensors. The automotive market can potentially derive great benefit from the advances in consumer electronics MEMS markets, but automotive accelerometer performance requirements are vastly more stringent than those of consumer electronics.

Although the electronics content of vehicles continually increases, the vehicle market itself is expanding more slowly than in previous years. The net effect is a steady increase in the value of the automotive electronics market. The use of IC-compatible process lines for MEMS production is growing. There are clearly high benefits in using CMOS or BiCMOS-compatible manufacturing lines to target high-volume MEMS markets. Some specific reasons to select monolithic integration as a solution are its cost advantages, commercial imperatives, performance advantages, and pragmatism. Nevertheless, hybrid solutions are still sometimes preferred, as they offer distinct advantages (more flexible manufacturing strategy, shorter development time, a simpler and more easily available manufacturing process, etc.). The integrated MEMS market share is expected to

grow in the future. However, there are still bottlenecks to be overcome (process standardization, MEMS is not always a large volume market, etc.).

MEMS in Medical and Biotechnology Markets: Bio-MEMS

Introduction

The per capita annual expenditure for health care is about $3000 in the United States, $3200 in Germany, and $2000 on average in Europe. The total world health market in 2004 was about $793 billion, and large segments include $280 billion for medical instruments and $484 billion for pharmaceuticals. In vitro diagnostics, surprisingly, represented only 3% of the total world health market (i.e., $27.6 billion) (see Figure 10.40; Adams Business Associates).

The biotechnology market covers a large area and has vast and crucial branches among them. Some of these biotechnology branches are genetically modified food (GMO), antibody technologies, antisense technology, biomaterials, biopharmaceuticals, drug discovery, DNA arrays, environmental/industrial biotechnology, enzymes, gene therapy, genetic engineering, genomics, informatics, instrumentation and equipment, molecular biology, pharmaceuticals, proteomics, tissue engineering, and wound care. The biotechnology market had strong growth in recent years, generating a total revenue of $126.3 billion in 2005. It represented a compound annual growth rate (CAGR) of 12.8% for the period 2001–2005. Currently, the United States dominates the market and accounts for 54.3% of the global market revenue.

Medical and biotechnology markets show strong promise for MEMS solutions in a wide variety of applications, as illustrated in Table 10.8. The medical/biotechnology application field of MEMS is referred to as bio-MEMS. In bio-MEMS, one deals with life-saving disposable diagnostic sensors, DNA arrays systems speeding up drug discovery, smart pills improving drug delivery, μ-TAS and LOC, etc., and most of those devices come with disposables. Therefore, the sales volume in bio-MEMS is expected to be eventually significantly larger than that for mechanical MEMS, where sensors are typically in use anywhere from three to ten years.

Progress in the medical and biotechnology area today is only to a minor extent derived from silicon-based bio-MEMS. It has become clear that the straight application of Si technology is very often inadequate in this field compared with its successful application in the automotive world or in mechanical and optical applications in general. Whereas in the automotive application the market for Si micromachined products is mature enough to warrant projection of realistic future revenue numbers, predicted revenues for micromachined sensors and instrumentation for biomedical and biotechnology applications mostly refer to speculation as commercial successes still remain short of realization. In the bio-MEMS field, to succeed as well as mechanical and MOEMS did, one will need to develop mass manufacturing techniques based on very large batches or roll-to-roll manufacturing on paper or plastic substrates. Through a merging of traditional micromanufacturing methods (e.g., electroplating and micromolding) with less traditional methods (e.g., lithography) and continuous manufacturing methods (R2R), we expect more bio-MEMS market breakthroughs.

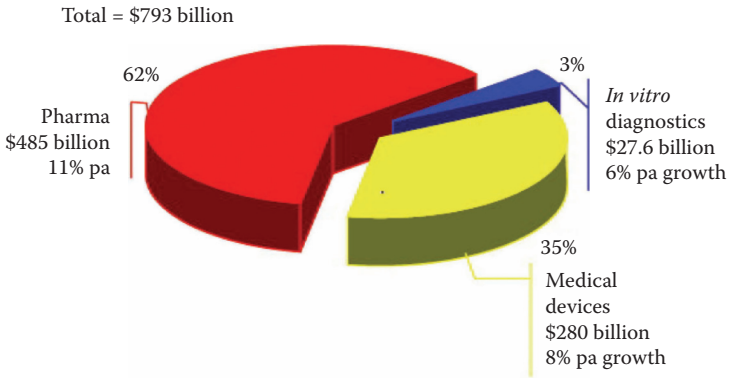


FIGURE 10.40 Global health care market in 2004. (From Adams Business Associates, aba@a-b-a.co.uk.)

Table 10.8 Medical/Biotechnology MEMS Applications

Drug Delivery Systems	Typical MEMS Example	Cardiology	Typical MEMS Example
Patches	Cygnus GlucoWatch Biographer (Figure 1.13 this volume)	Pacemakers	One- or two-axis accelerometers for activity monitoring
External and implantable pumps	Animas-Debiotech belt-worn drug delivery pump (Figures 1.10 and 1.11 this volume).	Angioplasty catheters	TRW NovaSensor, Motorola, and Siemens offer disposable pressure sensor (<$1)
Smart pill	Smart pill (Volume II, Figures 4.99 and 4.100)		
Monitoring		**Biotechnologies**	
Point of care testing	i-STAT handheld blood analyzer (Figure 10.42 this volume)	Polymerase chain reaction (PCR)	Polymerase chain reaction (PCR) efforts e.g., at Cepheid
On-line monitoring of blood gases	Dual lumen-based pH, CO_2, and O_2 sensor array (Volume II, Figures 4.92 to 4.96)	Genetic tests and therapy	Nanogen's NanoChip (Volume II, Figure 8.51)
On-line monitoring of pressure	TRW NovaSensor, Motorola, and Siemens offer disposable pressure sensors (<$1)	Electrophoresis	Caliper and Aclara electrophoretic chips
		Liquid handling systems	RotaPrep's CD sample preparation station (Figure 10.51 this volume)
Prostheses/artificial organs		**Minimally invasive surgery**	
Orthopedics, ophthalmology	Nanotexturing of surfaces for enhanced biocompatibility of implants (Chapter 5, this volume)	Cutting tools	Painless ultrasonic cutting tools
Neurology	MEMS neural stimulator probes		

After a write-up on the market character in this diverse field, we look at bio-MEMS applications in in vitro and in vivo diagnostics while highlighting bio-MEMS in biomedical microinstrumentation [micrototal analysis system (μ-TAS) or lab-on-a-chip (LOC)]. Micromachining efforts for building micro-instrumentation are seen today as a better approach to solving chemistry and biology problems compared with building more selective chemical sensors. Some underlying reasons are that the resulting instrument may justify significantly higher costs than disposable chemical sensors and help solve the shortcomings of chemical/biological sensors by enabling calibration, separating of bound and unbound species, cleaning the detectors, filtering out unwanted compounds, and obviating the need for very selective chemical sensors.

We conclude with a summary on progress in μ-TAS and LOC in sample preparation for molecular diagnostics, as it is felt that this is the area with the biggest need for bio-MEMS innovation.

Market Character

In terms of the business climate and funding sources in bio-MEMS, the playing ground is uneven and depends very much on the specific medical or biotechnology niche addressed. One overall trend is a move away from the initial application of μ-TAS and LOC to high-throughput screening (HTS) to its current application to point of care (POC) diagnostics instead. Even in the technology boom years (before 2001), it used to be difficult to raise money for in vitro diagnostic bio-MEMS work; for novel in vitro glucose sensors it was especially difficult, and for in vivo diagnostic devices it was harder yet. This was not only because the diagnostics industry is very entrenched and conservative but also because investors had already invested large amounts in that area with little return and because of the uncertainty of health care reimbursement. Moreover, disposable diagnostics were not necessarily perceived positively because of the waste disposal issues. Finally, one had

to demonstrate a new diagnostic sensor/instrument 80% toward a finished product in order to even be able to start talking to skeptical investors or conservative large diagnostic companies.

In the case of drug discovery, investors were, up to 2001–2002, more willing to take major risks at an early stage of the technology development. Significantly, from Figure 10.40, medical devices and pharma are a much larger market than in vitro diagnostics. Consequently most microfluidic start-up companies working on μ-TAS or LOC concepts were targeting drug discovery rather than the diagnostic area for early market penetration with bio-MEMS products. The 2001 downturn in the U.S. securities market greatly reduced the value of all of these companies, who must now make money the traditional way—by making products that someone wants to buy.

Today, the situation is reversed with diagnostic applications more attractive than research tools for biotechnology. Diagnostic products are indeed more and more appreciated as important tools in reducing costs of health care. In the past, diagnostic tests were merely used as support data for the real diagnosis, often to confirm the physical diagnosis of the doctor or for legal protection that the correct procedures had been carried out. A major change over the last 5–7 years has been the increased use in true diagnosis and differentiation of conditions in order to achieve the best treatment, i.e., theranostics bringing together the diagnostic and pharmaceutical industries. The driving force for this shift in role for diagnostic tests is the large costs for health care throughout the world at a time when the economies are less robust. The attraction of investing in the development of diagnostic products is being boosted by this change in importance and the recognition that the overall risks are reduced together with an earlier return on modest investments (Adams Business Associates, aba@a-b-a.co.uk).

With this improved business climate, improvements in bio-MEMS fluidics technologies may find now a more sizable market in better diagnostic systems, such as two-point sensor calibration schemes, sensor cleaning devices, sample preparation strategies (see further below), etc., than in HTS, where microfluidics are often only a minuscule part of a large robotic system. Diagnostics is a mass market; HTS is at most a research market (see above in the section Preamble on the MEMS market). We believe that after more significant progress is made in microfluidic platforms, there will be a resurgence of interest in chemical and biological sensors and methods to manufacture them inexpensively for diagnostic applications.

Perhaps this will produce more i-STAT (from Abbott) type products. Abbott's i-STAT, with its planar electrochemical sensor array for in vitro blood electrolyte and blood gas analysis (see below, Figure 10.42), is one of the few survivors of the many companies in the area of miniaturized diagnostics that were started in the 1980s.

The two most promising bio-MEMS areas in the short term are both in molecular diagnostics, and they involve sample preparation for DNA or RNA extraction from a wide range of samples with the same instrument (see Sample Preparation section, below) and the integration of DNA arrays with microfluidics to speed up the hybridization reaction.

In Vitro Diagnostics*

Introduction

Biomedical sensors and instrumentation for in vitro diagnostics represent a huge opportunity for micromachining in the broad sense of the word, through traditional and nontraditional miniaturization methods, continuous manufacturing methods and combinations thereof. To date, the majority of routine clinical tests are performed in large independent clinical laboratories with advanced, automated, analytical equipment. As the amount of resources spent on health care continues to soar, there is an increasing pressure for the decentralization of the POC and toward easy-to-use, miniaturized, inexpensive diagnostic machines and devices. It is in this shift that miniaturization science has so much to offer. With microinstrumentation (μ-TAS, LOC), one envisions that analyses will be performed significantly faster, without the need of

* The diagnostics market (also called *in vitro diagnostics*) includes the instruments and reagents used for the screening, diagnosis, monitoring, or prognosis of disease by laboratory techniques. The products are used in hospitals and private laboratories, in physician offices, and at other point-of-care sites (emergency rooms, intensive care units, outpatient clinics, nursing homes), as well as in home testing (primarily diabetic glucose tests and pregnancy tests).

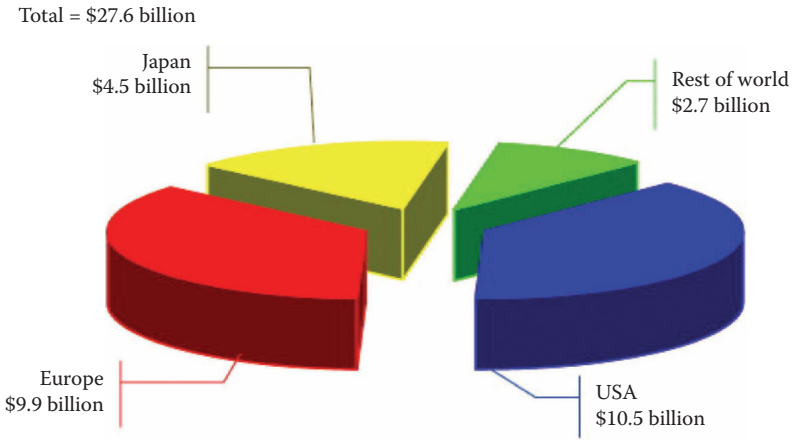


FIGURE 10.41 In-vitro diagnostic markets in 2004. (From Adams Business Associates, aba@a-b-a.co.uk.)

highly trained technicians (e.g., in remote villages, small health stations, etc.), only minute samples will be required, and dead volume in connections will be eliminated. Current in vitro diagnostic tests are categorized as clinical chemistry, blood gas/electrolytes, hematology/flow cytometry, coagulation, DNA tests (molecular diagnostics), microbiology, diabetes, urinalysis, histology/cytology. As shown in Figure 10.41, the worldwide in vitro diagnostics market was about $27.6 billion in sales in 2004, and the annual growth rate was 6% (see also Figure 10.40, which shows that in vitro diagnostics was only 3% of the global health care market for 2004).

The main areas of above-average growth (>6%) in in vitro diagnostics, identified in the 2003 report from the European Diagnostic Manufacturers Association (EDMA; see http://www.edma-ivd.be), are shown in Table 10.9.

As examples of MEMS contributions to in vitro diagnostics, we detail electrochemical sensors for blood electrolyte and blood gas analysis, glucose sensors and DNA arrays. We also cover MEMS contributions to the fabrication of miniaturized well plates, flow cytometry, and surface plasmon resonance (SPR) sensors.

TABLE 10.9 In Vitro Diagnostics with Above-Average Growth (European Diagnostic Manufacturers Association, http://www.edma-ivd.be)

Glucose tests	>10%
Cardiac tests	>20%
Nucleic acid-based tests	>20%
Genetic testing	>>20%, but from a small base

Other in vitro diagnostic applications are reviewed throughout the book. For example, contributions to the future of immunosensing were reviewed in Volume I, Chapter 7. DNA and protein arrays are also covered in Volume II, Chapter 8 (their fabrication and typical use), and the Nanogen electronic DNA array is covered in Example 8.2 of that chapter. Generic challenges concerning miniaturization in biomedical equipment, such as correct partitioning issues, were addressed in Chapter One of this volume.

Electrochemical Sensors for Blood Electrolytes and Blood Gases

Background: MEMS in Tabletop and Handheld In Vitro Diagnostics

During the past 25 years, the commercialization of MEMS for in vitro diagnostics failed to keep pace with earlier projections. We believe that an overemphasis on integration of chemical sensors with electronic functions on Si hampered the development. We suggest speeding up the progress by applying micromachining tools more correctly. This includes nonsilicon manufacturing, hybrid manufacturing, and roll-to-roll manufacturing, as covered in Chapter 1, this volume.

Progress in sensors for tabletop diagnostic instruments based on nonsilicon manufacturing techniques has been significant. The fully automated Kodak Ektachem is a tabletop clinical blood analyzer based on planar electrochemical sensor slides, and it has been on the market for some time. The Ektachem, the Seralyzer from the Ames Division of Miles Laboratory, the Refletron Plus system from

Roche Diagnostics/Boehringer Mannheim (with a sixteen-test menu), the Analyst from E. I. du Pont de Nemours, the Nucleus from Nova Biomedical, and other, similar instruments all cover a wide range of blood chemistries.[28,29] However, none of these products involve Si micromachining bio-MEMS technology. The progress enabling their development mainly was based on the development of solid-state stabilized reagents for so-called *dry electrochemistry*,[30] on miniaturized electrochemical sensors, and to a minor extent on advances in thin-film technology. Despite the many earlier claims, contributions from micromachining (Si-based or not) are mostly at the component rather than the system level, and many machining technologies are in the experimental stage. Systematic analysis of different machining options as advocated in this book (see Chapter 1, this volume) is only now starting to be applied. In the case of diagnostics, this leaves little room for Si-based technology. As a result of these analyses, one comes to the conclusion that bio-MEMS clearly counts as a fragmented set of supporting technologies rather than a monolithic new industry.

It was anticipated that Si micromachining would play a more prominent role in the development of sensors for the smaller, next-generation, handheld, and portable diagnostic instruments. In reality, micromachining and Si sensor content in the more advanced, handheld and portable, biomedical products is also surprisingly limited. Some such products for analysis of blood electrolytes and gases on the market include the GEM-Premier from Mallinckrodt Sensor Systems, the similar STATPal from PPG, and the Portable Clinical Analyzer from Abbott (the i-STAT system), providing data on sodium, chloride, potassium, urea nitrogen, glucose, and hematocrit in less than 2 min. Another handheld instrument of interest is the AccuMeter cholesterol test from Chemtrak, designed to screen lipids to evaluate cardiovascular disease risk, and the portable blood analysis systems by Diametrics Medical Inc. (electrochemical) and AVL (optical fluorescence). Abaxis introduced the Piccolo™ for general blood chemistry analysis, a small tabletop centrifugal instrument with a spinning reagent disk and optical read-out. Biologix Inc. developed the electrochemistry-based handheld pHacs STAT for blood pH, potassium, and oxygen.[28,31] Of all these products, only the i-STAT sensors are Si based. The Si avenue toward tabletop and handheld instruments has been pursued for more than thirty years with little success to report. Innovation in biomedical sensors for in vitro diagnostics has come mainly from the invention of new chemical detection schemes rather than from clever new micromachining methods. Actually, in biomedical sensors, a reversal toward hybrid and continuous manufacturing methods took place rather than a move toward more integration in Si embodiments.

Blood Electrolyte and Blood Gas Sensors

For blood electrolyte (pH, Na^+, K^+, Ca^{2+}, etc.) and blood gas (O_2 and CO_2) analysis with tabletop or handheld instrumentation, one wants to develop inexpensive, multiuse, or disposable chemical sensors. The principal approach so far has been via electrochemical technology. Microfabrication of electrochemical sensors, using IC technology, has been extremely challenging, mainly due to process incompatibility issues, packaging problems, failure to incorporate a true reference electrode, and the difficulties involving patterning relatively thick organic layers such as ion-selective membranes and hydrogels (see also Volume I, Chapter 7). Most researchers have come to the conclusion that for chemical sensing, especially for in vitro applications, a modular, hybrid approach rather than an integrated Si approach marks the road to progress. Of all the handheld biomedical analyzers on the market, only the i-STAT Portable Clinical Analyzer (Abbott, see http://www.abbottpointofcare.com/) involves Si in the sensor construction itself, primarily because of a historical production mode. Single-crystal Si, it was thought at that time, was the way to manufacture biosensors, and after the equipment investments were made, it was hard to reconsider. In Figure 10.42, we highlight the i-STAT instrument, together with a disposable sensor cartridge.

The Portable Clinical Analyzer from Abbott uses Si in the disposable chemical sensor cartridge only as a substrate and a contacting base; electronics are kept in the handheld reader, clearly a more economical approach than putting the electronics on the disposable (Figure 10.42). The sensor electrodes, gels, and membranes are all deposited with semiconductor-

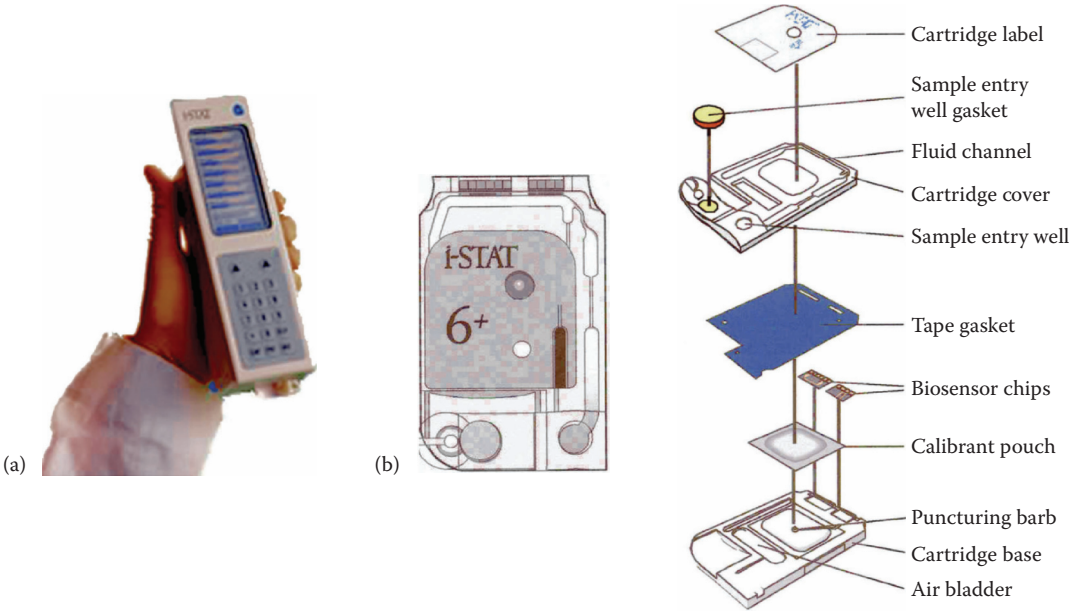


FIGURE 10.42 (a) The i-STAT portable clinical analyzer (now Abbott, see http://www.abbottpointofcare.com/) is used in conjunction with disposable cartridges for determination of a variety of parameters in whole blood. The analyzer stores up to fifty patient records and permits on-screen viewing of test results, as well as transmission of records to a data management system using infrared signals. (b) An i-STAT disposable cartridge.

type processes.[32] Without electronics onboard, the continued use of Si as a support has mostly a historical explanation. It should also be noted that the fluidic channels in the disposable i-STAT cartridge are made from molded plastic and not with Si micromachining. Abbott's manufacturing methods of the Si sensor chip in the i-STAT, after twenty years, still make it difficult to be profitable. This may be related to the manufacturing approach being silicon based and the design being integrated rather than modular. An alternative approach for the manufacture of a microfabricated, modular chemical sensor arrays is introduced in Chapter 1, Example 1.1. Epocal in Toronto uses a very different approach to making biosensor arrays compared to the methods used in manufacturing the i-STAT sensors (http://www.epocal.com). Epocal sensors are fashioned on flex circuits in a single continuous in-line manufacturing process. Flex circuits are adapted from industry standard smart card modules fashioned on 35 mm tape-on-reel format. The Epocal approach is very much the type of progress in biosensor manufacture we promote in Example 1.1.

While the number of new tools available for microfabrication, as gleaned from Chapter 1 of this volume, has grown dramatically, few methods have been gainfully applied to biosensor construction for clinical applications (some of these few methods have been reviewed in Volume II, Chapter 8). In contrast to mechanical sensors, where the excellent mechanical properties of single crystalline Si often tend to favor Si technology, the choice of the optimum manufacturing technology for chemical sensors is far less evident. For chemical sensors, the Si in the sensor often serves no other role than as a passive substrate. Consequently, the following trend emerges: whereas mechanical sensors (pressure, acceleration, temperature, etc.) are moving toward more integration embodied in CMOS-compatible surface micromachining, chemical sensors and biosensors are moving away from integration on Si and toward hybrid technology on plastic or ceramic substrates with silk-screening or drop delivery systems for the application of the organic layers.

Glucose Sensors

In vitro glucose sensing represents the single largest biosensor market, and a significant amount of all research and development in biosensors is geared toward improving this technology. Glucose self-testing by diabetics is one of the fastest growing segments of the world diagnostic market (see Table 10.9).

Diabetes mellitus is a chronic metabolic disorder that is caused by a failure of the body to produce insulin and/or an inability of the body to respond adequately to insulin. Type 1 diabetes occurs most often in children or young adults and accounts for 5–10% of the diagnosed diabetes patient population. Type 2 diabetes, or adult-onset diabetes, accounts for 90–95% of diagnosed diabetes cases worldwide, typically developing in middle-aged adults. Diabetes effects approximately 170 million people worldwide and is increasing, with the World Health Organization (WHO) predicting 300 million diabetics by 2025. The United States alone has 20.8 million people suffering from diabetes. This equates to approximately 6% of the population. Diabetes is the sixth leading cause of death in the United States (based on number of death certificates; National Institute of Diabetes and Digestive and Kidney Diseases). The global diabetes market was worth $18.6 billion in 2005, which was an 11.5% increase from 2004 sales of $16.6 billion. The United States has the dominant share in the global diabetes market, with 49.6% of 2005 global sales. The large U.S. market share is mainly due to the large population and the high prevalence of diabetes, the increasing incidence of diabetes associated with Western diets and an aging population. Kelly Close of Close Concerns Inc. in San Francisco, a consultancy devoted to diabetes and obesity research, estimates that on the basis of industry estimates worldwide sales of blood glucose meters were almost $5.9 billion in 2004. Roche Diagnostics leads, with sales of about $2.3 billion. LifeScan/J&J is second, with $1.7 billion. Together they control two-thirds of the blood glucose marketplace. Bayer, with $800 million, and Abbott Diabetes Care, with $790 million, are neck and neck for third place. The worldwide market for blood glucose test strips and monitors is currently almost $3 billion, and the market is growing rapidly at a sustained rate of 13% a year (for some example blood glucose monitors see Chapter 1, this volume, Figure 1.7). The most prominent method currently available for collecting blood glucose remains the painful and messy finger prick procedure. The recommended sampling rate is four times a day, further discouraging compliance. The need for a painless testing glucose sensor (minimally invasive) is a challenge that many MEMS engineers have worked on. A new method that permits painless blood testing while providing results that meet current standards of accuracy and reliability is key to success (for two examples of blood/interstitial fluid collections devices, see Figure 1.15a with a lancet for blood collections device, and Figure 1.15b for interstitial fluid collection using a laser puncture).

Three generations of glucose sensor technologies are reviewed in Volume I, Chapter 7. A breakthrough came with the 1987 introduction of the amperometric glucose sensor by Medisense (now owned by Abbott), which proved that small, planar electrochemical probes can compete with colorimetric paper strips. The Medisense/Abbott glucose sensor, the Precision QID (shown in Volume I, Figure 7.80), is based on an innovative mediator chemistry and a traditional thick-film manufacturing process. The sales of mediated amperometric biosensors now represents a dominant 65% of the world biosensor market.[33] A comparison of manufacturing techniques for these types of sensors reveals the difficulty of competing using a Si-based approach. One often forgets that in biosensors such as glucose sensors, one is not competing with an expensive serial manufacturing process but with continuous, printing-type processes (see Chapter 1, this volume), making products even cheaper than in a Si type batch production. The cost target for a disposable glucose sensor is about $0.10, about a factor of 10 less than an unpackaged physical sensor die. Unless a significant performance advantage arises, Si technology cannot compete. It is our belief that, besides new chemistries, continuous processes to manufacture disposable electrochemical sensors will be the ultimate answer to the pressing need for disposable, inexpensive diagnostics (see Chapter 1, Figure 1.17, where we used the manufacture of a disposable glucose sensor to illustrate our streamlined decision-making process). A way to go about continuous biosensor manufacture is demonstrated in Chapter 1, Example 1.1: "Proposed Scenario for Continuous Manufacture of Polymer/Metal-Based Biosensors." Besides Abbott's ExacTech blood glucose monitoring system, there are similar products by Johnson & Johnson (Lifescan) and Wampole Laboratories (Answer).[28]

The future of the glucose sensing business is quite intertwined with new sampling methods. Research is emphasizing minimally invasive, noninvasive, and continuous monitoring of blood glucose levels. MEMS microneedles are under development for minimally invasive sampling by several research groups in academia. One example of a minimally invasive method, recently taken off the market, is the GlucoWatch biographer from Cygnus (see Chapter 1, this volume, Figure 1.13). In this device, glucose is extracted through the skin by electro-osmosis, and the glucose is measured amperometrically.[34] Kumetrix (http://www.kumetrix.com) replaces the typical lancet and blood-testing strip with a battery-powered, handheld diagnostic instrument using disposable MEMS cartridges, consisting of a microneedle with a cross-section of about 100 µm, a silicon cuvette (a chamber for the blood sample), and a small window. The device uses capillary action to painlessly draw approximately 100 nL of blood into the cuvette, where it is mixed with a chemical reagent. A laser then measures the glucose and the blood's glucose level is displayed. The MiniMed continuous glucose monitoring system (CGMS; MiniMed Inc., Sylmar, Calif.) is the first commercially available continuous glucose monitor. Each sterile, disposable sensor (again an electrochemical type) used with the CGMS can provide a continuous glucose profile for up to 72 hours. Glucose concentration is measured in the interstitial fluid every 10 seconds, and an average glucose value is stored in memory every 5 minutes. A communications device enables data stored in the monitor to be downloaded and reviewed on a personal computer. A performance evaluation of this very promising technology is reviewed by Gross et al.[35]

The glucose sensor field became relatively stale after the introduction of the MediSense technology. A promising new research direction is touched upon in the section on in vivo diagnostics below and involves glucose binding protein for the detection. Roe and Smoller wrote a very comprehensive review on the state of the art of bloodless glucose measurements as of 1998,[36] and a review of noninvasive glucose monitoring was presented by Klonoff in 1997.[37] Heller and Feldman recently reviewed in vitro electrochemical sensors.[38]

The next generation of electrochemical sensors will be planar, thick-film, hybrid devices, which will have the cost and fabrication advantages of solid state technology but will likely still have some of the same limitations as present electrochemical sensors, that is, repeated calibrations to compensate for long-term drift and the need for periodic maintenance. An important lesson in reviewing this micromachining application is that there is as much need for further optimization of deposition of chemical sensitive layers as for exploration of new manufacturing options. Investigating continuous printing processes for electrochemical sensor manufacture will be most rewarding in the future (see Chapter 1, this volume, Example 1.1). In the meantime, a modular approach with thick-film deposition processes is advisable. Some of the calibration and cleaning problems will be resolved by implementing more sophisticated plastic molded microfluidics.

DNA Arrays

Molecular diagnostics is another one of the fast growing in vitro diagnostic applications (see Table 10.9). With microarrays or DNA chips, thousands of genes or representative oligonucleotide fragments can be arrayed on a single microchip and hybridization with complimentary strands can be sensitively detected (see Volume II, Figures 8.16 and 8.17). The fabrication methods and uses of DNA arrays are reviewed in Volume II, Chapter 8. The two most significant applications for DNA chips are believed to be in gene expression profiling and single-nucleotide polymorphisms (SNPs). In expression profiling, the chip is used to examine the amounts of messenger RNA, which controls how different genes are turned on or off to produce certain types of proteins and cells. If the gene is expressed in one way, it may result in a normal muscle cell, and if expressed in another way it may result is a tumor (see Volume II, Figure 8.23). By profiling the minute variations in a person's DNA called *single-nucleotide polymorphisms* (SNPs) and correlating these to a person's ability to metabolize a drug, personalized drug therapy might become a reality.

When biotechnology started flirting with DNA arrays and microfluidics-based companies, the MEMS market character changed overnight. The opening

salvo was perhaps given on March 2, 1995, when Glaxo Wellcome (now GlaxoSmithKline) acquired Affymax, a DNA array company, including its ownership interest in Affymetrix, for $650 million, before there was even a product on the market. This was uncommon for the "traditional" MEMS companies who had always needed to show product revenue and even then got rewarded with only modest initial public offerings (IPOs). Microfluidic companies that were early players in the DNA market fray include Aclara Biosciences, Affymetrix, Nanogen, Caliper, Cepheid, and Orchid. There is some overlap between microfluidics companies and DNA array companies, and there will be much more in the future. The author expects great merit in merging DNA arrays with microfluidics, as it will enable sample preparation for diagnostic DNA arrays and enable working from cruder samples for high-throughput screening or gene expression. All the listed companies did go public and enjoyed large capitalizations due to the explosive stock market and generous evaluations associated with biotechnology companies in general. The "technology bubble" of 2001 took down much of these gains, but it is expected that these array companies will recuperate. The wide acceptance in the United States of genetic testing is not about "if" but about "when," and it will be a burgeoning bio-MEMS application. It is not clear today, though, what is the right format to present molecular diagnostic panels to the market; should it be in small panels and handheld readers for the point-of-care (POC) market with a variety of individual cartridges/test panels (say a chip to screen for population linked genetic diseases, a separate heart disease predeposition chip, etc.), or should it be a high-throughput robotized clinical diagnostics laboratory with hundreds of test panels?

Until large market molecular diagnostic tests, preferably with an instrument including sample preparation, become available, the importance of DNA microarrays in research for disease diagnosis, understanding drug interactions and determining drug candidates is expected to keep the DNA array market growing more slowly. Molecular diagnostics today is less dominated by giant multinationals than traditional diagnostics, and tests are performed in many different specialty labs and are still quite expensive (from a few hundred dollars to several thousands of dollars), so there seems to be a more positive outlook for potential new players.

In 2005, revenue for the total U.S. DNA microarray market totaled approximately $446.8 million, and it is expected to rise to $532.1 million by the end of 2012 (Frost & Sullivan). The compound annual growth rate (CAGR) for 2005 to 2012 is estimated at 10.9%.

Protein chips are also expected to emerge, driven by advances in proteomics. Yet despite their apparent similarity to DNA arrays, protein arrays represent a much smaller market, precisely because they are as finicky as the protein elements spotted/patterned on their surfaces. The biological activity of the capture protein elements, whether antibodies or proteins, must be maintained. One has to be able to print them while preserving their activity, and for a commercial product, the arrays have to be stable for years. As a result, there are relatively few protein arrays on the market. Those that are available come in two basic flavors: general protein arrays and antibody arrays. Protein arrays from Invitrogen, ProtoArray microarrays, are finding utility in various applications, including discovery of autoantibody biomarkers, identification of novel kinase and ubiquitin ligase substrates, and elucidation of protein interactions. Antibody arrays serve a different purpose. Available from such suppliers as Clontech, Thermo Scientific, and Sigma-Aldrich, these products enable protein profiling, examining how levels of several hundred proteins change in response to some treatment. Thermo Scientific's LabVision antibody microarray includes 720 antibodies, and Clontech's Ab Microarray includes 512. Typically, protein samples are labeled with Cy3 and Cy5 dye, and this allows the arrays to be visualized using most standard DNA array scanners.

Miniaturized Well Plates

Almost any new fluidic platform faces the daunting challenge of how to either incorporate or displace the current standard method of dealing with small amounts of chemicals, i.e., the use of miniaturized well plates and associated liquid dispensing systems, as shown in Figure 10.43.

Economic and technological forces are driving the continued push toward further miniaturization

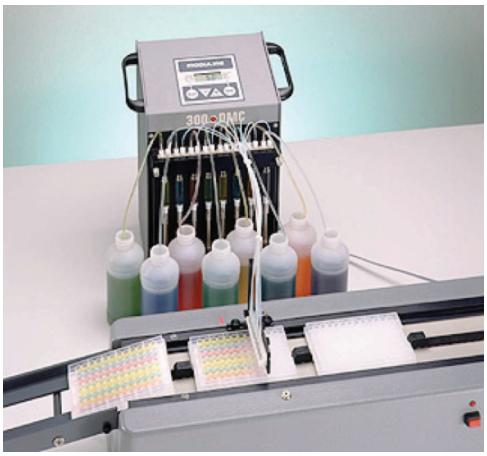


FIGURE 10.43 The 300 DMC dispenser configured for filling 96-well plates using eight different reagents (http://www.varivest.com).

of assays in plate wells for high-throughput screening (HTS) or high-volume diagnostics (Figure 10.2). In HTS, the number of potentially useful molecules is increasing, and the time during which the discovered molecules have a proprietary, strategic advantage is decreasing—in other words, the evaluation of more samples in a shorter amount of time with finite resources compels miniaturization. The number of samples has increased dramatically, primarily because of the concurrent advent of combinatorial and automated synthesis methods together with more precise biochemical or genetic definition of targets using genomic approaches. A first step in the miniaturization process for HTS and high-throughput diagnostics was from test tubes (1-mL range) to 96-well microplates (0.1-mL range) and, more recently, to 384-well (10-µL range) and 1536-well plates (1-µL range). With a 1536-well plate, ultra-high throughput (arbitrarily defined as 10^5 samples per day) becomes feasible. There are some reasons not to miniaturize and push to 1536-well plates and beyond. Certain assays become impossible if wells are too small. Cell-based assays, for example, cannot be miniaturized below 100–1000 cells without cell-to-cell variability becoming the dominant source of sample-to-sample variation, and cells are hard to accommodate in microfluidic systems. There is also substantial capital investment required for laboratories that are not already outfitted with pipetters and automated systems capable of reliably addressing 1536-well plates. For reading, 1536-well compatible readers tend to be more expensive than those limited to 384 wells, and the speed advantage of 1536 is difficult to obtain without expensive charge-coupled device (CCD camera) detectors. Some assays will suffer degradation in signal-to-noise ratio upon miniaturization (e.g., those that employ colorimetry), but many of the assays being developed have sufficient dynamic range to support a smaller format.[39] Engineering challenges associated with handling microliter volumes include: 1) evaporation, which limits the time of exposure to ambient humidity air and can cause sample-to-sample variability across the plate; and 2) fluid handling, which becomes increasingly difficult at smaller and smaller scales. As miniaturization proceeds still further, into the nanoliter-volume assays, evaporation becomes a much more significant issue. At a certain stage, a transition to fully enclosed systems, such as the microfluidic lab-on-a-chip (LOC) systems, becomes a necessity for evaporation and fluid metering. Xeotron developed the XeoChip, a bulk micromachined chip that contains thousands of three-dimensional nanochambers in an area smaller than a postage stamp (http://www.touchbriefings.com/pdf/855/fdd041_applied_tech.pdf). Each chamber serves as a protected reaction vessel for DNA synthesis and subsequent bioassay for use with conventional microarray scanners. The same system can be adapted for the synthesis of RNA, peptides, and other organic molecules.

Flow Cytometry

Flow cytometry is a method for high-throughput measurements of physical and chemical characteristics of cells and other microscopic biological particles. Typically, in commercial cytometers, optical measurements are made as particles are passing, single-file, through a sensing region. The particles are constrained to such a single line-up by hydrodynamic focusing. In this method, the sample flow is a narrow stream formed as the innermost flow of two concentric fluid flows (see Figure 10.44A). The faster the outer sheath flow, the narrower the inner sample flow is focused. Hydrodynamic focusing requires absolute laminar flow in all positions in the flow cell; from some high speed on, the inner stream will defocus through turbulence. Existing commercial

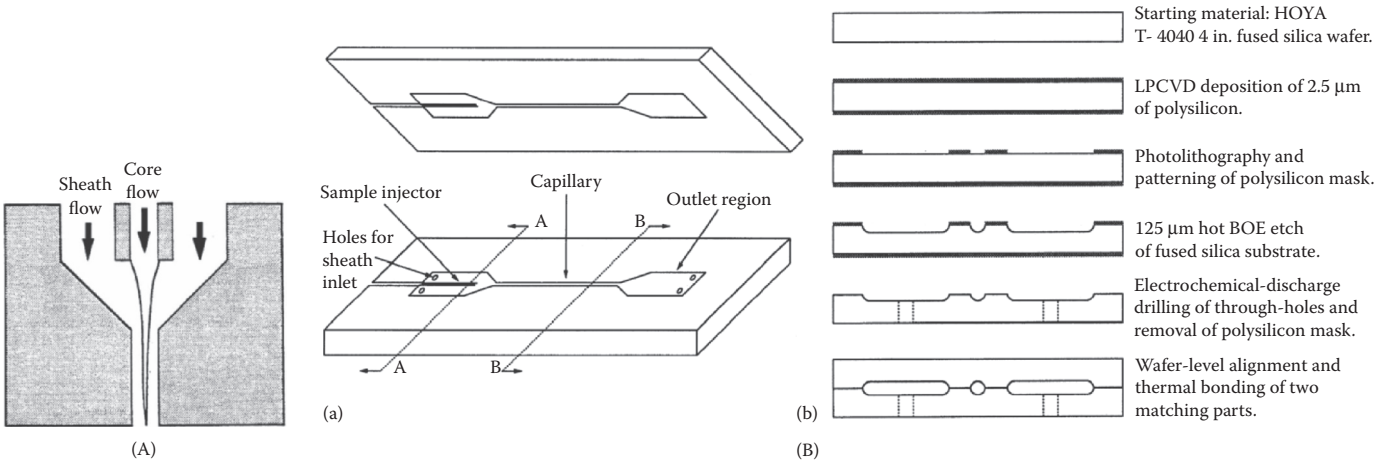


FIGURE 10.44 (A) Schematic of a flow chamber for cytometry. (B) (a) Exploded view of a fused silica flow chamber. The top and bottom have matching injection and outlet areas with a capillary joining them. Through-holes in the bottom plate are for sheath inflow and outflow. The sample is introduced through the injector into the center of the sheath flow. (b) The fabrication sequence shown along cross-section A-A. (After Sobek, D., S. D. Senturia, and M. L. Gray. 1994. In *Technical Digest: 1994 Solid State Sensor and Actuator Workshop,* 260–63. Hilton Head Island, SC.)

equipment requires frequent adjustment of the optical components, but simpler instrumentation is expected through the introduction of optical fibers replacing lenses for sample illumination and light collection. Shapiro and Hercher, in 1986,[40] suggested integration of flow channels, optical elements, and signal conditioning electronics as the next step toward more user-friendly equipment. The first such attempts led to complex micromachined sheath flow cells; for example, a five-layer stainless steel/glass laminate[41] and a four-wafer Si micromachined flow chamber.[42] Sobek et al.[43] have made a much simpler micromachined flow cytometer, which seems to put us back on the right track in terms of correct use of available micromachining tools. Cost consideration did result in a $30 \times 15 \times 1$ mm chip made from two symmetrically machined HOYA T-4040 synthetic quartz (fused silica) parts. These two parts, bonded at 1000°C, define the ovoid cross-section of the sheath flow channel, the circular cross-section of the sample injector, and the ovoid measuring capillary (Figure 10.44B). A smooth, flat surface of the capillary is essential for ensuring good optical measurements in the measuring region. The quartz starting material used here comes in 4-inch wafers, just like Si wafers. This is an important point, as most processing equipment is built to take Si wafers only. The resulting cytometer can focus a sample stream, introduced into the injector with a syringe, to a 10-μm width and can maintain this focusing to a mean velocity up to at least 10 m/s.

The next step in this development would be to find the correct package for the cytometer and an interface to the macroworld. Connectors to the sheath flow ports and the sample injection mechanisms are areas ripe for innovative approaches.

Surface Plasmon Sensor: Spreeta™

Many opportunities for innovative micromachining applications in optical immunosensors present themselves. A prime example can be found in surface plasmon sensors (SPRs). Incident plane polarized light at a specific angle can be almost totally absorbed into a thin metal film (e.g., 500 Å Ag or Au) deposited onto a prism, fiber, or waveguide surface because of resonant coupling of the light energy into a free-electron cloud within the metal (see SPR theory in Volume I, Chapter 5). This effect is known as *surface plasmon resonance* and can be extremely sensitive to the refractive index changes surrounding the metal film. By adsorption of species on the metal film, the specific angle of the light needed to create the resonant absorption changes (Figure 10.45).[44] This adsorption could be from gases or the binding of macromolecules onto the metal film. The main limitation is that the sensitivity depends on the molecular weight, or more correctly, the optical thickness of the absorbed layer, meaning that small

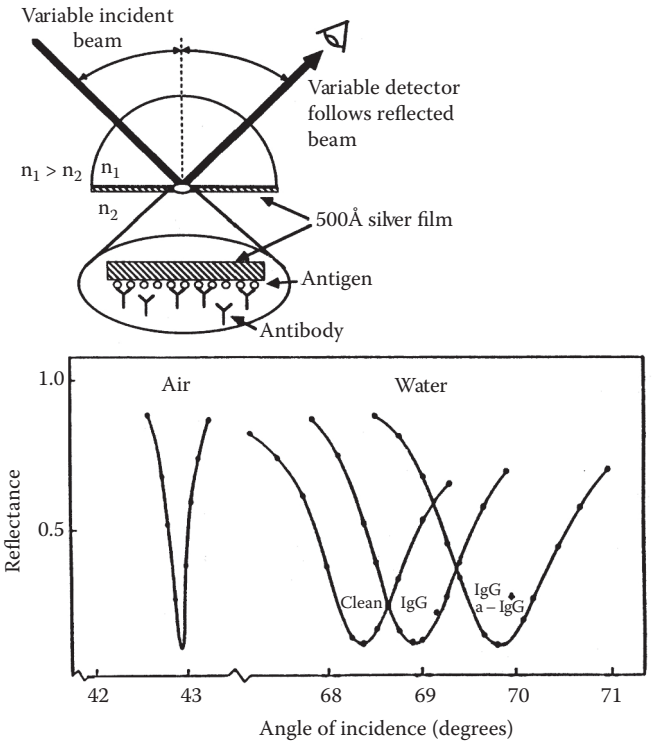


FIGURE 10.45 Measurements with a surface plasmon resonance sensor based on a prism. (From Sutherland, R. M., and C. Dahne. 1987. In *Biosensors: Fundamentals and applications,* eds. A. P. F. Turner, I. Karube, and G. S. Wilson, 655–78. Oxford, United Kingdom: Oxford University Press. With permission.)

molecules will not be good candidates. The SPR-based "immunosensor" from Amersham Pharmacia (Sweden) is a general-purpose, expensive, and rather large instrument incorporating a prism for detection and four parallel flow channels in removable sample chambers (in 2004 Amersham Biosciences, was acquired by GE Healthcare).

A significant amount of research effort is directed toward trying to miniaturize this large optical system. Comparison of diffraction gratings and prism systems clearly indicates that the diffraction grating is preferable for miniaturization. Garabedian et al.[45] attempted to micromachine the whole surface plasmon sensing system by implementing such a grating, but their resulting microinstrument is much too complex for manufacturing. An elegant alternative to both the grating and prism approaches may come from using an optical fiber or a planar optical waveguide. For example, Jorgenson and Yee[46] simplified SPR sensing dramatically by using such a fiber optic approach. As a sensing element, they used a multimode fiber with a section of the cladding removed and a thin layer of highly reflecting metal symmetrically deposited directly on the fiber core (see Figure 10.46a). This design has potential for simplified, robust, remote, and disposable sensing. Unlike traditional SPR measurements, employing a discrete excitation wavelength while modulating the angle of incidence, the SPR fiber optic sensor system uses a fixed angle of incidence and modulates the excitation wavelength. The refractive index can then be determined from the resonance spectrum in the reflected or transmitted spectral intensity distribution. In Figure 10.46b, the SPR sensing surface is enclosed in a flow-cell. Using micromachined planar waveguides in a flow-channel instead of the set-up shown here might result in a promising commercial product.

The most advanced example of a miniaturized SPR instrument is embodied in the miniature SPR sensor from Texas Instruments shown in Figure 10.47.[47–49] This sensor is also an apt illustration of "functional" packaging. As seen in the schematic in Figure 10.47a, light from a light-emitting diode (LED) is reflected from a gold surface in the LED

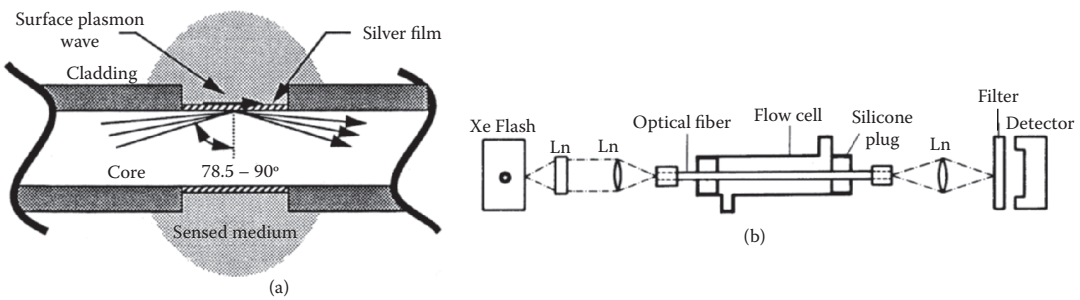


FIGURE 10.46 Simplified surface plasmon resonance (SPR) measurements. (a) Illustration of the SPR fiber. (b) Illustration of the SPR fiber enclosed in a flow-cell. (From Jorgenson, R. C., and S. S. Yee. 1993. A fiber-optic chemical sensor based on surface plasmon resonance. *Sensors Actuators B* B12:213–20. With permission.)

FIGURE 10.47 A miniature surface plasmon resonance sensor from Texas Instruments: Spreeta. (a) Spreeta geometry. (b) Packaged Spreeta unit. (c) Beta version of a cartridge-based Spreeta handheld instrument. (Courtesy of Dr. J. Elkind, Texas Instruments.)

optical plastic substrate/package and is picked up on a photodiode array embedded in the same package. Also contained in the same package is a polarizer for the light emitting diode and a temperature sensor. A plot of reflected light intensity versus angle is read out from the linear diode array as illustrated in Figure 10.48. The current noise floor of the sensor is 7×10^{-7} refractive index units. A beta-version of a cartridge-based Spreeta handheld instrument is shown in Figure 10.47c. Pricing for the disposable today is $30, and a handheld reader might cost about $250 (one-tenth as expensive as a Biacore).

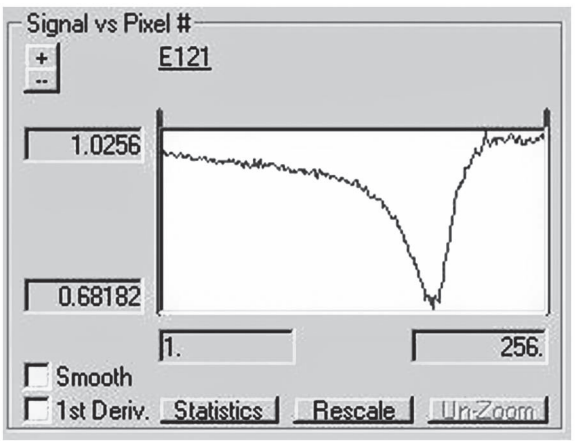


FIGURE 10.48 Surface plasmon resonance curve: reflected light intensity versus pixel number on the photodiode array. (Courtesy of Dr. J. Elkind, Texas Instruments.)

In Vivo MEMS Applications

Introduction

The many problems involving biocompatibility of in vivo devices were addressed in Chapter 5, "Biocompatibility and Surface Modification," where we learned that micromachining might actually provide a way of rendering surfaces more biocompatible by patterning them. Chemical surface modifications are also discussed. After the retrenching from in vivo devices in the 1990s, progress in materials and bio-MEMS has been so swift in the last decade that the problem of in vivo

sensing is ready for a new attack. This move back to in vivo sensing is further made easier by the progress in miniaturized telemetry and micropower systems (see Chapter 9, this volume).

In Vivo pH and Blood Gases

The cost of disposable sensors for in vivo use is not as big a concern as for in vitro disposable sensors; size is more critically important because the smaller the sensor, the smaller the intrusion on the patient. It is estimated that a catheter-based pH, CO_2, and O_2 sensor array, as shown in Volume II, Example 4.2, "An Electrochemical Sensor Array Measuring pH, CO_2, and O_2 in a Dual Lumen Catheter," may cost up to $250 and would serve a market of well over $600 million a year in the United States alone. Unfortunately, the warm, wet, saline in vivo environment perhaps puts forth the most severe environment in which Si sensors might be used. No Si-based in vivo chemical sensors are commercially available today. Biocompatibility is the single most complex issue facing in vivo sensor development, and it needs to be addressed up front in the sensor design. This is in line with our philosophy of starting a sensor design from the package. The package for an electrochemical in vivo pH, CO_2, and O_2 sensor typically is a dual lumen catheter modified to incorporate the sensor structure with the required chemically sensitive materials (see Volume II, Example 4.2). In building such a sensor, the author at first neglected to investigate biocompatibility issues up front and instead concentrated on the embedded micromachined electrochemical sensors.[50] The measurements obtained in saline solutions were encouraging, but the funding for continued research dried up before serious biocompatibility testing could be started. The development of the disposable in vivo pH, CO_2, and O_2 probe, the room-temperature micro electronic chemical smart sensor (RT-MECSS) took four years and $6 million. The RT-MECSS puts chemistry and electronics on separate planes while keeping the signal line as short as possible (see Volume II, Figure 4.99).[51] The latter represents a considerable improvement over both ISFET and EGFET (see Volume I, Chapter 7, "Ion-Sensitive Field-Effect Transistors, Extended Gate Field-Effect Transistors, and Isn't FET").

In Vivo Glucose Sensing

There have been many different technologies proposed for in vivo and ex vivo detection of glucose, and a number of ex vivo instruments and devices have been commercialized. While some of these, in particular the ex vivo ones (see above), have been very successful, a technology that meets the requirements of continuous in vivo detection of glucose and can be coupled to a miniaturized responsive delivery of insulin would be superior. Insulin delivery systems that have been developed are external and quite large in size. Ideally, miniaturized drug delivery systems would be in vivo and feature an integrated sensor for closed-loop control (see the responsive drug delivery pill in Volume II, Figure 4.100). Developing systems that are capable of both in vivo detection of biomolecules and delivering a drug is complex and requires the merging of technologies from chemistry, chemical engineering, micromachining, biomaterials engineering, and pharmaceutical science. For example, an ideal in vivo biosensor must be capable of long-term, rapid, reproducible detection of its target biomolecule at physiological levels, must be biocompatible, and must be amenable to miniaturization/microfabrication. Also, the insulin delivery system, controlled by the glucose sensor, should be capable of fast response and hold a sufficient amount of insulin to justify the long-term implantation of the device. The low success level achieved in designing in vivo miniaturized systems so far may partially stem from the fact that when developing these systems, researchers have placed more emphasis on one particular aspect than on others. Some of the persistent difficulties associated with glucose sensors include manufacturing reliability and consistency, short life spans, and cost. The number of biosensors with demonstrated capabilities for in vivo sensing is also limited.[52] Most sensors in the market today utilize glucose oxidase since the enzyme is stable and the cost is low. However, major disadvantages with using glucose oxidase are that a number of in vivo endogenous species are electroactive at the applied potential required for peroxide formation[52] and the production of H_2O_2 renders the eventual disintegration of these sensors.

As we saw in Volume I, Chapter 7, binding proteins associate with a particular analyte in a highly specific manner. Daunert et al. utilized this property in the development of reagentless optical biosensors.[53–56] Specifically, one of the many proteins of interest is the glucose and galactose binding protein (GBP). Glucose binding induces a hinge motion in GBP, and the resulting protein conformational change constitutes the basis of the sensor development. Labeling this protein with a fluorescent probe allows for the monitoring of this change without the addition of any external reagents or substrates. In order to obtain maximal signal versus background fluorescence, it is important that the fluorescent probe be attached to a single site (e.g., a unique cysteine) and at a position where maximum conformational change occurs. This approach has been shown to yield sensors with improved detection limits.[56]

As mentioned above, the most undesirable effect of current daily glucose monitoring is the frequent fingerstick glucose monitoring that the patient needs to perform. Even the more successful implantable devices, such as the MiniMed Continuous Glucose Monitoring System (currently being evaluated in several clinics throughout the United States), need frequent recalibration. This is due to the direct correlation between sensor response and the time-dependent enzyme activity. The use of GBP labeled with fluorescent compounds that have a long lifetime can afford sensors with longer service lives. Lifetime measurements are dependent on the glucose concentration but independent of the concentration of the fluorescently labeled protein present.[57] Thus, since these sensors do not require an active enzyme, a reduction in calibration frequency can be achieved. It is this type of sensor we envision combining with the smart pill in Volume II, Figure 4.100 (pharmacy-on-a-chip).

μ-TAS and LOC in Sample Preparation for Molecular Diagnostics

Introduction

Micromachining and genomics parallel each other in perceived value to society and much is expected from a merging of both fields. Bio-MEMS, in principle, holds the potential to revolutionize molecular diagnostics, as it could lead to much more compact, potentially disposable, and well-integrated molecular diagnostic systems. As we saw in Volume II, Chapter 8, the use of photolithographic techniques has already permeated molecular biology; they are used, for example, to make DNA arrays (e.g., Volume II, Figure 8.16). To enhance DNA hybridization speed, DNA arrays have been embedded in all types of fluidic structures. The principal advantages associated with a bio-MEMS, μ-TAS/LOC approach include the following:

- Small instrument footprint
- Multiplexing so that several analyses can be completed simultaneously
- Microfluidics for fluid manipulation, often obviating the need for traditional macropumps
- Excellent potential for integrating various components onto the device
- Reduced volume requirements for sample and reagents (often expensive)

Here we analyze specifically bio-MEMS contributions for sample preparation in genomics. Like packaging in the semiconductor industry, sample preparation is an area that has been less vigorously pursued than developing the assay itself; perhaps it is perceived as somewhat less glamorous, as is packaging in the IC industry. But, like IC packaging, sample preparation is of overwhelming importance to the end customer. The benefit in developing miniaturized devices for sample preparation in genomics is the low dead volume requirement associated with accuracy, costs of reagents, and small overall footprint, and the difficulty is having enough copies for the analyte to be detectable. The latter implies a flow system reducing the volume during the sample preparation step.

Types of Samples

Procedures for preparing samples for chemical analysis vary dramatically from sample to sample and from analysis to analysis. To reduce the complexity of the discussion, we single out for closer scrutiny the challenges faced in preparing samples for molecular diagnostics. Specifically, we investigate the feasibility of performing sample preparation in a small-footprint integrated instrument capable of performing all the aspects of a common biological DNA assay. The preparation of biomedical samples

prior to the diagnostic procedure is often complex, laborious, and time-consuming. It varies, even within the more limited application of DNA testing, depending on the sample type, for example, urine, whole blood, plate bacterial cultures, serum, stool, tissue, etc. Today, the preparation of such samples involves several different instruments and intensive manual operations such as centrifugation, heating and cooling, mixing with reagents, filtration, etc. For most clinical diagnostics, sample preparation is actually the rate-limiting step. In the case of whole blood samples, red blood cells (RBCs) must be separated from leukocytes [white blood cell (WBC) count for 0.2% of human blood cells: 1 μL of whole blood contains 5×10^6 RBCs and 7×10^3 WBCs) and other cells (bacteria, virus particles, etc.). Red blood cells do not have nuclei, the source of nucleic acids, but they do contain contaminants, which interfere with nucleic acid amplification [polymerase chain reaction (PCR)] and other enzymatic manipulations (e.g., both proteins and metal complexes such as hemoglobin bind with nucleic acid). Isolated leukocytes and bacteria samples must be lysed before the extraction of nucleic acid can take place. A traditional procedure for blood DNA isolation comprises separation of the white cell fraction from the whole blood, lysis with detergent, and digestion of proteins with proteinase K (1–6 hours) to purify the target material from contaminating proteins, followed by extraction with phenol-chloroform and precipitation of DNA with alcohol. DNA is not soluble at high ethanol concentrations, so it precipitates out as long strands. Salts, such as sodium chloride, also greatly aid in precipitating DNA. Tissue samples primarily contain cells and connective tissue, which also must be degraded before DNA extraction. The first step in the DNA extraction procedure from tissue is typically to boil a small amount of the tissue in a 5% Chelex solution. The boiling helps to break up the cells in the preparation and denatures proteins. The Chelex protects the DNA sample from DNAases* that might remain after the boiling and that could subsequently contaminate the samples. In the diagnostic approach of the future one envisions a small footprint instrument and a test specific disposable cartridge providing a "sample to answer" solution.

* DNAases are enzymes that occur naturally in all body tissues. They cut DNA, rendering it unsuitable for PCR. Magnesium ions are essential cofactors for DNAases. Chelex resin binds with cations, including Mg^{2+}. By binding with the magnesium ions, the Chelex resin renders DNAases inoperable, thus protecting DNA from their action.

General Sample Preparation Protocols

Sample Size

Miniaturization is a mixed blessing in the case of the amount of sample required to detect a given analyte concentration. That volume is determined by the following:

$$V = \frac{1}{\eta N_A C_i} \tag{10.1}$$

where η = sensor efficiency, with a value between 1 and 0

N_A = Avogadro's number (6.02×10^{23} mol^{-1})

C_i = the concentration of analyte i (moles/L)[58]

The sample volume required is fundamentally dictated by the concentration of the analyte one wants to measure. Manz et al. examined the use of very small sample sizes and concluded that microfluidics is a desirable approach for many diagnostic applications.[59] They summed up their analysis in a graph of target analyte concentration versus required sample volume. This graph shows that many biological chemicals associated with clinical chemistry assays (between 10^{14} and 10^{20} copies/mL) and immunoassays (between 10^7 and 10^{18} copies/mL) might be readily assayed with very small sample volumes, in the range between picoliters and microliters. However, Petersen et al. pointed out that numerous chemicals (and organisms) are routinely present at much lower concentrations, from less than 100 to 10^7 copies/mL.[58] These low-concentration samples include most sources of DNA, which must be detected and analyzed in a growing number of new diagnostic tests. In Figure 10.49, Petersen et al. expand the boundaries of Manz's original graph, which spanned a concentration range from 10^8 to 10^{21} copies/mL and a volume range from 10 mL to 10^{-18} L, and clearly show that the minimum sample volume required for accurate DNA assays is a relatively large volume of 100 μL.

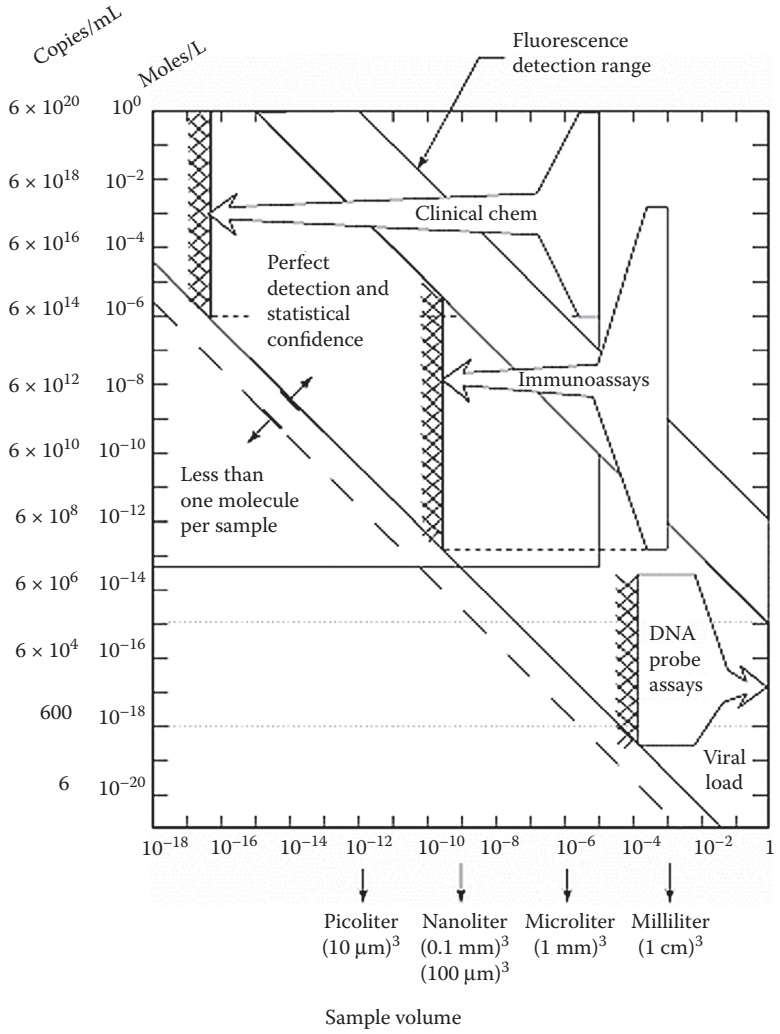


FIGURE 10.49 Sample volume. Scaling of concentrations and volumes.[58] (Courtesy of Dr. Kurt Petersen, Cepheid.)

The amount of original sample required, depending on the analysis method, and the number of DNA copies needed for detection may range from some microliters to several tens of milliliters. Figure 10.49 demonstrates that typical clinical chemistry assays (between 10^{14} and 10^{20} copies/mL) and immunoassays (between 10^{7} and 10^{18} copies/mL) are readily carried out with very small sample volumes, in the picoliters and microliters range, but numerous DNA assays deal with less than 100 to 10^{7} copies/mL[60] and may require milliliter sample volumes. To summarize, modern DNA diagnostic assays face mounting demands to detect organisms or DNA mutations at very low concentrations, often less than 100 copies per milliliter, in raw biological samples such as blood or urine. Such sensitivity requirements set fundamental physical limitations on the minimum quantities of the starting sample. Using the current nucleic acid hybridization formats and stringency control methods, it remains difficult to detect low copy number nucleic acid targets even with the most sensitive reporter groups (enzyme, fluorophores, radioisotopes) and associated detection systems (fluorometers, luminometers, photon counters, scintillation counters), necessitating a DNA amplification step, for example, using polymerase chain reaction (PCR) (see Chapter 2, "Polymerase Chain Reaction"). PCR is usually considered sensitive enough to detect as few as twenty copies of a target in any given, purified, biological sample. But if the required assay sensitivity is 100 copies per mL, then a minimum of 200 μL of the original sample must be processed to ensure statistical confidence in the result of the assay. If microfluidics and small PCR chambers are used, this means that DNA preconcentration is needed, for example, on silica-type capture surfaces

or on arrays of micromachined Si posts.[60] The latter reduces the sample size from macro to micro, which can be accommodated in a microsystem by using some type of a flow system. A modular approach, with a variety of disposable cartridges containing different fluidic options and sensor chemistries in a small fixed instrument, is the most probable solution.

Lysing

Sample preparation for DNA analysis in an integrated device includes releasing the nucleic acid from the sample by cell disruption or lysis, removing contaminants (purification), and recovering/concentrating the nucleic acid. There are many mechanical and nonmechanical cell lysis or disruption methods available. The results of these may be evaluated in terms of the activity level of a cellular enzyme released to the suspending medium measuring the efficiency of the disrupting process. We briefly review both types of lysing approaches and consider their relative merit for integration in microsystems.

Mechanical Methods

Currently, in large nonintegrated tabletop equipment, intracellular products are released from microorganisms or tissue mainly by mechanical disruption of the cells. In this process, the cell envelope is physically broken, releasing all intracellular components into the surrounding medium. There are several types of equipment for mechanical cell disruption commercially available. In a high-pressure homogenizer, a pump forces slurry of sample through a restricted orifice valve. High pressure (up to 1500 bar) is followed by an instant expansion through a special exiting nozzle. Cell disruption is accomplished by three different mechanisms in this system: impingement on the valve, high liquid shear in the orifice, and sudden pressure drop upon discharge, finally causing an explosion of the cell. This approach, although very efficient at cell disruption, is difficult to miniaturize. In a bead mill, cells are agitated in suspension with small abrasive particles. Cells break because of shear forces, grinding between beads, and collisions with beads. This approach is slightly more amenable to miniaturization. Another widely applied method is cell lysis through high-frequency ultrasound. These devices generate intense sonic pressure waves in liquid media. Under the right conditions, the pressure waves cause formation of microbubbles, which grow and collapse violently. Called *cavitation*, the implosion generates a shock wave with enough energy to break cell membranes. Modern ultrasonic processors use piezoelectric generators made of lead zirconate titanate crystals (see Chapter 8, this volume). Vibrations are transmitted down a titanium metal horn or probe tuned to make the processor unit resonate at 15–25 kHz. A relatively new development is a cordless ultrasonic processor supplied by BioSpec Products (Bartlesville, OK). With a tip diameter of 1/8 inch, it easily fits into microtubes and ninety-six-well titer plates. Ultrasonic disintegrators generate considerable heat during processing. For this reason, the sample should be kept ice-cold, and one should be aware that free radicals can be generated during sonication and that these radicals may react with biomolecules. Ultrasonic disruption is an attractive approach for miniaturization of the cell lysis process.

Mechanical disruption methods suffer several drawbacks. Because cells are broken completely, *all* intracellular materials are released. Therefore, the DNA must be separated from a complex mixture of proteins and cell wall fragments. Cell debris produced by mechanical lysis often consists of very small fragments, making the solution difficult to clarify. Moreover, complete product release often requires more than one pass through the membrane disruption device, exacerbating the problem by further reducing the size of the fragments. These fragments are difficult to remove by continuous centrifugation because the throughput of a centrifuge is inversely related to the square of the particle diameter. Filtration is complicated by the gelatinous nature of the homogenate and by its tendency to foul membranes. Because of these drawbacks, several nonmechanical cell lysis approaches have gained favor.

Nonmechanical Methods

Nonmechanical lysis includes thermal, electronic, and chemical methods. Chemical permeabilization

of cells can be achieved in numerous ways.[61,62] The most important are enzymatic lysis (e.g., with lysozyme), detergents, and osmotic shock. Chemical methods extract intracellular components from microorganisms by permeabilizing outer cell wall barriers. This can be achieved with organic solvents that act by creation of channels through the cell membrane. Some of the solvents that have been used for this purpose are toluene, ether, phenylethyl alcohol, dimethyl sulfoxide (DMSO), benzene, methanol, and chloroform.

Chemical permeabilization can also be achieved with antibiotics, thionins, surfactants (e.g., NP-40, Triton X-100, Brij, Duponal, CHAPS, SDS, etc.), chaotropic agents, and chelates. An important example in the latter category is EDTA, widely used for permeabilization of Gram-negative microorganisms. Its effectiveness is a result of its ability to bond the divalent cations Ca^{++} and Mg^{++}. These cations stabilize the structure of outer cell membranes, by bonding the lipopolysaccharides to each other. Once these cations are removed by EDTA, the lipopolysaccharides are removed, resulting in increased permeability of the outer cell walls. Chaotropic agents, such as urea and guanidine, are capable of bringing hydrophobic compounds into aqueous solutions. The relative ability of various anions to influence the structural stability of protein (and DNA) in general follows the series $F^- < SO_4^{2-} < Cl^- < Br^- < I^- < ClO_4^- < SCN^- < Cl_3CCOO^-$. This ranking is known as the *Hofmeister series* or *chaotropic series*. Fluoride, chloride, and sulfate salts are structure stabilizers, whereas the salts of other anions are structure destabilizers. Salts that stabilize proteins enhance hydration of proteins and bind weakly, whereas salts that destabilize proteins decrease protein hydration and bind strongly. In other words, the denaturing effect of chaotropic salts is related to destabilization of hydrophobic interactions in proteins. High concentrations (6–8 M) of compounds that tend to break hydrogen bonds, such as urea and guanidine salts, also cause denaturation of proteins. These substances apparently disrupt hydrogen bonds which hold the protein in its unique configuration. However, there also is evidence suggesting that urea and guanidine hydrocholoride may disrupt hydrophobic interactions by promoting the solubility of hydrophobic residues in aqueous solutions. Detergents break open (lyse) cells by destroying the fatty membrane that encloses them. This releases the cell contents, including DNA, into the solution. Detergents also help to strip away proteins that may be associated with the DNA. Enzymes can also be employed to permeabilize cells, but this method is often limited. Enzymes used for enzymatic permeabilization are glycanases, proteases, and mannase. The main drawbacks of employing enzymatic means for recovering intracellular products in large-scale processes are cost and the necessity of removing the lysing enzyme from the product. Basic proteins, such as protamine, or the cationic polysaccharide chitosan can permeabilize yeast cells. Antibiotics that affect bacterial cell-wall biosynthesis, causing loss of viability and often cell lysis (penicillins and cephalosporins, bacitracin, cycloserine, vancomycin), may also be used. Similarly, mammalian cells can be permeabilized by exposure to several natural substances such as streptolysin or even viruses.

While cells exposed to slowly varying extracellular osmotic pressure are usually able to adapt to such changes, cells exposed to rapid changes in external osmolarity can be mechanically injured. This procedure is typically conducted by first allowing the cells to equilibrate internal and external osmotic pressure in a high-sucrose medium and then rapidly diluting away the sucrose. The resulting immediate overpressure of the cytosol is assumed to damage the cell membrane.

Repeated freezing and thawing will thermally lyse a cell, and in some instances, the target DNA may be released by elevating temperatures simultaneously with the denaturing step performed in PCR.[63]

Electronic lysis of cells has been reported by Cheng et al.[64] and Lee and Tai.[65] Of the different types of nonmechanical lysing, the electronic and thermal approach is the most amenable to miniaturization. Chemical approaches should be ranked according to the amounts of chemicals required; the lesser the amount, the more suited the approach to miniaturization

Table 10.10 summarizes some lysis methods and subsequent purification and recovery steps.

Table 10.10 Cell Lysis, DNA Purification, and Recovery

Process	Technical Difficulty	Costs	Efficiency	Ease of Miniaturization
Lysis				
Mechanical	High	High	High	Of the mechanical lysis methods reviewed, ultrasonic is most amenable
Chaotropic	Medium	Medium	Medium	Depends strongly on the volume of reagents involved
Thermal	Low	Low	Low	Easy to miniaturize (sometimes this step can be combined with polymerase chain reaction)
Purification				
Ion exchange	Medium	Medium	High	Easy to implement on a miniaturized platform
Beads	Medium	Medium	Medium	Somewhat more cumbersome
Molecular weight cutoff membranes	Low	Low	Low	Easy to implement on a miniaturized platform
Recovery				
Silica elution	Low	Low	High	If a flow-through system is used, miniaturization of low-salt buffer is easy to accommodate
Bead elution	Low	Low	Medium	More cumbersome
Membrane elution	Low	Low	Low	Easiest to miniaturize

Examples of Integrated Sample Preparation

Over the last ten years, several attempts have been made to miniaturize molecular diagnostics, inclusive of sample preparation, in the areas of biological warfare agents and point of care.[58,66–74] These efforts involve microfluidic systems, in which the entire protocol from sample collection through PCR, and in several cases all the way to detection, is embedded in one integrated instrument. In Table 10.11 we review a few of these prototype DNA analysis devices (expanded from Yuen et al.[75]). Extensive efforts pertain to isolating white blood cells from whole blood through the use of micromachined filters (arrays of posts, tortuous channels, comb-shape, and weir-type)[66,76,77] and different designs of silicon microfilters (3–10 μm) across microchannels.[78,79] All these designs feature a high efficiency for separating red from white blood cells. More recently, Yuen et al.[75] CNC-machined a Plexiglas-based module integrating white blood cell separation and nucleic acid amplification reactions. This module comprises a heater-cooler for thermal cycling and a series of 254 × 254 μm microchannels for transporting human blood and reagents in and out of an 8–9-μL dual-purpose (cell separation and PCR) glass-silicon microchip. The filter in this case was of the weir type (3.5-μm filter gap). A region in the human coagulation Factor V gene (226-bp) was directly amplified by the microchip-based PCR from DNA released from white blood cells isolated from the weir filter section.

In the sample preparation approach of the company Tecan Boston (now defunct), a compact disc (CD) (the LabCD; see Volume I, Figure 6.88 and Figure 10.50 below), was used as a sample preparation stage (see also Table 10.11 below). The physics behind this centrifugal microfluidic platform is covered in Volume I, Chapter 6, and its prototyping is detailed in Chapter 5, this volume. The Tecan Boston prototype combined sample preparation and PCR, while detection was performed separately on an ethidium-bromide-stained gel electrophoresis slab. The disposable disc is shown in Figure 10.50. The sample preparation protocol employed is as follows: 1) mixing raw sample (5 μL of whole blood or *Escherichia coli* suspension) with 5 μL of 10 mM NaOH; 2) heating to 95°C for 1–2 minutes to lyse cells, releasing DNA and denaturing proteins inhibitory to PCR; 3) neutralization of lysate by mixing with 5 μL of 16 mM Tris-HCl (pH 7.5); 4) mixing of neutralized lysate with 8–10 μL of liquid PCR reagents and user-selected primers; and 5) thermal cycling. Samples of whole blood or *E. coli* were run on the discs shown in Figure 10.50 using a rotational profile that sequentially drives fluids past capillary valves *d*, through mixing channels *e*, to receiving reservoirs. Reservoirs *a*, *b*, and *c* contain sample, NaOH, and Tris-HCl, respectively. At the lowest rpm,

TABLE 10.11 Prototype Integrated DNA Analysis Devices

Sample Preparation	Biochemical Reactions	Detection	Fabrication Methods for Device	Source
RNA purification from 1 mL of serum samples spiked with different copy numbers of an RNA transcript that contained a partial HIV sequence	Reverse transcription-PCR, nested PCR, DNase fragmentation and dephosphorylation, and terminal transferase labeling	GeneChip hybridization	Conventional CNC machining	Anderson et al.[67]
White blood cells isolation from <3 μL of human whole blood	PCR using integrated heater-cooler to provide thermal cycling	Not applicable	Conventional CNC machining (standard microfabrication techniques for microchips)	Plexiglas microchip module[75]
5 μL of whole blood or *E. coli* suspension	PCR using thermoelectric devices on a spinning personal computer board platen	Detection is performed separately by electrophoresis	CNC machining of polycarbonate or poly(methylmethacrylate) disc combined with lithography in polydimethylsiloxane for smaller fluidic features	Tecan Boston[74]
Starting from the DEP collection of intact bacteria in a 70-μL sample to DNA hybridization assay (*SLT1* gene of *E. coli*),	Strand displacement amplification, dielectrophoretic collection of *E. coli* bacteria, electronic DNA hybridization	Electric field driven DNA assay	Laminated flexible substrates with electrode arrays (using integrated circuit lithography)	Nanogen[73]
Spores in paper products (0.5 gm in 10 mL of water)	PCR on material eluted from a Si DNA extraction chip (set of 200-μm-high columns, 18 μm in diameter, with a pitch of 34 μm)	Gel electrophoresis in a separate step	Acrylic or polyester plastics and Si chips	Cepheid[60]
30 green-fluorescent protein-transfected THP-1 cells in 25 μm of whole blood	Real-time-PCR	Real-time PCR	Droplets containing surface-functionalized superparamagnetic particles	Institute of Bioengineering and Nanotechnology[80]

Source: Expanded from Yuen et al.[75]
CNC, computer numerical control; PCR, polymerase chain reaction.

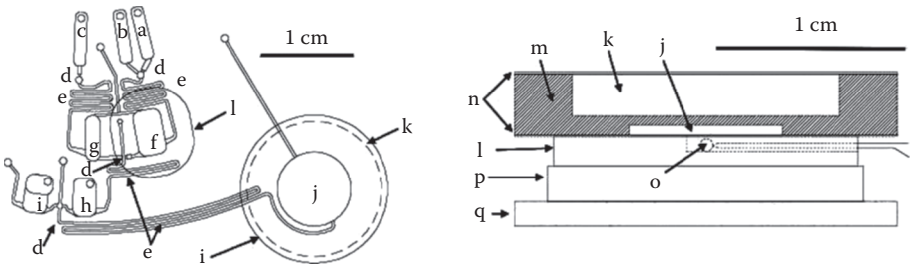


FIGURE 10.50 On the left is the polymerase chain reaction (PCR) structure. The center of the disc is above the figure. The elements are as follows: (a) sample; (b) NaOH; (c) Tris-HCl; (d) capillary valves; (e) mixing channels; (f) lysis chamber; (g) Tris-HCl holding chamber; (h) neutralized lysate holding chamber; (i) PCR reagents; (j) thermal cycling chamber; and (k) air gap. Fluids loaded in (a), (b), and (c) are driven at a first rpm into reservoirs (g) and (f), at which time (f) is heated to 95°C. The rpm is increased and the fluids are driven into (h). The rpm is increased and fluids in (h) and (i) flow into (j). On the right, the cross-section shows disc body (m), air gap (k), sealing layers (n), heat-sink (l), thermoelectric (p), personal computer board (q), and thermistor (o). (Courtesy of Dr. G. Kellogg, Tecan Boston.)

sample and NaOH mix in mixing channel *e* and move to reservoir *f* while Tris-HCl moves to holding reservoir *g*. Chamber *f* is the lysis chamber, which heats the sample/NaOH mixture up to 95°C for 1–2 minutes to lyse the cells and denature the proteins. Heating of the lysis chamber is accomplished by a thermoelectric element mounted on a personal computer-board platen *q* on the spindle of the rotary motor and connected by a slip-ring to a stationary power supply and a temperature controller. The lysis chamber sits on top of a brass heat-sink *l* mounted directly onto the thermoelectric element *p*. A thermistor, *o*, mounted in the heat-sink, enables closed-loop temperature control. By driving up the rpm, the Tris-HCl from the holding chamber *g* and the lysed material in reservoir *f* mix in the serpentine mixing channel neutralizing the lysate and then move into chamber *h*. At yet a higher speed fluids in *h* and *i* flow into the PCR chamber *j*. The PCR reagents used were either reconstituted beads or a solution of necessary reagents held in reservoir *i*. Rapid cycling with small temperature gradients in chamber *j* is achieved by confining the fluid to a thin layer, 0.5 mm thick, adjacent to another thermoelectric element. Above the fluid, an air pocket, *k*, serves as a thermal insulator. As a result, most of the heat transfer occurs via the thermoelectric. Slew rates of $\pm 2°C/s^{-1}$ have been achieved with fluid volumes of 25 µL; the temperature gradients across the liquid are less than 0.5°C (from simulation and thermocouple measurements). With slight design changes, twelve structures of the type shown in Figure 10.50 may be placed on a 12-cm plastic disc. As we remarked when reviewing fluid propulsion options in Table 1.8 of this volume, a centrifuge platform is an attractive sample preparation stage (wide dynamic range, relatively insensitive to fluid composition, etc.). There is one major issue yet to be resolved, though: unless one works with all dry reagents (and has wetting take place only upon first use of the disc) the liquids held in the sample reservoirs will evaporate and distribute over time as there is no vapor barrier to stop them from expanding within the whole available open volume in the disc. The fluid gating in the disc works only for fluids, not for vapors. Electrokinetic and acoustic fluidic platforms suffer the same problem, so a major remaining challenge for developers of diagnostic platforms based on microfluidics is to implement both fluid and vapor gating. One possible solution is shown in Chapter 1, this volume, Figure 1.12, where reagents of interest are sealed in small blister pouches in a plastic laminate and only when the liquid is needed does a mechanical roller break the liquid and vapor seal. In Figure 8.52 we suggest yet another solution involving ferrowax valves on a CD platform. An external magnet can move these ferrowax plugs to pump or valve fluid plugs on the CD platform.

Kido at RotaPrep, Inc., developed the OrbiPrep system, an integrated, centrifugal sample preparation system. This station does not involve PCR yet, so it is not listed in Table 10.11 but PCR will be integrated in the future. It combines magnetism,[81] centrifugal acceleration, dynamic reagent dispensing,[82] and the Coriolis effect[83] in a system that performs automated, simultaneous milling/homogenization of up to twelve biological samples, followed by the extraction, washing, and elution of purified DNA from each sample. The system consists of a computer-controlled centrifuge-like device that accepts either two-sample rotors or twelve-sample rotors. Figure 10.51a shows a two-sample rotor on the device with its main components labeled. The device is also able to dispense accurate volumes of up to four discrete reagents into the center ring-shaped distribution trough of the rotor while it rotates. This is accomplished by fixing four dispensing needles to the lid of the device. When the lid closes (Figure 10.51b), the tips of the needles enter the trough to rest below the top plane of the rotor. Walls within the trough divide the entering stream of liquid equally among the sample preparation units of the rotor. Figure 10.52 shows the different stages of operation of the OrbiPrep. After loading the samples with buffer into the milling chambers through the sample loading ports, the milling stage involves a slow rotation at 400 rpm that causes the free moving magnet within each milling chamber to oscillate in the radial dimension as the rotor turns through the static magnetic field. The yellow arrows depict the trajectory of the free-moving magnet. The resulting milling action homogenizes biological samples, releasing their contents. The slow rotation prevents the liquid in the milling chamber from moving to the next

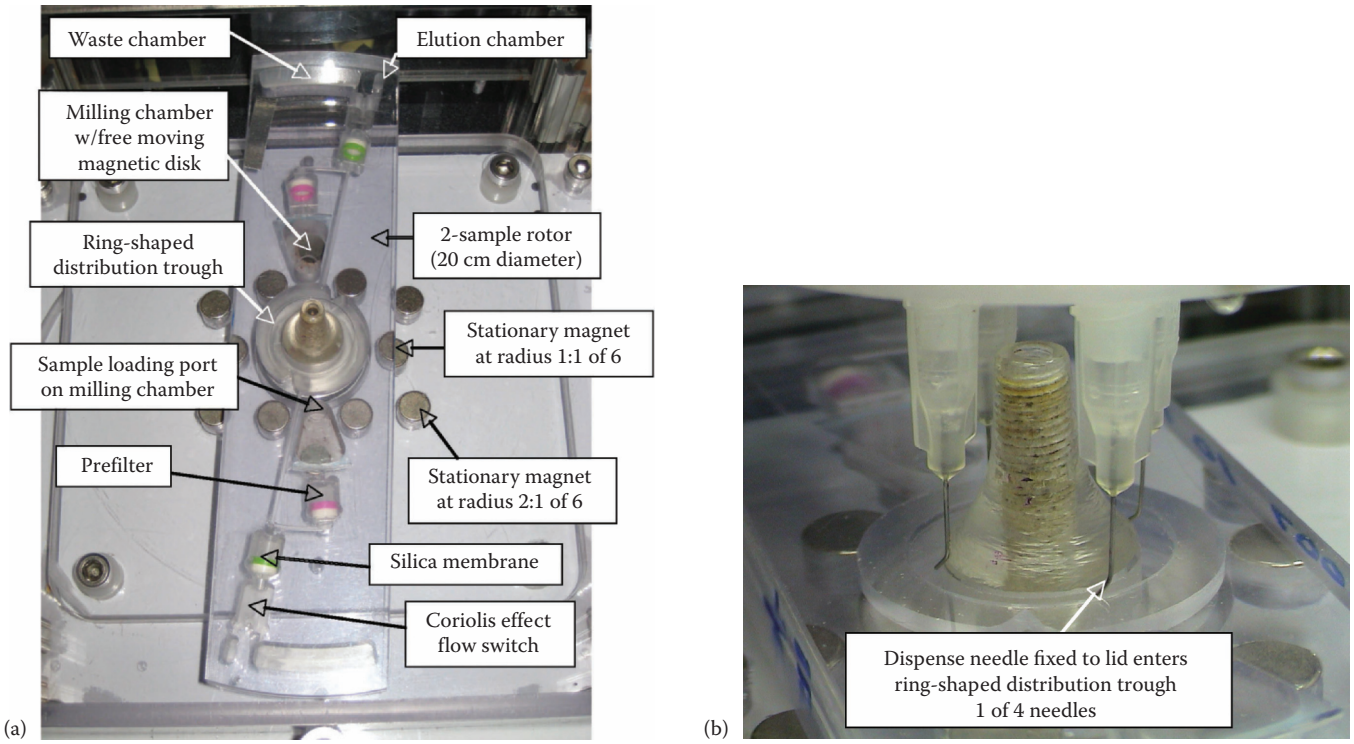
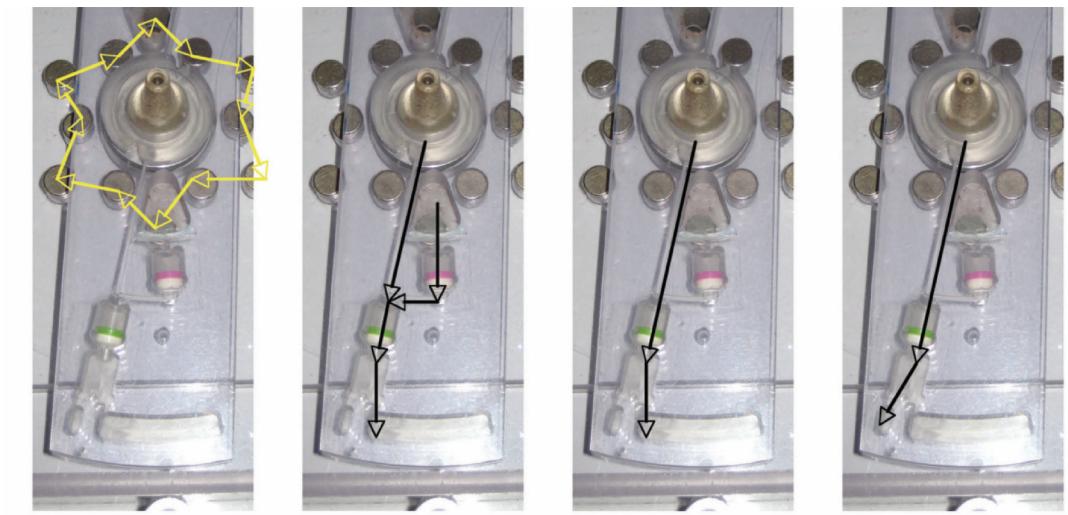


FIGURE 10.51 OrbiPrep system (RotaPrep, Inc.): an integrated centrifugal sample preparation system. (a) Sample preparation station with open lid. (b) Lid closed.



Milling	DNA binding	Washing 1 & 2	DNA elution
400 RPM	8,000 RPM	8,000 RPM	8,000 RPM
CW rotation	CW rotation	CW rotation	CCW rotation
1 min	1 min	2 min	1 min

FIGURE 10.52 OrbiPrep system (RotaPrep, Inc.): an integrated centrifugal sample preparation system. Shown are the different stages of operation of the OrbiPrep.

chamber held there by capillary forces (i.e., the rotation rate is below the burst frequency of the capillary valve separating the two chambers, see Volume I, Equation 6.141). In the DNA binding step, the rate of rotation is increased to 8000 rpm. This causes liquid from the milling chamber to flow downstream (i.e., the rotation speed is now increased above the burst frequency of the capillary valve), through the prefilter, toward the silica membrane. Before the liquid reaches the silica membrane, it is mixed with ethanol that started to be dispensed through the first dispensing needle into the distribution trough. The black arrows depict the flows of the respective liquid streams and the mixing point immediately upstream of the silica membrane. The silica membrane extracts DNA from the combined liquid stream that passes through it down into the Coriolis effect flow switch. Since the rotor rotates clockwise, the stream enters the waste chamber. Washes 1 and 2 follow, as wash buffers 1 and 2 are dispensed from needles 2 and 3 in series. Since the liquid from the milling chamber has been drained, the liquid flow into the silica membrane comes only from the distribution trough, as the black arrows indicate. The rotation continues to be clockwise, so the stream from the silica membrane is directed into the waste chamber. The rotor reverses direction and resumes rotation at 8000 rpm before the elution buffer is dispensed from needle 4. The elution buffer enters the silica membrane and dissolves the DNA from it before flowing downstream into the Coriolis effect flow switch. The counterclockwise rotation causes the stream to flow into the small elution chamber, ready for use in downstream processes such as amplification by PCR. The whole process takes approximately 5 minutes to complete.

Pipper et al.[80] made a free droplet containing surface functionalized superparamagnetic particles into a virtual μ-TAS with a (sub)microliter volume. Aside from acting as the force mediators for actuating the droplet in a magnetic field, the superparamagnetic particles serve as a solid support for the sequential performance of biochemical processes. Depending on the task, the droplet becomes a pump, valve, mixer, extractor, or thermocycler. In an all-automated experiment, thirty green fluorescent protein (GFP)-transfected THP-1 cells are isolated from 25 μm of blood, 100-fold preconcentrated, purified, lysed, and subjected to a real-time PCR (RT-PCR) targeting the transfection vector all within 17 minutes. Fast thermocyles of 7 seconds take place on a disposable substrate under timespace conversion by rotating the droplet clockwise over different temperature zones.

Beyond the prototype stage, Cepheid, in 2001, introduced the GeneXpert for pathogen detection (http://www.cepheid.com). This instrument goes from raw biological sample to PCR and detection within 30 minutes. The detection technology is based on DNA amplification and real-time, fluorometric identification and quantitation. The real-time optical reading of the amplified product is based on four colors (for four targets) per disposable cartridge (the ICORE, which stands for intelligent, cooling/heating, optical reaction). Four ICORE can be loaded at the same time into a GeneXpert instrument. Cepheid's test, which looks for telltale genetic sequences, takes as little as 72 minutes. The GeneXpert machine costs $30,000 to $150,000, depending on how many tests it can run at once. Each individual test cartridge is priced at $42.

MEMS in Industrial Automation, Including Environmental Monitoring Markets

Introduction

According to the Nexus III report, industrial automation, including environmental monitoring, constituted 5% of a $12 billion MEMS market in 2004 and will be a 4% of a $25 billion MEMS market in 2009 (see Figure 10.16). MEMS pressure sensors, gas sensors, flow sensors, accelerometers, infrared imagers, angular rate sensors, temperature probes, etc., all have utility in modern industrial processes and automation. MEMS is also expected to play an important role in environmental applications. Primary environmental MEMS products are expected to be in the area of distributed sensor networks for detection of air and water pollution and in downscaled instruments such as microspectrometers, micro-gas chromatography systems, micro-ion mobility spectrometers, infrared detectors, and perhaps even electronic noses and tongues.

As a wide variety of non-silicon and traditional technologies are involved in the construction of novel industrial and environmental monitors it is harder to come up with accurate MEMS market data in industrial/automation and environmental control. Silicon technology has not penetrated this application as deeply as it has the automotive world. With a plethora of solutions and the unifying theme of silicon-based sensing missing, MEMS market numbers vary much more than for automotive sensing.

Research in industrial and environmental instrumentation miniaturization is especially challenging, as science requirements call for advances in performance with simultaneous large reductions in device mass, volume, and cost. Many conventional instruments operate at or near theoretical limits, and reducing the scale (volume and mass) often leads to sharp reduction of performance. This situation calls for new fundamental principles for instrumentation technology development. In other instances, we shall see, miniaturization is making for better instruments.

From the time a picture of a gas chromatograph (see Volume I, Figure 1.22) the size of a matchbox was printed on the cover of the April 1983 issue of *Scientific American*[84] (Terry and Angell built a first prototype between 1974 and 1975), there has been a continued effort to micromachine analytical instruments. Few MEMS-based microinstruments reached the market. They are still relatively large, tabletop or suitcase size, with only a few key components based on micromachining. One well know example is the Agilent MicroGC (Figure 10.8), looking significantly less integrated than the device suggested in the 1983 *Scientific American* article (the MTI gas chromatograph column itself is not micromachined; only the inlet and detector are). The wisdom of using Si micromachining for microinstrumentation is now in doubt, but looking at the miniaturization job from a precision engineering perspective, one becomes aware of the many alternative manufacturing options available.

MEMS Solid-State Gas Sensors

Most current solid-state gas sensors operate on the principle of modulating the resistance of a metal-oxide thick film by adsorption of gas molecules to its surface at high temperature (see Taguchi sensor, Volume II, Figures 9.11 to 9.13). Typical metal-oxide sensor materials include SnO_2, TiO_2, ZnO, WO_3, and Fe_2O_3. Each metal oxide is sensitive to different gases. For example, tin oxide is most effective at detecting alcohol, hydrogen, hydrogen sulfide, and carbon monoxide, and zinc oxide is useful for detecting halogenated hydrocarbons. Unfortunately, most metal oxides are adversely affected by humidity, which must be controlled at all times. In addition, variations in material properties require that each sensor be individually calibrated.

The market in 2000 for small solid-state gas sensors and sensor systems in the toxic and combustible gas arena was some $120 million in the United States alone. Broken down by product, the U.S. market for these chemical sensors comprises about $25 million for hydrogen sulfide, $20 million for total hydrocarbon monitoring, $12 million each for carbon monoxide and oxygen deficiency, and small percentages for a myriad of other gases. The use of a ceramic tube in a classical Taguchi sensor (see Volume II, Figure 9.11) maximizes the utilization of the power from the coiled Pt wire heater inside the tube, so that most of the power is used to heat the tin oxide covering the tube. The tin-oxide paste is applied on the outside of the ceramic body over thick-film resistance measuring pads by dip coating and is sintered at high temperature. The sintering stabilizes the intergranular contacts where the sensitivity of the gas sensor resides. The structure is not a design suited for mass production, as it involves excessive hand labor. The problems with the thick-film structure are mainly in the area of reproducibility: compressing and sintering a powder, the deposition of the catalyst (e.g., the size of palladium crystallites and how close they are to the intergranular contacts), and the use of binders and other ceramics (e.g., for filtering) are all very difficult to control. It is obvious from our discussion of the Taguchi sensor in Volume II, Chapter 9, that the application of IC techniques to Taguchi sensor manufacture can improve the state of the art dramatically in terms of the required power budget; a power reduction by a factor of ten is quite common. Unfortunately, as explained in Volume II, Chapter 9, no thin-film tin-oxide gas sensor can yet compete with the thick-film-based traditional Taguchi gas sensor in terms of sensitivity. If one could make a

thin-film tin-oxide with the same gas sensitivity as a thick film, then micromachining the heater elements, and thin-film deposition of the gas-sensitive material would further improve reproducibility and make for an excellent product. For long-term stability, relatively low temperature semiconductor catalyst should be used (perhaps operating at temperatures <200°C). Figaro Engineering Inc. of Japan pioneered the Taguchi sensor technology in 1962. This company is still one of the leaders today.

A major obstacle to more rapid progress with thin-film micromachined gas sensors remains that the sensitivity of thin films to gases is not yet very well understood. The currently available devices, already quite small and inexpensive, will only be replaced quickly when newer technology improves upon the existing product in terms of reproducibility, cost, sensitivity, response time, concentration range, specificity, reversibility, stability, and power consumption.

At this time, the best course for future progress in the Taguchi-type sensor is to continue to optimize sensitivity of thin-film oxide semiconductors, use planar Si devices (reproducibility and cost), and concentrate on implementing less power-consuming microheater elements. Along this line MicroChemical Systems (MiCS; http://www.microchemical.com) produces micromechanical Si thin-film tin oxide gas sensors. Details about the nature of the thin-film tin oxide are scarce (Figure 10.53).[85] For very low-cost and less demanding applications, the MiCS sensor might succeed, but for more demanding applications, planar thick-film structures on ceramic do present a better option.[86]

An interesting new research direction in the MEMS gas sensor arena might be to take advantage of the micromachined ceramic (Al_2O_3) structures, as illustrated in Chapter 5, Figure 5.9, for both heaters and catalyst support.

The MGS1100 carbon monoxide sensor from Motorola (Schaumburg, IL)[87] is not on the market anymore. The device incorporated a tin-oxide, thin-film sense resistor over a polysilicon resistive heater. The sense resistor and the heater reside over a 2-µm-thick silicon membrane to minimize heat loss through the substrate. Consequently, a mere 47 mW was sufficient to maintain the membrane at 400°C. There is a total of four electrical contacts: two connect to the tin-oxide resistor, and the other two connect to the polysilicon heater. The simplest method to measure resistance is to flow a constant current through the sense element and record the output voltage. Although the electronics were implemented correctly, as you would expect from Motorola, it seems like the fact that a gas sensor is only as good as the selective catalytic nature of its surface was forgotten in this sensor implementation.

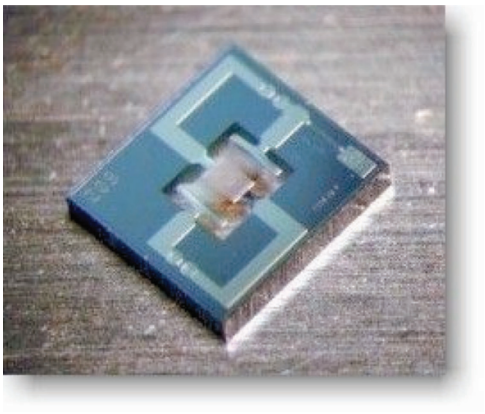


FIGURE 10.53 Microelectromechanical system gas sensor by MicroChemical Systems (MiCS; http://www.microchemical.com). Top view of the sensor chip. The key elements of the sensor are a heater that maintains the gas sensitive catalyst layer at a specific temperature (I), a thin supporting thin dielectric layer allowing very low power consumption (II), and a sensitive catalyst layer that changes resistance upon exposure to gases (III). MiCS uses proprietary technology to deposit precise amounts of nanoparticle metal oxide material on the micro hot plate. The very small grain size provides high stability and sensitivity.

The market for humidity sensing in the United States is about the same size as or larger than the total toxic gas sensor market ($120 million). A particularly interesting set of micromachined gas sensors, including humidity sensors, is based on thin-film-coated chemiresistors and acoustic devices.[88] We discussed the Jet Propulsion Laboratory (JPL) microhygrometer in Chapter 8, "SAW Example Application: Dew Point/Humidity Sensor" (Figure 8.26).

Solid State Gas Sensor Arrays: Electronic Noses and Tongues

The selectivity of a chemical sensor is the ability to measure only one chemical in the presence of many other chemicals. Because of the lack of perfect selectivity of chemical sensors, sensor arrays are often

implemented. The solid-state gas sensors described above are one of the types of sensors used in arrays in electronic noses to fingerprint complex gas mixtures (other types use conductive polymers, coated quartz microbalances, or carbon black in nonconductive polymer matrix). The sensitivity of Taguchi sensors to different gases depends principally on the choice of catalytic metals and the operating temperature. Although the selectivity of the individual sensors is limited, qualitative and quantitative gas analysis can be performed using pattern-recognition techniques. Optimally, the sensors have a wide dynamic range and a small correlation to the other sensors in the array so that each contributes as much new information as possible. When applying pattern-recognition techniques to chemical analysis one defines chemometrics. The combination of multiple gas sensors and signal analysis using pattern-recognition techniques is the concept behind the *electronic nose* or *E-nose* (Figure 10.54). The sensor array is part of a neural network sensor array (NNSA) consisting of the array, a sampling system, a flow control system, a preprocessor, and a pattern-recognition system.

An electronic nose is an example of biomimetics at work (see also Chapter 2, this volume). In animals, the olfaction system (smell) is perhaps the most mysterious of the senses: even the sense of taste, for example, consists in large part of smell. Two American scientists, Dr. Richard Axel and Dr. Linda B. Buck, who suggested an answer to the enigma of how, with only a limited number of receptors (350), people can smell more than 10,000 different odors, and recall them later, were awarded the 2004 Nobel Prize in Physiology or Medicine.[89] The initial event in the process of olfaction is the recognition of an odorant by olfactory receptors—proteins found in

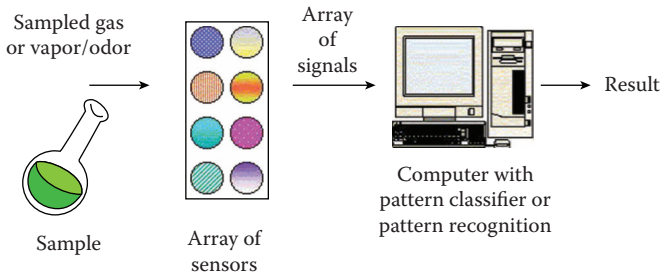


FIGURE 10.54 The electronic nose: an example of biomimetics. Electronic noses were touched upon in the context of neural network software in Chapter 9, this volume.

the olfactory epithelium. Olfactory receptors convert the "information" in odorant molecules into electrical signals that are interpreted by the brain. In animals a layer of mucus dissolves the arriving scents and separates out different odor molecules so that they arrive at the receptors at different speeds and times, so there is some type of *snot chromatography* involved as well. The brain is able to interpret this pattern to distinguish a diverse range of smells. In contrast, an artificial nose consists of a much smaller array of chemical sensors, typically between six and twelve, connected to a computer with neural net software capable of recognizing patterns of molecules. (Electronic noses are also touched upon in the context of artificial neural network software in Chapter 9, this volume.)

E-noses were popular at the end of the twentieth century and have been successfully used in a number of applications, e.g., the quality estimation of ground meat, the identification of different paper qualities, the classification of grains with respect to microbial quality, the screening of irradiated tomatoes. NASA used a JPL developed electronic nose on the space shuttle; the manufacturers of Coors beer, Evian water, and Starbucks coffee use electronic noses to test the quality of their products; and E&J Gallo Winery is trying to find out whether they can spot bad corks. E-noses also have been tested to monitor pollution, spot explosives and drugs, and diagnose diseases.

Unfortunately, electronic noses have had a hard time finding acceptance in the marketplace. A lot of training is required of both sensor arrays and the technical staff running the equipment, and the instruments are still large and expensive. The chemical sensors used in the sensor arrays drift too much, so that too-frequent calibration makes the use of the instruments cumbersome. Recognizing this problem, surviving electronic nose companies are now adding mass spectrometry to their sensor arrays (Figure 10.55).

Alpha MOS, Neotronics Scientific Lt., Illumina Inc, AromaScan Inc., and Cyrano Sciences, Inc., were all important players in this field in the 1990s. Alpha MOS, a French company, and AromaScan Inc., a British manufacturer, were the first to develop the electronic nose, or sensor array systems (SAS)

FIGURE 10.55 The Prometheus from Alpha MOS is the world's first odor and VOC analyzer that combines a highly sensitive fingerprint mass spectrometer with a large choice of patented multisensor array technologies to create an extremely flexible analysis platform (http://www.alpha-mos.com/).

technology, for commercial use. At Pittcon 2000, Cyrano showcased a portable electronic nose, the Cyranose 320, which distinguished odors using an array of thirty-two sensors. Each sensor was made of a composite of conductive carbon black and a nonconductive polymer. Today, only Alpha MOS continues as an electronic nose company. There is little left of all the other original contenders. In 2004, the Pasadena-based Cyrano Sciences was acquired by Smiths Detection for $15 million plus earn-out. Neotronics has been amalgamated into the Zellweger Analytics consortium, Illumina Inc. went on to do great things in DNA arrays (see Volume I, Chapter 7, "Bead Arrays for Detection and Analysis of DNA Sequences"), and AromScan became Osmetech, focusing on molecular diagnostics (http:// www.osmetech.com) instead.

The scientific principles of the electronic tongue are similar to the electronic nose but are applied to liquids rather than vapors. An electronic tongue incorporates electrochemical membrane-based sensors (see Volume I, Chapter 7, "Ion-Selective Electrodes") mimicking the five types of taste buds, which sense sourness, saltiness, bitterness, sweetness, and umani (deliciousness). The sensors in the array, like our taste buds, exhibit global selectivity rather than absolute selectivity; in other words, they are not selective to one chemical species, but to a wide variety. The ASTREE Electronic tongue from Alpha MOS is a liquid taste analyzer designed for taste assessment in solutions or solids dissolved in liquids. Like the human tongue, it provides a global "taste fingerprint" of a complex mixture of organic/inorganic compounds. For all tastes, it performs analysis with similar or better detection threshold than the human tongue.

Electrochemical Gas Sensors and Optochemical Gas Sensors

Solid-state semiconductor devices have been heralded as the replacement for electrochemical sensors; however, the electrochemical gas sensor sales are experiencing a growth rate of 30% per year. The latter is probably due to the slow rate of improvement in the selectivity of solid-state gas sensors and superior selectivity of electrochemical sensors brought on the market by companies such as City Technology (http://www.citytech.com) and Neotronics.[90] There is quite some consolidation going on in gas sensor companies. Neotronics and Sieger have been amalgamated into the Zellweger Analytics consortium, and more recently, First Technology has acquired City Technology and Capteur Sensors, while EEV is now part of E2V Technologies. There is plenty of opportunity in the application of MEMS to room-temperature electrochemical sensors, although these sensors are typically small already, they are not made in batches yet.

An optochemical gas sensor approach developed by Quantum Group Inc.[91] was adopted by First Alert as an inexpensive and selective fire alarm (http://www.qginc.com/products.html). This room-temperature sensor comprises a porous, semitransparent substrate (e.g., a glass frit) into which a self-regenerating chemical sensor reagent, selective for CO, is impregnated. Reaction with CO is fast, very selective and changes the color of the reagent. This color change is picked up by a diode photo-detector. Oxygen in the air regenerates and clears the substrate slowly back to its original color (~30 minutes). In this fashion, the sensor is biomimetic, simulating human response to airborne toxins such as carbon

monoxide, mercury, ethylene oxide, volatile organic materials, and hydrogen sulfide. This is a very inexpensive and promising approach, and with further fine-tuning of the chemistry of various sensor reactants, we might expect a large family of products based on this optochemical approach, with a selectivity akin to an electrochemical sensor but at a cost similar to that of a solid state gas sensor.

The biomimetic sensor has a functional lifetime of over one year, and the mimicking of the human response to toxins is achieved by the use of a molecular encapsulant that encapsulates at least one gas component. Work is concentrating on extending the sensor's lifetime and speeding up the regeneration of the sensor. The optochemical gas sensor from Quantum Group Inc. could be miniaturized by using nontraditional manufacturing techniques, but finding selective chemistries for other gases, such as NO_2, O_3, and H_2S, is a higher priority.

ISFETs

ISFET technology was covered in Volume I, Chapter 7, "Ion-Sensitive Field-Effect Transistors, Extended Gate Field-Effect Transistors, and Isn't FET." In the United States, ISFETs are manufactured only at the University of Utah, Salt Lake City. Orion Fisher, Corning, Sentron, and Rosemount, as well as UniFet, Inc. (San Diego, CA) sell ISFET-based pH sensors in the United States but do not produce them in-house. Honeywell sells the DuraFET, and in a detailed evaluation study, Sandifer and Voycheck called this sensor remarkable rugged, very stable, and accurate.[92] Orion and Rosemount buy their pH ISFETs in research quantities from Neuchatel (Switzerland) and target the food industry where glass pH electrodes cannot be used for food (pH measurement in dough, for example). Unisensor recently acquired Sentron (www.unisensor.ch). These commercially available ISFETs are for pH only and feature a simple pH sensitive inorganic membrane on the gate. To modify FETs with polymer membranes for detection of other ions is difficult because of the thickness of the membrane and the need for a stable internal reference electrolyte.

As we pointed out in Chapter 1, this volume, usually it is simpler and less expensive to build electrochemical sensors on a nonsilicon substrate (see, for example, the reference electrode in Chapter 1, this volume, Figure 1.18).

Gas Chromatography

Gas chromatography is a technique widely used for the separation and analysis of gaseous samples. A typical gas chromatograph (GC) consists of a carrier gas supply, a sample injection system, a separating fused silica column, an output detector, and a data processing unit, as depicted in the block diagram in Figure 10.56.[93] Separation of the sample vapors is achieved via their differential migration through a capillary column (see Volume I, Chapter 6). A precise and reproducible volume of sample vapor is injected at the input of the column by a valve and swept through it by an inert carrier gas. The

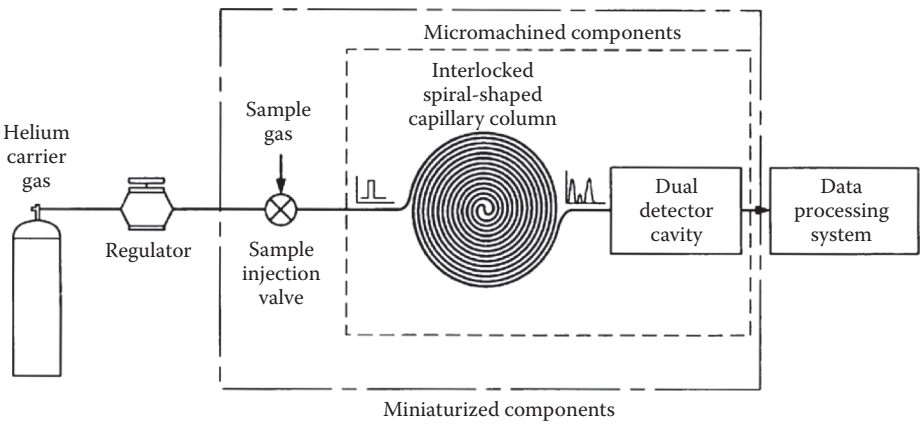


FIGURE 10.56 Functional block diagram of a gas chromatography (GC) system. (After Kolesar, E. S., and R. R. Reston. 1995. Separation and detection of toxic gases with a silicon micromachined gas chromatography system. *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications, National Aeronautics and Space Administration (NASA) and The Aerospace Corporation*. Houston, TX, October 30–November 3, 1995, pp. 102–14.[93])

column is lined with a liquid stationary phase, a substance capable of absorbing and desorbing each of the component vapors depending upon their unique partition coefficients. The migration rate of each vapor along the column depends on the sample injection pressure, the carrier gas velocity, the temperature, and the degree to which the vapor is absorbed by the stationary phase. The column's output is thus a series of vapor peaks separated by regions of pure carrier gas. To detect those peaks, the output gas stream from the column is passed over a detector that measures a particular property of the gas, such as the thermal conductivity, which can be related to the concentration of sample vapor in the carrier gas. The detector produces a signal, which is amplified and used in the analysis of the sample mixture; the identity and quantity of each vapor in the mixture can be determined from, respectively, its retention time in the column and the area under its output peak.

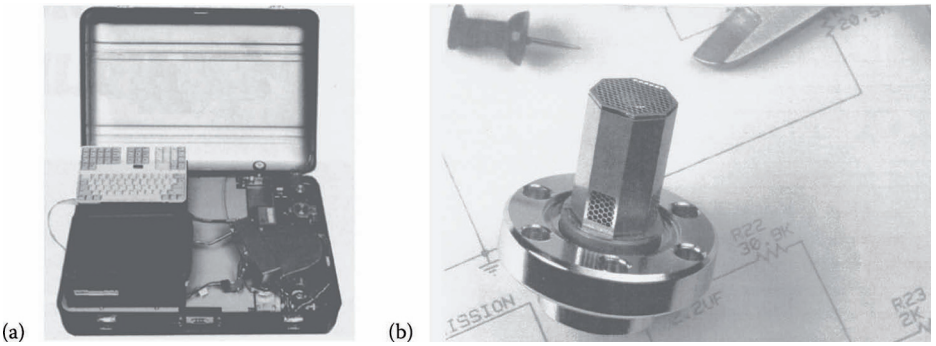
Gas chromatography systems tend to be large, fragile, and bulky pieces of laboratory equipment, and miniaturization would present many advantages:

- Integration of the instrument into a process
- Portable and handheld units
- Fast response time (1 minute or less) in comparison with a large unit with comparable plate numbers (see Volume I, Chapter 6, "Scaling in Analytical Separation Equipment")
- Less calibration and carrier gases needed
- Low dead volume
- Higher accuracy

The first integrated GC, consisting of a long separating column, a sample injection valve, and a thermal conductivity detector (TCD), fabricated on a single Si wafer, was developed at Stanford as early as 1974 (see Volume I, Figure 1.22). The Agilent MicroGC, based on that original work, uses a traditional fused silica capillary column but retains the other micromachined components (Figure 10.8). By the time MTI produced its first GC, with a 100-µm diameter, fused silica GC columns had become commercially available, and they proved to be superior to the Si micromachined columns. The reasons behind their superiority must be sought in the perfect symmetry of the monolithic silica columns. The smoothness of the stationary phase coating the inside of a capillary column is a function of the profile symmetry. The uniformity of coating thickness is extremely important to maintain very sharp separation peaks. The micromachined designs described in the literature usually consist of a cover sheet of Pyrex glass, which is anodically bonded to a U-shaped, V-groove, or rectangular groove etched in Si.[94] Columns also have been made by joining two U-grooved halves. All these designs have a major potential deficiency: they are bound to accumulate coating on the corners or joints, leading to a less perfect separation.[95] Analytes spend a longer time in areas where the coating is thick than in places where it is thin, leading to a broadening of peaks. Thermal gradients might also be induced by joining dissimilar materials, again causing peak broadening. It also should be realized that the column has to be replaced on a regular basis because of aging and fouling. Therefore, the price of a total integrated system would have to be very low.[96]

Research efforts for making better micromachined GC columns are continuing,[93] but since existing fused silica columns are perfectly adequate, the focus would be more gainfully put on other GC problem arenas. Specifically, micromachining could be of great immediate use in the arena of better, more reproducible, miniaturized valves and fast heating and cooling of thermally isolated column and injector for fast temperature-programmed chromatograms.

A Hewlett-Packard (HP) development of a handheld GC may be the only effort that combines both excellent micromachining capabilities with a long-standing instrumentation knowledge. RVM Scientific, Inc. has also built a novel handheld miniaturized gas chromatograph for detection of contraband drugs in cargo containers and chemical weapons treaty verification. There are several small GC products now being introduced to the market. The SnapShot, for example, is a handheld GC from Photovac International Inc. (Deer Park, NY). The instrument has a battery pack providing 4–6 hours of operation and may be used for personal exposure monitoring. Besides Agilent, Stanford Research Instruments is offering rugged GCs for field use.

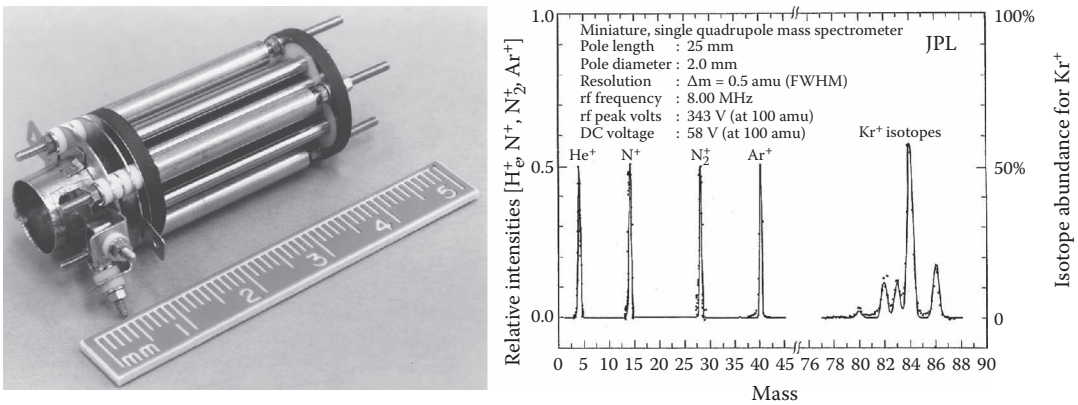


(a) (b)

FIGURE 10.57 (a) A gas chromatography-mass spectrometry system built in a suitcase for real-time field use. (Courtesy of Lawrence Livermore National Laboratory, California). (b) A 1-cm scale array of quadrupoles for gas analysis (traditional mechanical precision machining). Built by Ferran Scientific, California.

Mass Spectrometry

Mass spectrometers (MSs) are extremely versatile tools that have found, at least for the moment, their greatest use in laboratory science. The determination of the component atomic and molecular masses and their abundance is indeed essential for the understanding of almost any chemical and/or process analysis. Lack of broader use is attributed to their large mass, volume, and power requirements. Especially powerful would be a miniaturized gas chromatography-mass spectrometry (GC-MS) combination, with the GC for the isolation of the various compounds of a complex mixture and the MS for further identification. Innovative vacuum systems, smaller GC ovens, new sample inlets, and advanced electronics are now making some of this possible. For example, a GC-MS system built in a suitcase by Lawrence Livermore National Laboratory in California for real-time field use is shown in Figure 10.57a.[97] Ferran Scientific has successfully implemented a 1-cm scale array of quadrupoles for gas analysis (Figure 10.57b), and a similar MS is under development at Jet Propulsion Laboratory (JPL) in Pasadena, California. JPL researchers suggest that the Ferran instrument lacks in mechanical precision and hence resolution. In their own work, a miniaturized quadrupole mass filter, one machined with traditional but very precise machining methods (Figure 10.58a) and one dramatically smaller in size fabricated employing LIGA (not shown), was pursued. The JPL research team deemed quadrupole and ion-trap mass spectrometers most suited for miniaturization, as their operation does not require a permanent magnetic field, and they can be fabricated as parallel arrays of spectrometers.[98] The array concept is extremely important considering the following simple scaling insights. Miniaturization of an instrument dimension, *l*, by a



(a) (b)

FIGURE 10.58 Miniature mass spectrometry. (a) Jet Propulsion Laboratory's small quadrupole array made by traditional, but very precise, machining. (b) Spectrum generated from the single quadrupole mass spectrometer in (a).

factor k leads to a cubic mass and volume reduction (l^3 scaling) and an input aperture area reduction (l^2 scaling), and thus sensitivity reduction of k^2. To offset the lost sensitivity, one can construct an array of k^2 spectrometers, all working in parallel. One ends up with a spectrometer array with a sensitivity comparable to its larger cousin but with mass and volume reductions of k. A present quadrupole or magnetic-sector instrument may be of the order of 10^4 cm^3 volume and 10–15 kg mass and may require about 20 W power for operation. The JPL traditionally machined microspectrometer, a 4 × 4 array of nine miniature quadrupole mass analyzers (Figure 10.58a), has a mass of less than 0.8 kg (including electronics), is 20 cm^3 in volume, and draws less than 8 W power. The rod alignment, straightness, and roundness accuracy for a mass range of 1 to 250 amu and a mass resolution of 0.5 amu (FWHM) at 200 amu must be 0.15% (see mass spectrum in Figure 10.58b). Initially, the rods were machined at 2 cm long, but they can be further reduced to 1.5 or even 1 cm. The alignment was carried out using a machinable ceramic alignment jig to accurately position the rods. The holes in the jigs are machined using a conventional, high-precision lathe (±0.0002-inch tolerance). With the LIGA technique, the attempt is to make the quadrupole rods 1 mm high and 80 µm in diameter, but these specifications have not yet been met. LIGA might provide more accuracy for the straightness of vertical walls, but the tolerances in the x- and y-directions are still dependent on the aligner. As we may deduce from Volume II, Figure 6.5, a 0.15% tolerance on dimensions below 100 µm is quite a challenge.

Optical Absorption Instruments

The most common analyzers used in analytical chemistry for both gases and liquids are spectroscopic absorption techniques. A considerable amount of effort has been put into miniaturization of such equipment with both traditional and nontraditional machining tools.

Consider, as an example of an environmental application, the detection of methane by means of infrared absorption (based on Beer's law, see Volume I, Equation 7.133). From Volume I, Chapter 7, we know that to obtain a better sensitivity, a longer optical path is required. To accomplish this on a small-footprint instrument of a few square centimeters, a set of LIGA-machined micromirrors may be positioned in a folded configuration, as illustrated in Figure 10.59. The infrared mirrors are preferably made from solid electroplated gold or by evaporating gold on poly(methylmethacrylate) (PMMA) walls. The vertical height of the mirrors is limited by the spot size of the infrared laser beam. The light intensity loss in the miniaturized instrument is due to absorption and scattering by the micromirrors. Gold is the preferred material for the micromirrors because of its high reflectivity in the mid-infrared region. Superpolished gold surfaces can reflect as much as 98% of an incident infrared beam in the specular direction. However, micromachined gold mirrors, electroformed or evaporated on PMMA walls, will likely have an inherent surface roughness, which may cause scattering of incident beams. Working in the infrared range presents an advantage in this case, as the roughness of the mirrors will typically be small compared with the wavelength used. With an angle of incidence close to the normal direction, one can typically accommodate up to 100 reflections before serious losses occur. With a 1.6-cm distance between mirrors, this means 1.6 m of optical path-length, which may not be adequate for a 0.1-ppb sensitivity but would serve well for applications that require only a few parts per million of sensitivity. The conclusion is that one might well be able to miniaturize a gas absorption cell, but a miniaturized infrared analyzer of this sort will be suited only to less demanding applications. No products of this

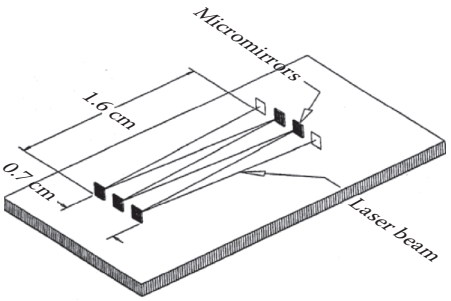
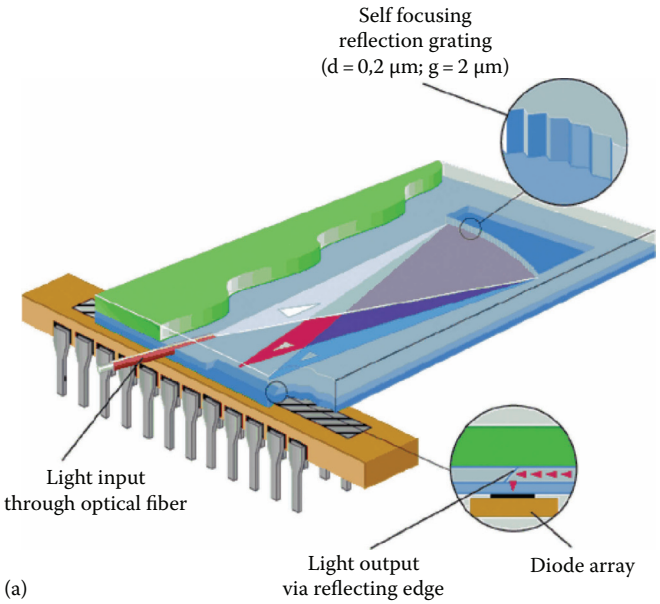
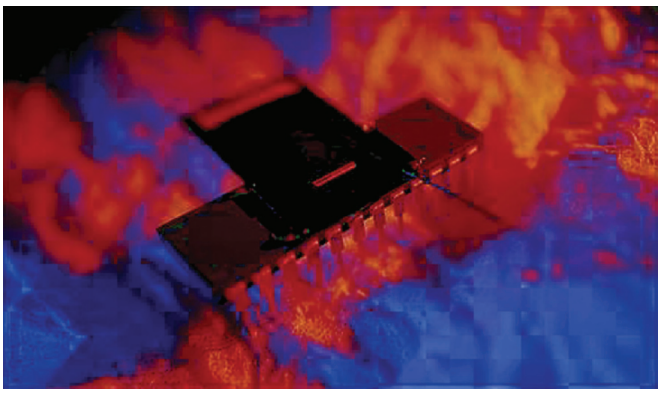


FIGURE 10.59 Layout of a folded optical system. (After Desta et al. 1995. *Microlithography and metrology in micromachining.* Austin, TX: SPIE. With permission.)



(a)



(b)

FIGURE 10.60 Microspectrometer from STEAG microParts. Schematic (a) and photograph (b).

sort are on the market yet. In Volume I, Chapter 7, "Scaling in Absorption Spectroscopy and Cavity Ring-Down Spectroscopy," we introduced the cavity ring-down spectroscopy (CRDS) technique, which represents a way around the sensitivity issue when scaling down an optical instrument. As far as we are aware, there is no MEMS version available of the CRDS.

The STEAG microParts VIS microspectrometer is a monolithic dielectric slab waveguide with integrated focusing echelette grating produced by a micromolding technique. The part sells for $100–$300. Light to be analyzed is coupled into the device by an optical fiber. Light reflected from the grating is detected by a diode array. The detector attached at the focal plane is a Hamamatsu photodiode array with 256 pixels. The STEAG microParts microspectrometer is shown in Figure 10.60.

MEMS in IT Peripherals and Telecommunications Markets

Introduction

In the past, automotive applications were setting the pace for increases in manufacturing volumes of MEMS sensors. The future will present quite a different scenario: huge growth rates in MEMS sensors are expected for applications in consumer electronics (CE) products, with a focus on IT and communications, and extensions into health care, life sciences, security systems, domestic appliances, and others. Once a MEMS device succeeds in the automotive application, with its very stringent specifications, the door is wide open to other more relaxed markets, even the consumer market. Analog Devices, for example, is now putting its surface micromachined accelerometers in video games, car and notebook alarms, smart pens, etc.

IT peripherals is a general term used to describe technologies that help produce, manipulate, store, communicate, or disseminate information. From Figure 10.16, this application area today constitutes already the largest MEMS market, dominated by hard disc drive heads, inkjet print heads and microdisplays. Interestingly, at the moment of their market introduction, the first two of these products were not even considered MEMS (see also the Preamble on the MEMS Market section, above).

To illustrate that there are many more MEMS opportunities in this market segment, consider that the mobile phone market crossed the 1 billion unit level in 2006. MEMS applications in phones include microphones/speakers (Knowles Acoustics), accelerometers (Analog Devices) for functions such as scrolling, gyroscopes (Invensense) for in-phone camera-lens stabilization, system timing products/oscillators (SiTime for myriad functions), and RF MEMS for band switching, filters (TeraVicta), and displays (Qualcom/Iridigm). The challenge for mobile phone MEMS manufacturers will be to meet the stringent price constraints demanded by the phone companies while still maintaining necessary reliability and product performance.

Read-Write Heads, Ink-Jet Heads, Microdisplays, Optical Switches, and Accelerometers

Read-Write Heads

Read-write (RW) heads still represented around 51% of the MEMS market in 2009: although the traditional application of RW heads in PCs will grow only moderately, the RW market is experiencing a renaissance in consumer electronics, as hard discs are entering music players (e.g., in many iPod models) and smart phones (Samsung introduced the first cell phone with HDD in 2004), as well as digital video cameras, set-top players, and DVD recorders. Memory read/write technology is an area where MEMS will continue to contribute and astound with its results; an important example we encountered in this regard is the IBM thin film read write head and the millipede (see Volume II, Figure 2.4, and Figure 1.1d, this volume). Major U.S. players in this market segment are Seagate, Conner, IBM, Read-Rite, and Applied Magnetic, in Asia there are companies like Yamaha, TDK, Alps, and a few others. Global revenues from shipments of HDDs were expected to reach $27.7 billion in 2010, up 18.4% from $23.4 billion in 2009. This represents a 2010 total unit shipment forecast of 674.6 million compared to 549.5 million units in 2009 (http://www.emsnow.com/npps/story.cfm?pg=story&id=42187).

It is important to note though that MEMS will lose a very important market when hard disc drives are replaced by solid-state disc (SSD) memory (flash drives), as both RW heads and accelerometers will fall away.

Ink-Jet Heads

At a greater than 25% share, inkjet print heads represent the largest portion of the total MEMS component market. Ink-jet heads continue to be one of the most profitable markets for MEMS. Cartridges for ink-jet printers were introduced in the early 1980s by HP and Canon. In Volume II, Figure 5.30, we show inkjet-printer heads machined with excimer lasers. As shown in Volume II, Figure 8.21b, some of those printer heads are very sophisticated, combining fluidics (at high temperatures!) with integrated electronics, sensors, and actuators. There are two major types of ink-jet print heads. First are the less expensive and disposable bubble-jet printheads, such as the ones made by HP, Canon, Ollivetti, Lexmark, Xerox, etc. Then there are the more expensive piezo print heads, manufactured by companies like Epson, MIT (in Sweden), Brother (Japan), and Frankotyp-Postalia (Germany). The main new driver for ink-jet heads in the next couple of years is expected to be the printing of digital photos—especially from cell phone cameras—growing at 15% to 20% per year. An important new trend in desktop printers is the integration of nondisposable inkjet heads in the printer instead of disposable print heads in cartridges. After Epson and Canon, HP is now also starting to ship printers with integrated print head. Although this will slow down the growth in units, the value of these nondisposable print heads is 3 to 5 times higher.

As competition in the printer marketplace continues to intensify, manufacturers of MEMS inkjet heads must reduce fabrication costs while improving product performance. The vast majority of inkjet head production in 2008 was on 6-inch silicon wafers, but increasing inkjet heads sizes—and therefore die sizes—as well as a general trend toward larger wafer diameters were expected to shift almost 10% of production to 8-inch wafers by 2010. These MEMS devices require high-quality, high-speed slotting of silicon wafers that are diced into the barrier chips, through which ink passes from the reservoir into the nozzles. Currently implemented MEMS techniques for silicon wafer slotting have a number of drawbacks. Wet etching processes require a mask step, making them both slow and costly, while sandblasting creates holes with conical edges, limiting their diameter and density. Dry laser techniques, although somewhat more effective, typically generate debris and microcracks that hamper device quality and performance. Today, key MEMS technologies to fabricate inkjet-heads include increasing use of deep reactive ion etch (DRIE) and SU-8 resins, but new MEMS manufacturing techniques are being explored. The Synova Laser MicroJet (Volume II, Chapter 5, "Water Jet-Guided Laser Machining," Figure 5.23), for example, overcomes some of the machining challenges, ablating the silicon material quickly and efficiently

without causing any residual damage. In production environments, the Synova water-driven laser has proven superior to competitive approaches, reducing the cutting time to just 2.5 seconds per slot—faster than any previously implemented slotting technique—with no damage.

It is to be expected that the fluidics and their integration with electronics pioneered in this application field will produce rewarding spin-offs in other application areas such as drop delivery systems, micromachined PCR units, etc., for medical/biomedical applications.

MEMS Microdisplays

MEMS microdisplays are the new MEMS blockbuster, and their revenues were expected to overtake inkjet heads in 2009. Millions of DMDs™ were built into projectors (the first one came out in 1996 from InFocus) and big-screen televisions (the first came out in 2002 from Samsung). Microdisplays found their way from projectors and high-definition, large-screen TVs to microdisplays for mobile handsets (e.g., the iMoD—interferometric modulator display—developed at Qualcomm). It was projected that by 2009 MEMS displays would become the second largest MEMS market, after thin-film RW heads (according to the Nexus report).[1] A major advantage for Texas Instruments (TI) is that DLP chips are an exception to the high-volume/low-price rule, typically for all ICs and most MEMS devices: the 2004 asking price for a DLP chip was $400 in 2004; it was estimated at $190 in 2009 (Nexus).[1]

In an imaginative new application of TI's micromirror technology, NimbleGen Systems developed a set-up that enables researchers to fabricate their own DNA arrays without having to go through the expense of making large sets of expensive photomasks as in the case of the Affymetrix system (see Volume II, Figure 8.19 for a schematic of NimbleGen's virtual mask generation system).

Another elegant MEMS display technology, from Silicon Light Machines, Sunnyvale, California (now owned by Cypress Semiconductor), is the grating light valve, or GLV™ (http://www.siliconlight.com/htmlpgs/homeset/homeframeset.html). In this technology, invented at Stanford, closely spaced parallel rows of reflective ribbons are suspended over a substrate. The gap between the ribbons and substrate is one-quarter wavelength, and in the resting state, the ribbons appear as a continuous reflective surface. With an electrostatic voltage pulling down alternate rows of ribbons, light reflected from the pulled down ribbons travels twice the gap or 1/2 wavelength farther than if reflected from the resting ribbons. The ribbons are arranged such that each element is capable of either reflecting or diffracting light. This allows an array of elements, when appropriately addressed by control signals, to vary the level of light reflected off the surface of the chip. This control of light can be analog (variable control of light level) or digital (switching of light on or off). Because a GLV™ device is reflective, it is highly efficient in its use of light. Besides Silicon Light Machines and TI, Samsung and Sony are also manufacturers of digital mirror technology.

In May 2008, Qualcomm MEMS Technologies, demonstrated the first reflective interferometric modulation (IMOD) color Mirasol™ display for mobile phones. IMOD technology requires no backlighting and reflects light so that wavelengths interfere with each other, creating pure, vivid colors. Mobile devices such as mp3 players and mobile phones stand to benefit from Mirasol displays, which require significantly less power and harness ambient light sources to automatically scale for optimal viewing in virtually any lighting condition. The interferometric modulator element (IMOD) is a simple MEMS device that is composed of two conductive plates. One is a thin-film stack on a glass substrate, and the other is a reflective membrane suspended over the substrate. There is a gap between the two that is filled with air. The IMOD element has two stable states. When no voltage is applied, the plates are separated, and light hitting the substrate is reflected as shown in Figure 10.61. When a small voltage is applied, the plates are pulled together by electrostatic attraction and the light is absorbed, turning the element black. This is the fundamental building block from which Qualcomm Mirasol™ displays are made (Figure 10.61).

MEMS obviously figures central in various designs for next-generation flat panel displays. Illustrative is also the thin-film field emitter work from Capp Spindt (the author's first mentor in MEMS at SRI

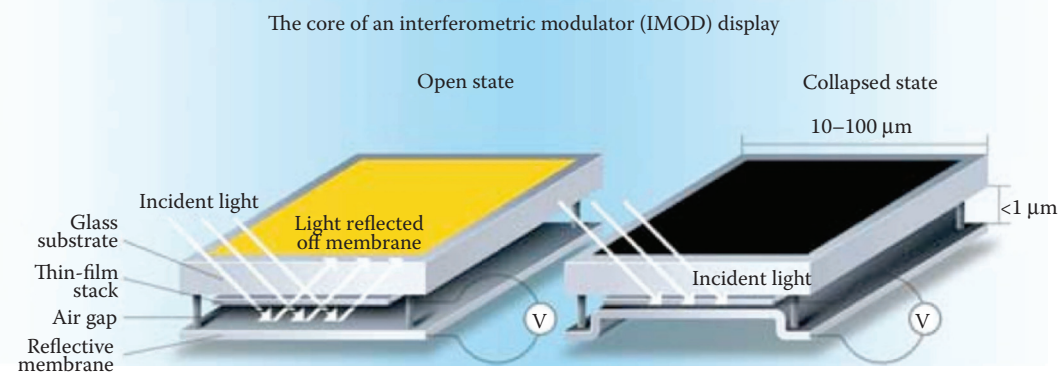


FIGURE 10.61 Interferometric modulation fundamental building block from which Qualcomm Mirasol™ displays are made (http://www.qualcomm.com/qmt/).

International), which is illustrated in Volume II, in Chapter 7, Figure 7.90.

Optical Switches

IC-based processing techniques have been used for a long time in the fabrication of complex optoelectronic ICs, and MEMS has been employed in the fabrication of simple fiber alignment devices. Passive alignment devices such as V-grooves are either fabricated in polymer or Si, with the polymer embodiments enabling high-volume manufacture of low-cost components. The newest developments in optical telecommunications require a new generation of additional components, including the following:

- Wavelength filters
- Micromirror-based all-optical switches
- Configurable add/drop multiplexers to add or drop signals to generate thousands of different frequencies of light
- Tunable filters
- Multiplexers and demultiplexers that direct multiple light beams onto one fiber
- Stabilized optical sources
- Optical polarization controllers
- Integrated optical pumps and MOEMS amplifiers for attenuation, dynamic gain equalization, and dispersion compensation

MEMS is expected to play a role in the manufacture of several of these optical devices. Here we reiterate briefly the role MEMS is playing in just one such application, i.e., large optical switches for telecommunications. As we remember from earlier in this chapter (see under Preamble on the MEMS Market), this application unfortunately led to the so-called *MOEMS bubble* in 2001–2002.

Miniature optical switches, optical cross-connects (OXCs), were expected to change how data are transmitted in everything from communication systems to computers. Miniaturization methods enable the creation of arrays of very small, high-capacity optical switches (see Example 4.4, "Self-Aligned Vertical Mirrors and V-Grooves for a Magnetic Micro-optical Matrix Switch"). These optical switches provide a way of routing light paths from one fiber to another without converting the signals to the electronic domain. OXCs seemed to have arrived at just the right time; optical fiber cables connecting network hubs were indeed quickly outpacing the capacity of electronic switches, and this electronic bottleneck could be solved by this new generation of MOEMS switches. Four example approaches to optical switching are illustrated in Figure 10.62: mechanical (bending of an optical fiber) (a), planar waveguides (where heaters and Peltier devices change the index of refraction to cause the beam path to shift) (b), moving mirror types (c), and the bubble approach (bubbles deflect the light between criss-crossing waveguides) (d).

The router developed at Lucent, for example, could switch 256 input light beams to 256 output fibers without any intermediate optical-electrical conversion (Figure 10.63). Supporting 256 wavelengths, each carrying 40 Gbit/second of traffic, this corresponds to an aggregate throughput of more than 10 Tbit/second. Tiny double-hinged mirrors are used to redirect optical signals by adjusting the mirror tilt.

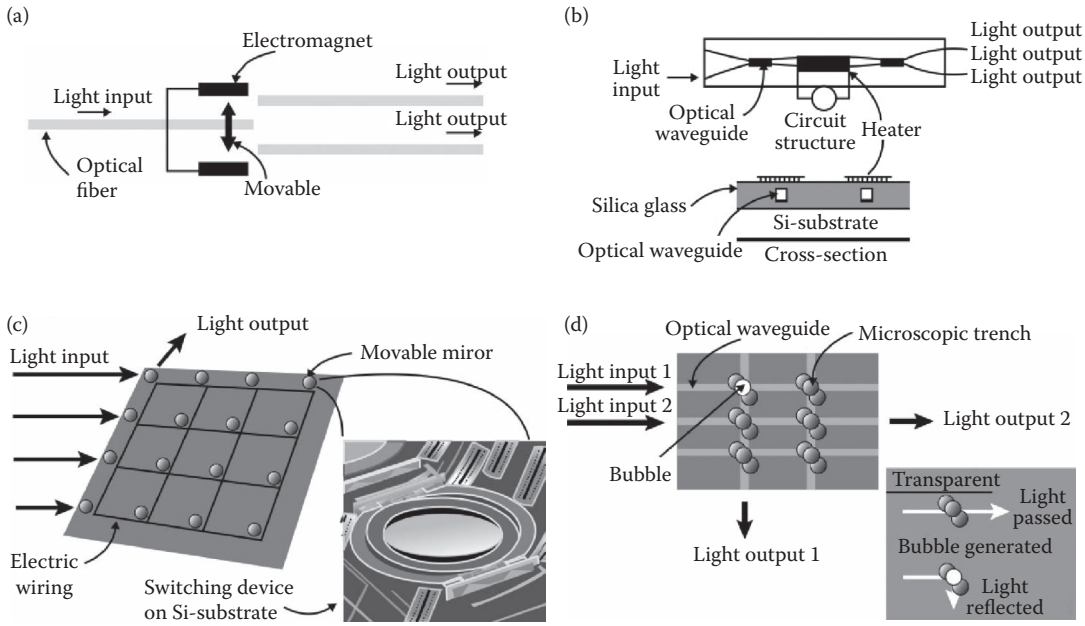


FIGURE 10.62 Example principles of optical switching devices. There are four major principles of operation involved. (a) The mechanical type, where the optical fiber is bent with an electromagnet to change the direction of the light beam. (b) The planar optical waveguide, where heaters and Peltier devices are located near the optical waveguide to change the waveguide temperature. When the temperature changes, so does the index of refraction, causing the beam path to shift. (c) The mirror type. The angles of micromirrors are changed to select different light paths. Mirrors may move on one axis (two-dimensional) or two (three-dimensional). (d) The bubble approach. An oil-filled wall is created along the light path, and bubbles are generated there to reflect the light beam.

A first application for these new switches was expected to be in dense wavelength division multiplexing (DWDM). To squeeze more traffic into a network, one of the strategies is to deploy DWDM. A programmable DWDM add/drop node "drops" one or more wavelengths out of a multiwavelength channel and reroutes them to a different destination while potentially "adding" traffic from another source to the channel.

One of the most impressive optical switches was a 1152 × 1152 OXC by Xros (bought by Nortel and now defunct). The product was projected to be on the market in 2001. Another player was Ilotron a spin-off company from the University of Essex. Most major players in MEMS optical switching were backing the traditional micromirror approach, but Agilent Technologies has a fluidic approach based on its ink-jet technology. In the Agilent switch, bubbles deflect the light between criss-crossing waveguides, and at Corning, liquid crystals are used. In Table 10.12, the four most popular optical switch approaches are compared.

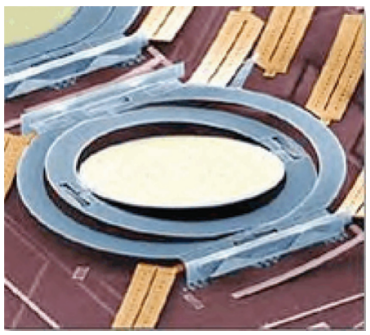


FIGURE 10.63 Lucent free-space microelectromechanical system optical cross-connect mirrors with two-axis, gimbal-mounted micromirrors.

It was expected that optical switches would achieve full commercialization by 2004. The great promise that these truly enabling MOEMS devices seem to offer was at first validated by the acquisition of a series of MOEMS companies. Among these were XROS (a 95-person company), acquired by Nortel Networks for $3.25 billion in stock; IntelliSense, 67% of which was acquired for $500 million by Corning; and Cronos Technologies, acquired for $750 million in stock by JDS Uniphase. However, as we explained in the beginning of this chapter, the more than 10 billion dollar investment by these large telecommunication companies in 1999–2000 became a textbook example of "irrational exuberance." There

Table 10.12 Major Characteristics of Optical Switching Devices

Type	Maximum Channels	Switching Time	Insertion Loss	Cross-Talk (Interference with Other Channels)	Device Dimensions	Manufacturers
Mechanical	1 × 2 to 2 × 4	10 milliseconds or less	0.3 dB	–65 dB or less	Two switches in 8.3 mm × 15.6 m × 28 mm	Hitachi Metals, Seiko Instruments, etc.
Planar waveguide	16 × 16	10 milliseconds or less	8 dB for an 8 × 8 switch	–50 dB or less	10 cm square	NTT, etc.
Mirror	256 × 256	About 50 milliseconds	Unknown	Unknown	Mirrors about 500 μm in diameter	Lucent Technologies, A&T, Japan Aviation Electronics, etc.
Bubble	32 × 32	10 milliseconds or less	0.07 dB per trench	–60 dB or less	2 cm square	NTT Electronics, Agilent Technologies, etc.

were not enough customers for the more than eighty start-ups, and the three largest investing companies pulled back. What was left of the all-optical switching market, in the period between 2002 and 2004, was littered with the broken shells of startups that often tried to survive by doing MEMS foundry work. Today, some surviving MOEMS companies, building all-optical switches, are successfully targeting more modest switch applications; say, for apartment buildings or university switchboards, and for new emerging applications such as semiconductor test equipment. Companies such as Glimmerglass claim to have solved the problems previously associated with the MOEMS switches and maintain that their optical switches are more reliable, less complicated, and more cost-effective. The Reflexion eighty-switch pack from Glimerglass is available as either a small core module tailored for communications-systems manufacturers or as a larger evaluation system. The evaluation system includes the Reflexion module, as well as optical-channel power monitors, a fiber patch panel, and an ethernet* control interface. It is all packed into a 4.5- by 9.5- by 7.0-inch rack-mountable case that consumes just 20 W (see Figure 10.1).

As we saw above, instead of switching signals among very large arrays of fiber optic cables for telecommunication switches, TI targeted its digital light processing (DLP) technology to projectors and video displays. Because it focused on the right mass consumer application from the get-go, TI's digital micromirror device (DMD) turned out to be a big-scale MEMS success (see above).

Today, even the large all-optical optical switches are back. Late-stage startup Lambda OpticalSystems (Reston, VA), for example, recently unveiled the LambdaNode 2000, an integrated optical-switching and DWDM transport platform. The system supports 256 wavelengths, each wavelength running at 10 Gbit/s for a total capacity of 2.5 Tbit/s. The device is running live traffic today in a 10-Gbit/s transcontinental network built by the U.S. Naval Research Laboratory. Also, Calient Networks announced details of an interoperability demonstration it conducted with Cisco Systems.

Accelerometers

Recently, the market for accelerometers in consumer electronics (CE) (computers, games, phones, etc.) has also started to explode. Consumer electronics have the potential to dwarf the automotive application, not only of accelerometers but also of pressure sensors and gyros. The first accelerometer application in consumer electronics was for hard drive protection in notebooks. IBM (now Lenovo) puts a dual-axis accelerometer in its notebook: the accelerometer detects falling motion and places the HDD in safe mode prior to impact. Samsung (and others) have experimented with motion-based applications in handsets (gaming, user interfaces, power management, navigation). Early in 2007, Apple announced the iPhone, with motion-based features.

* Ethernet: LAN (local area network) connects your computer to computers close by and around you.

Consequently, MEMS growth is expected by some to follow a hockey stick curve, with the value of products rising to $40 billion in 2015 and $200 billion in 2025.[99] (See also an estimate by Dr. Marinis from Draper Laboratory: http://www.sem.org/PDF/s01p01.pdf.) The problem is now one of profitability rather than the absence of real product opportunities: how to earn money on, say $0.32 microphones or $1.35 triaxis accelerometers for mobile handsets? Only by becoming more similar to the IC industry can MEMS devices achieve this goal, i.e., through more integration on larger (8-inch) Si wafers. In a sense, as far as mechanical and MOEMS are concerned, the IC industry has now swallowed the MEMS field.

RF-MEMS

Wireless communications is showing an explosive growth of emerging commercial and consumer applications of radiofrequency (RF), microwave, and millimeter-wave circuits and systems. Just think that the number of mobile phones was predicted to reach 4.6 billion in 2010 (http://www.cbsnews.com/stories/2010/02/15/business/main6209772.shtml). Several of the new and improved functionalities in mobile phones will be enabled by the use of MEMS products, such as microphones, lenses, and miniaturized data and energy storage devices. Analyst Bourne predicts that by 2010, three-quarters of all cell phones will benefit from MEMS (http://www.eetimes.com/electronics-news/4071288/MEMS-exec-sees-billion-dollar-markets) in 2009. This is making the consumer industry more and more the driving force behind many technology developments in MEMS. RF-MEMS will help further reduce the mobile phone size, power consumption, and cost. According to a market study from WTC covering the 2005–2009 period, the market for RF MEMS components was $126 million in 2004, confirming WTC's forecast from 2002. It was forecasted to grow to over $1.1 billion in 2009 (Figure 10.64).

In this application area, a situation has been reached where the presence of the expensive, off-chip passive RF components, such as fixed capacitors and tunable capacitors (varicaps), high-Q inductors, RF switches and filters, plays a limiting role. MEMS promises to provide higher levels of integration at lower costs and reduce the power consumption. In RF-MEMS the compatibility of the MEMS process with active circuitry processing is critical, and the ideal MEMS solution must have little if any impact on the wafer/electrical yield of the active circuitry. This capability will enable component suppliers to offer increased performance designs while maintaining or reducing device cost. An additional benefit of the totally integrated approach is the inherent reliability of the IC-based processing. One approach is the innovative "above-IC" (AIC) technology that enables the placement of RF-MEMS devices directly on top of the IC by using a thick copper technology compatible with CMOS, BiCMOS and gallium arsenide processes.[100] Over time, integration of RF-MEMS devices may be the means to replace all passive RF chips with on-chip devices, achieving the long-awaited single-chip RF solution. Current technology and process limitations prevent replacement of all passive components with on-chip MEMS components. But placing even some components on-chip offers significant space and cost savings, allowing smaller form factors, benefiting cell phones for example, or added functionality such as internet connectivity.

Commercially available MEMS-based switches (e.g., from Raytheon), with low-loss metallic structures for lower insertion losses and higher linearity, achieve actuation by using either electrostatic or electromagnetic techniques. Infineon, for example, is aiming for electrostatic switches with 10- to 100-microsecond switching times. Electrostatic actuation consumes no current but requires higher actuation voltage (>10 V), whereas electromagnetic actuation requires a lower actuation voltage, but its current consumption is significantly higher (see Chapter 8, this volume). The use of magnetic thin films introduces slightly more processing complexity. Electrostatic switches offer the most promise as reconfiguration switches (e.g., antenna switches or frequency band selection switches), where the key factor is low power consumption.

MEMS manufacturers are also fabricating inductors using a thick copper-on-insulator process to achieve higher quality factors and resonant frequencies. Similarly, vendors are fabricating MEMS-based

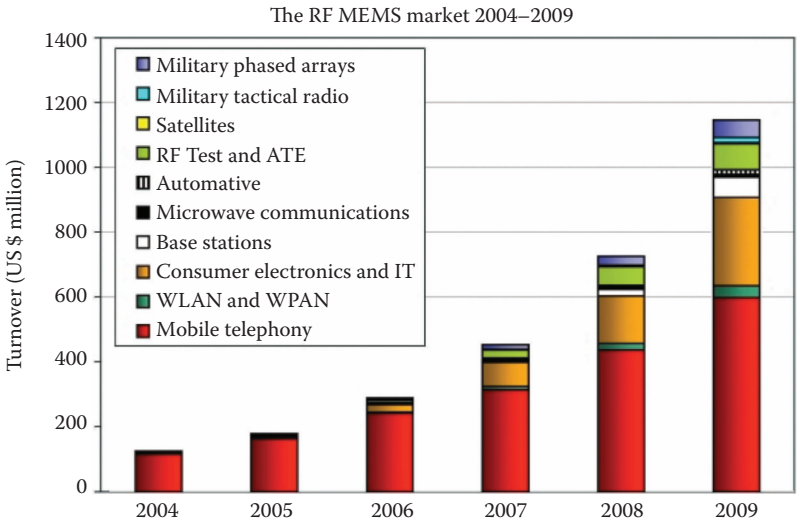


FIGURE 10.64 Radiofrequency microelectromechanical systems in wireless communication for the period 2004–2009 (WTC).

variable capacitors by using a thick copper-on-insulator process to minimize the ohmic losses. MEMS inductors and capacitors are typically suspended with a small air gap insulating them from the lossy silicon or GaAs substrates. The manufacturers achieve tunability of these components by using movable metallic structures with actuation processes similar to the ones used for switches. Typically, the quality factor of inductors using silicon technology have been less than 10 at 2 GHz, so far; this means that most applications have still required off-chip inductors to realize high-performance RF circuits. The objective remains to improve the quality of these devices and the tunability range of the circuits. MEMSCAP has designed high-quality inductors from thick copper on insulator; they range from 1.5 to 15 nH.[101] A quality factor of between 50 and 80 at 2 GHz was obtained. Variable capacitors from the same group offer a high tuning range of more than 2.5, with values ranging between 1 and 10 pF. The combination of inductor and capacitor capabilities optimized for low loss and flexible tunability range is important for the design of voltage-controlled oscillators (VCOs) and filters. Currently, foundries making on-chip RF-MEMS devices are aiming toward fabrication of all the RF-MEMS device types at the same time.

In Volume I, Chapter 4, we saw that for wireless networks such as Wi-Fi and Bluetooth, SiGe is becoming an important choice substrate (see also Chapter 5, this volume where we compare GaAs with SiGe). An example SiGe process is the 130-nm SiGe BiCMOS technology from Jazz Semiconductor. This process combines industry standard 130-nm CMOS with 200 GHz silicon-germanium (SiGe) heterojunction bipolar transistors (HBT) NPN transistors for high-performance RF and millimeter-wave ICs. This technology enables the design of the highest performance circuits in advanced high-speed optical, wireless, and millimeter-wave applications.

Currently, the bulk of the RF-MEMS market services mobile telephony applications, with bulk acoustic wave (BAW) (see Chapter 8, Figure 8.18) duplexers and filters supplied largely by Agilent and Infineon (this product was predicted to continue to dominate and constitute around 40% of the total market in 2009). The RF-MEMS market earned $330 million in 2006, with three companies shipping the newer FBAR (Figure 8.19) duplexers and filters: Avago, Infineon, and EPCOS.

Knowles Acoustics shipped 200 million surface-mount MEMS microphones (a two-chip solution in a single package) and confirmed its clear leadership in this sector, doubling its estimated revenues to $92 million for 2006. Knowles has been selling to cell phone makers for five years and is by far the biggest player in MEMS cell phones. MEMS Tech of Malaysia is a close second. Akustica, Sonion MEMS, and Infineon are other entrants in this area. Today, Akustica is the only maker of a single-chip MEMS microphone, and its rival Sonion MEMS A/S (Roskilde, Denmark) was set to begin volume shipments of MEMS microphones from its new

production facility in Ho Chi Minh City, Vietnam, in 2008. Akustica puts all the functions of a miniature electret condenser microphone—including a discrete FET, a separate operational preamplifier chip, and an A/D converter chip—onto a single CMOS MEMS microphone chip.

Another mass market emerged for consumer and IT applications in 2006–2007 with micromechanical resonators. Silicon MEMS start-up companies Discera, Silicon Clocks, and SiTime are currently trying to displace quartz from the oscillator market. They are building packaged silicon MEMS oscillators, consisting of a MEMS resonator and the electrical circuits for the oscillator function and for frequency and temperature compensation. Here, micromechanical resonators are used not as filters but rather as timing devices. MEMS oscillators sales started at the end of 2006 with the first million shipments by SiTime. Seiko Epson also started to ship its QMEMS oscillators. RF-MEMS switches were only in the sampling phase in 2006 (e.g., at Teravicta and Panasonic) but were jolting the market, with Teravicta shipping 10,000 units per month and plans to reach the million mark by the end of 2007. In Chapter 5, "Quartz Properties and Quartz Machining," we made a detailed comparison of Si oscillators with quartz oscillators. More than sixty companies are currently involved in RF MEMS manufacturing, and around a quarter of them were shipping commercial products or samples in 2005. Outside of mass markets, RF test and automated test equipment (ATE) offer the best opportunities, followed by military applications (see Figure 10.64).

MEMS Foundries

As we remarked earlier, the emerging MEMS consumer market is a strong driving force for the MEMS industry to develop MEMS devices integrated with ICs and is pushing the MEMS industry to even adapt 8-inch wafer size manufacturing. Capital requirements for a MEMS production line can be very, very large. And, as with semiconductors, the smaller the features, the more expensive the production line is. Many MEMS startups expended plenty of investor dollars wrestling with exotic materials and getting pilot production lines in place, exhausting their funds before their markets could develop. With this knowledge and the current MEMS upswing, even foundries are now becoming profitable. Along this line, Bosch, a MEMS pioneering company, founded a subsidiary named Bosch Sensortec in early 2005. This subsidiary is dedicated to engineering, developing, and marketing micromechanical sensors with foundry services for consumer electronics, medical technology, security systems, and logistics. The firm is backed by Bosch automotive MEMS development and production and owns a 4000-m^2 CMOS fab and a 4200-m^2 MEMS back-end fab. A new 8-inch fab was planned for 2009. The need to adhere closer to IC manufacturing rules has become even more urgent as few U.S. MEMS foundries have survived. As mentioned earlier, the MEMS foundries that did survive the MOEMS bubble can expect to be doing good business if they can scale up to compete with traditional IC fabs fast enough. In the author's own experience, MEMS foundries are hard to make profitable in the United States, as they need to keep on accepting a wide variety of jobs to keep the expensive clean-room operation going. In the process they do not stay long enough on the learning curve of one new product to become world leaders for that particular product. As a consequence, these labs are moving more and more to Asia, where a large company might have more patience to keep the doors open till the MEMS foundry turns profitable. Olympus, in Japan, for example, already runs the largest MEMS foundry in the world.[102]

Nanotechnology Markets

Introduction

As is clear from the preceding section, although not as large a market as projected, MEMS has made a successful transition out of the laboratory into the marketplace. If one excludes the contributions to nanotechnology markets from genetic engineering (e.g., genetically engineered proteins and DNA probes) and the latest ICs (e.g., chips with sub-100-nm critical dimensions that started shipping in 2003) and next-generation lithography equipment (e.g., extreme UV lithography exposure stations), nanotechnology (NEMS) is still in its infancy, and when a new field emerges it is often the tool-makers

and the conference organizers/consultants who make the most money. In his entertaining treatise on Nanotech Pioneers, Steven Edwards[103] compares this to the California gold rush, when the hardware stores and whorehouses made more money than the miners themselves. Nanotechnology quickly became a buzzword and a catchall term to attract funding, and as in the case of MEMS, many existing activities have been relabeled to raise their stock market value.

Whereas in MEMS, products/markets do exist (Bosch makes >100 million MEMS devices a year), in NEMS we deal mostly with manufacturing gear (e.g., carbon-nanotube chemical vapor deposition, dip-pen lithography, nanoimprint lithography, etc.), metrology equipment (Veeco's SPM, Zyvex Nanomanipulators, etc.), software (e.g., Atomostix, MEMulator™, etc.), and products based on nanomaterials (e.g., nano skin cream, suntan lotions, nano-enhanced tennis balls that bounce longer, etc.). Some market studies do also count nanodevices and next-generation lithography equipment in their market estimates. We advice against this practice, as co-opting the IC industry in nanotechnology market projections is bound to be misleading. The emphasis should be put on innovative nanofabrication methodologies making new nanomaterials and nanoproducts and not co-opting efforts that have been going on for a long time. Some will argue that, today at least, besides nanotools research tools, this leaves only low-tech nanomaterials-based products as a new realization. These are products with little technology content and scant manufacturing innovation, and cynics would say that a new transparent sunscreen hardly justifies the current level of investment in nanotechnology.

As with any exciting topic, there is a lot of speculation and exaggerated estimates about the impact nanotechnology will make over time. This author does believe that the ultimate impact of nanotechnology will be broader and deeper than the IC revolution ever was. This is because of the potential of nanotechnology to apply material innovations to so many different situations and challenges. Nanotechnology will profoundly change major industries such as electronics, medicine, transportation, defense, and consumer goods through the development of advanced materials that are lighter, stronger, and more versatile than existing materials. The current nanomaterials boom will be followed by a nanodevices upsurge—it just might be much slower than current market studies suggest.

Overall Market and Market Character

The National Science Foundation (NSF) projects that nanotech will be a $1 trillion industry by 2015; the accuracy of this claim is difficult to assess given the doubts expressed above about what is included in this estimate. Many policy makers have bought into this projection, though; governments across the world poured $4 billion into nanotech research in 2004, and in 2006, the worldwide investment in nanotech research and development was said to be $12.4 billion. There is even a stock index of public companies working on nanotechnology (launched by Lux Research in 2005).

According to a technical market research report, "Nanotechnology: A Realistic Market Assessment" from BCC Research (http://www.bccresearch.com), the global market for nanotechnology was worth $11.6 billion in 2007. This was expected to increase to $12.7 billion in 2008 and $27.0 billion by the end of 2013, with a compound annual growth rate (CAGR) of $16.3% (see Figure 10.65) This makes the NSF market projection for 2015 ($1 trillion) seem unrealistically high. However, as a counterpoint, Lux Research estimates that by 2014, $2.6 trillion in manufactured goods will incorporate nanotechnology—or about 15% of the total global

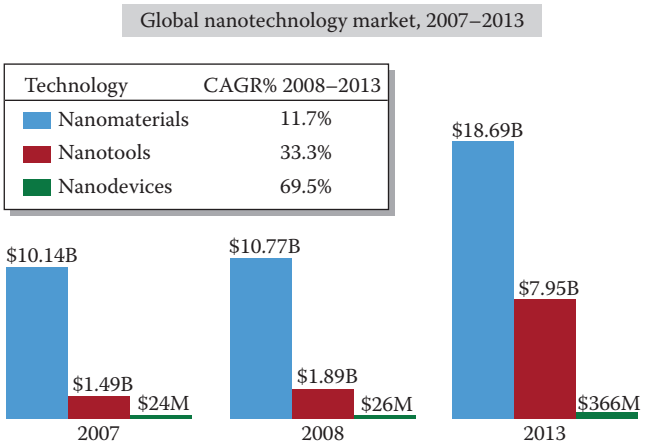


FIGURE 10.65 Global nanotechnology market, 2007–2013. (From BCC Research.)

output will be nanotechnology based! It is hard to figure out what this Lux Research number really means though in terms of the real nanotechnology market size (e.g., TiO_2 and ZnO nanopowders are cheap and would not really constitute a big market share or put many people to work). We actually believe that even the BCC Research numbers are too high. This research firm breaks down the market into nanomaterials, nanotools, and nanodevices. As stated above we are uncertain about the merit of including nanodevices in a nanotechnology market study as today the only nanodevices on the market are the next-generation ICs. These are based on the traditional top-down manufacturing methodologies that were ongoing long before and independent of the current nanofocus. More critical, in terms of coming up with acceptable numbers, is the unfortunate inclusion of current next-generation lithography (NGL) equipment in a market study that should be projecting the impact that new nanotechnologies are bound to make. By including NGL equipment one runs the risk of making projected numbers a self-fulfilling prophecy by broadening the definition of what passes for nanotechnology (this is what happened in the MEMS field). Such work will also not provide sufficient information in order to guide corporate or individual investment decisions. Investors require additional data, such as the size of specific nanotechnology markets, novelty of the manufacturing method (e.g., is it evolutionary or revolutionary), prices, and competition, as well as potential regulation. Nanotechnology market reports should point out specifically whether one is just dealing with an evolutionary extension of IC fabrication or with a revolutionary new nanofabrication method (there are plenty of sophisticated market reports covering the future IC market separately).

Of the three BCC Research nanomarket categories, nanomaterials dominated in 2007, accounting for 87% of the market. Worth an estimated $10.8 billion in 2008, this segment was projected to reach $18.7 billion in 2013, for a CAGR of 11.7%. Nanotools accounted for 12.8% of the market in 2007. Worth an estimated $1.9 billion in 2008, the segment was predicted to reach almost $8.0 billion by the end of 2013, for a CAGR of 33.3%. Nanotools, in which BBC Research includes the nanolithographic tools used to produce the next generation of semiconductors, are projected to grow at a much faster rate than nanomaterials. As a result, that market share was predicted to increase to 29.4% in 2013, while the nanomaterials' share would fall to 69.2% of the total market.

The nanodevices segment, according to BCC Research, will enjoy a CAGR of 69.5% between 2008 and 2013. It was expected to increase from a $26.2 million segment in 2008 to $366.2 million by the end of 2013.

The largest end-user markets for nanotechnology in 2007, according to BCC Research, were environmental remediation (56% of the total market), electronics (20.8%), and energy (14.1%). Electronics, biomedical and consumer applications have much higher projected growth rates than other applications over the next five years (i.e., 30.3%, 56.2%, and 45.9%, respectively.) In contrast, energy applications are projected to grow at a CAGR of only 12.6%, and environmental applications should actually decline by an average of 1.5% per year. It is not clear to this author why this should be the case. Depending on government policies on green technologies, the real numbers might again be quite different.

While it appears inevitable that nanotechnology will have a broad and fundamental impact on many sectors of the world economy, various technical, marketing, and other hurdles need to be overcome before nanotechnology fulfills this promise. These challenges and differences of opinion regarding the speed of the commercial implementation of new nanoproducts are reflected in the wide divergence among existing estimates of the U.S. and global nanotechnology markets. Global market estimates for 2004–2005, for example, range from $8 billion to $68 billion.

Polls show that most Americans know little or nothing about nanotechnology, despite the fact that, according to some estimates, nanotechnology was already incorporated into more than $50 billion in manufactured goods in 2006. Public perceptions about risks—real and perceived—can have large economic consequences. How consumers respond to these early products—in food, electronics, health care, clothing, and cars—is a litmus test

for broader market acceptance of nanotechnologies in the future. Along with high expectations, nanotechnology also evokes fear in some people. Bill Joy, cofounder of Sun Microsystems, published an article in 2000 in which he called nanotechnology (along with biotechnology and robotics) a "Faustian bargain" and a "Pandora's box," and called on mankind to relinquish all three technologies. In 2002, Michael Crichton's novel *Prey*, a fictional treatment of the dangers of nanotechnology, made the *New York Times* bestseller list. Interestingly, a recent nationwide survey conducted by the University of Wisconsin–Madison indicates that scientists are more concerned than the general public about the potential health and environmental risks of nanotechnology. Over 30% of scientists interviewed expressed concern that nanotechnology may pose risks to human health, whereas only 20% of ordinary people shared their fears (because they do not understand the technology?).

An important event that helps us appreciate the nanotechnology market dilemma today is what happened to Nanosys in 2004. On April 23 of 2004, Nanosys Inc., a Palo Alto-based developer of inorganic semiconductor nanocrystals, filed for an initial public offering (IPO). Merrill Lynch, Lehman Brothers, CIBC World Markets, and Needham & Co. were underwriting the offering. The company was seeking to raise up to $115 million on the Nasdaq market. But on August 4, 2004, Nanosys Inc. withdrew its IPO. "Based on adverse market conditions," the company explained "Nanosys has determined that it is not advisable at this time to proceed with the proposed offering." Press reports leading up to the IPO had lambasted Nanosys for having only $3 million in revenues in 2003. It was also pointed out that one of the few hard metrics available to value the firm—a price-to-revenue ratio over 120—suggested it was a speculative investment, and that if it performed poorly, it would undermine the ability of other nanotechnology firms to access the public markets. One must also realize that in 2004, the MEMS technology bubbles of 2001–2002 were still vividly remembered.

Nanosys later did raise approximately $40 million in a private equity financing. This financing was led by El Dorado Ventures and includes new investors Masters Capital, Medtronic, Inc., Wasatch Advisors, and others. In addition, there was strong participation from existing investors, including Alexandria Equities, ARCH Venture Partners, CDIB BioScience Ventures, CW Group, Harris & Harris Group, Inc., In-Q-Tel, Intel Capital, H. B. Fuller Company, Lux Capital, Polaris Venture Partners, Prospect Venture Partners, UOB Hermes Asia Technology Fund, and Venrock Associates. Nanosys is using this funding for the ongoing development and manufacturing scale-up of products that incorporate its proprietary, inorganic nanostructures with integrated functionality for multiple industries. Current product development programs include chemical analysis chips for pharmaceutical drug research, fuel cells for portable electronics, nanostructures for displays and phased array antennas, nonvolatile memory for electronic devices, and solid-state lighting products.

The type and number of California-based MEMS and NEMS start-up companies (see Table 10.13) has historically played an important role in what happens in these areas worldwide. In 2000, according to Janusz Bryzeck, there were 100 MEMS/NEMS-based companies in California alone, 33 of which were fabricating MOEMS devices. Decimated by the MOEMS bubble of 2001, "only" sixty-five California MEMS/NEMS companies survived in 2005. In the early years (say, from 1972 on), companies making mechanical MEMS devices (using wet bulk micromachining) played a pivotal role in the development of the market for these mechanical Si sensor products. Petersen's 1982 paper extolling the excellent mechanical properties of single-crystalline silicon helped galvanize academia's involvement in Si micromachining in a major way.[104] Before that time, timid efforts had played out in industry, and practical needs were driving the technology (market pull). Academics often explored micromachines that constituted gadgetry only, and, as a consequence, the field was perceived by some as a technology looking for applications (technology push). Polycrystalline surface micromachining-based (see Volume II, Chapter 7) MEMS companies started emerging around 1991. However, from the mid-1990s on, many of the MEMS companies became bio-MEMS oriented (application of MEMS to biotechnology), and single-crystalline and polycrystalline Si were often not the

TABLE 10.13 Some California Microelectromechanical System/Nanoelectromechanical System Companies

Company	Location	Date	Overview of Activity and Manufacturing Method
Sensym	Milpitas	1972	Made sensors from silicon and piezoresistive materials. Some cavities in products made by anisotropic etching. *Source:* U.S. Patents.
Endevco	San Juan Capistrano	1975	Produces sensors for vibration, shock, inertial motion, and dynamic pressure measurement using semiconductor technology and piezoceramic processing techniques. *Source:* Endevco.
Irvine Sensors Co.	Costa Mesa	1980	Manufactures some products in a batch process. Products are layered silicon and connected with gold. *Source:* Irvine Sensors Co.
Komag	San Jose	1983	Komag sputters thin-film disks using a vacuum deposition process. During the sputtering process, microscopic magnetic layers are successively deposited on the disk—these layers ultimately provide the magnetic storage capacity for data. *Source:* Komag.
Nanostructures	Santa Clara	1988	Produced thin films such as silicon membranes by electrochemical means or mask production with RIE. *Source:* U.S. Patents.
Abaxis	Sunnyvale	1989	Produces point-of-care blood analyzers from a bottom- and mid-layer of polymethymethacrylate plastic with a top layer of ABS made in a molding process. Some products made from joined thermoplastics through melting. *Source:* U.S. Patents, Abaxis.
Sentir	Santa Clara	1991	Produced sensors, including a pressure sensor made on a ceramic substrate with piezoresistive diaphragm and gel as the transfer medium to transducer. *Source:* U.S. Patents.
Affymetrix	Santa Clara	1992	The company's microarrays were made by a process that begins by coating a 5" by 5" quartz wafer. *Source:* Affymetrix.
Unisun	Newbury Park	1993	Unisun's techniques provide substantial cost advantages to traditional PV products based on silicon wafers and to vacuum deposition techniques typically used to deposit PV thin films. *Source:* Unisun.
Nanogen	San Diego	1993	Makes microarray cartridges the only Si based DNA arrays ever made. *Source:* Nanogen patents.
Front Edge Technology	Baldwin Park	1994	FET develops, manufactures, and markets next-generation, ultra-thin rechargeable batteries for card-type applications. *Source:* Front Edge Technology.
Aclara	Mountain View	1995	Some products made by thermally bonding plastics. *Source:* U.S. Patents.

(Continued)

TABLE 10.13 Some California Microelectromechanical System/Nanoelectromechanical System Companies (Continued)

Company	Location	Date	Overview of Activity and Manufacturing Method
Nanochip	Fremont	1996	Nanochip is developing the next generation of removable, portable memory chips for consumer applications. Using MEMs technology along with arrays of atomic force probe tips, Nanochip is developing storage chips that will store tens of gigabytes per chip. Our technology is not bound by lithography constraints. *Source:* Nanochip.
Caliper	Mountain View	1996	Microfluidic lab-on-a-chip made of quartz, glass, or plastic-bonded to each other. *Source:* Caliper.
Cepheid	Sunnyvale	1996	Made sample preparation products with the process of electrodeposition, LPVD, glass bonding, and RIE out of materials from the semiconductor industry. *Source:* U.S. Patents.
Fluidigm	South San Francisco	1997	Fluidgm's NanoFlex valve consists of a membrane that deflects under pressure to pinch off the flow of fluids in a microchannel. The valve is made from two separate layers of elastomeric rubber. *Source:* Fluidigm.
Carbon Solutions	Riverside	1998	Manufactures single-walled carbon nanotubes (SWNTs) in bulk quantities. *Source:* Carbon Solutions.
Zyomex	Hayward	1998	Products included fluidic technologies using an etching process. *Source:* U.S. Patents.
PolyFuel	Mountain View	1999	PolyFuel's membranes are based upon hydrocarbon polymers rather than fluorocarbon. *Source:* PolyFuel.
GeneFluidics	Monterey Park	2000	GeneFluidics disposable cartridges and sensor chips are injection-molded, using only plastic substrates and other low-cost materials; the GeneFluidics platform will dramatically reduce the cost of molecular analysis.
Molecular Nanosystems, Inc.	Palo Alto	2001	The company synthesizes carbon nanotubes on predefined locations with the desired orientation on different substrates, manipulates the properties of carbon nanotubes via functionalization, and integrates them into next-generation gas and biological sensors and other high-end electronics devices. *Source:* Nanovip.
Discera	San Jose	2001	Discera offers PureSilicon resonators, which is displacing quartz crystal solutions with systems-on-a-chip alternatives. *Source:* Discera.

Ahwahnee Inc.	San Jose	2002	Ahwahnee's industry-leading CNTs are revolutionizing a number of markets, including fuel cells, energy, and semiconductors. *Source:* Ahwahnee.
Apex Nanomaterials	San Diego	2003	Apex Nanomaterials's world first-class SWNT production technologies using both arcing and CVD methods are generating SWNTs with the best quality and most cost-effective at an industry scale. *Source:* Apex Nanomaterials.
Atomate	Santa Barbara	2003	Atomate Corporation develops complete systems, components, and materials that are optimized for the synthesis of nanostructures using CVD.
NanoSensors, Inc.	Santa Clara	2003	Develops and markets sensors and instruments to detect explosive (X), chemical (C), and biological (B) agents ("XCB") *Source:* Nanosensors.
InvenSense	Santa Clara	2003	InvenSense's gyroscopes takes advantage of major innovations in MEMS mechanical design and fabrication, mixed-signal ASIC design, and wafer-level packaging, which is based on the IC industry. *Source:* Invensense.
Intelligent Optical Systems, Inc.	Torrance	2004	Develops cutting-edge technologies in optical sensing and instrumentation. Current research and development is being conducted in various areas, including Nanoprobes and Materials for Sensing. *Source:* Nanovip.
Unidym, Inc.	Menlo Park	2006	Unidym's initial product is a transparent, conductive film of CNTs that replaces the approximately $1B of expensive and brittle metal oxide films currently employed in touch screens, flat-panel displays, OLEDs, and thin-film solar cells. *Source:* Unidym.
Biodent	Menlo Park	2006	Biodent is the first company to combine printed optoelectronic components with microfluidic systems. This combination is enabled by the fluid processability of organic semiconductor materials. Biodent allows for the first time to print and integrate full optical readout systems directly onto microfluidic devices. *Source:* Biodent.

preferred MEMS substrates for these applications. Later yet, from 1997 on, nanotechnology companies, mostly working on biotechnology type applications (bio-NEMS), started emerging, and here Si was often not involved at all.

For further reading on the history of MEMS and NEMS up to today see Volume I, Chapter 1.

Nanotechnology Tools

To estimate the size and to forecast the market for nanotechnology tools over the next five years, nanotechnology analysis firm Lux Research interviewed CEO's and marketing executives at 21 leading tool vendors worldwide and research and development leaders at forty-nine companies, universities, and national laboratories active in nanotechnology. They correctly excluded established uses of manufacturing and metrology tools in fields such as semiconductors and data storage. This firm believes that the market for tools used to fabricate emerging nanotechnology could grow from less than $20 million in 2004 to nearly $235 million by 2010. At the same time, they claim, the market for metrology tools to inspect matter on the nanometer scale, which currently dominates the $580 million nanotechnology tools market, is expected to see only tepid growth, although still command more than $750 million in 2010. These numbers are lower, and more realistic, than the BCC Research findings (see above) as they exclude the traditional IC tools.

Nanotechnologists basically use three kinds of tools to experiment with matter on the nanometer scale. First, metrology tools, reviewed in Chapter 6, this volume, such as atomic force microscopes (AFMs) and scanning electron microscopes (SEMs) visualize the nanometer-sized objects and represent the most mature and widely used category of tools for emerging nanotechnology. Second, fabrication tools such as nano-imprint lithography and dip-pen nanolithography create structures at the nanometer scale. Thirdly, modeling tools predict nanostructure properties to circumvent the costs and time associated with experiments. Lux Research predicts that the nanotechnology-tools market will see a CAGR of 11% to nearly $1.1 billion in 2010. However, growth will not be even across each of the three market segments. The current market for nanotechnology tools is dominated by inspection tools, which accounted for 95% of the 2004 revenues. The inspection-tools market saw dramatic growth during the early 2000s nanotech explosion, when many university nano centers were being built and equipped. However, the growth outlook for inspection tools looks slow because the university and lab markets are now saturated.

Nanofabrication tools have not yet found commercial use. Still, Lux Research expects these instruments to gain traction over the next five years and move beyond research and development labs to manufacturing floors. In particular, Lux Research sees nano-imprint lithography as the favorite candidate. Lux Research felt dip-pen nanolithography facing sharp resistance, as in many cases researchers would rather have a kit to modify an AFM they already own than purchasing a whole new instrument just to do dip-pen. No wide use for dip-pen nanolithography in production is anticipated before 2010.

Modeling tools claimed less than $10 million in 2004 in the face of entrenched skepticism from potential customers. In five years, Lux Research expects that market to expand to only roughly $40 million, a small fraction of the total nanotechnology-tools market. Still, in a decade the analysts predicted it could break $100 million as power, accuracy, and ease of use of modeling tools continues to improve and more customer success stories emerge.

Nanomaterials

Even though much of the research is still in the initial stages, nanoproducts based on nanomaterials have entered the commercial market. According to an update to the nanotechnology consumer product inventory maintained by the Project on Emerging Nanotechnologies (PEN) (*ScienceDaily*, April 25, 2008), new nanotechnology consumer products are coming on the market at the rate of three or four per week. The number of consumer products using nanotechnology has grown from 212 to 609 since PEN launched the world's first online inventory of manufacturer-identified nanotech goods in March 2006. Health and fitness items, which includes cosmetics and sunscreens, represent 60% of these products. There are thirty-five automotive products in the

PEN inventory, including the Hummer H2. General Motors Corporation bills the H2 as having a cargo bed that "uses about seven pounds of molded in color nanocomposite parts for its trim, center bridge, sail panel and box rail protector." Nanoscale silver is the most cited nanomaterial used. It is found in 143 products, or over 20% of the inventory. Carbon, including carbon nanotubes and fullerenes, is the second highest nanoscale material cited. Other nanoscale materials explicitly referenced in products are zinc (including zinc oxide), titanium (including titanium dioxide), silica, and gold. A list of nanotechnology merchandise—containing everything from nanotech diamonds to cooking oil to golf clubs to iPhones—is available at http://www.nanotechproject.org/consumerproducts. One of the new items among the more than six hundred products now in the inventory is Swissdent Nanowhitening Toothpaste, with "calcium peroxides, in the form of nano-particles." PEN Project Director David Rejeski cites Ace Silver Plus—another of the nine nano toothpastes in the inventory—as an example of the upsurge in nanotechnology consumer products in stores.

There are more than five hundred nanotechnology companies globally, and one-third are in nanoparticles. Of that one-third, more than fifty-five are in carbon nanotubes. Of the present carbon nanotube companies, about twenty are into large-scale production. Some representative example companies in the last category include the following:

1. Nanoledge—(France). Large-scale production (4 ounces per day)—tennis racquets, fibers.
2. Mitsui CNRI (Akashima, Tokyo)—new nanotube plant (120 tons per year at 10% of the present cost).
3. Frontier Carbon (Kitakyusyu City)—Mitsubishi large-scale manufacturing (400 kg per year and 1500 tons in 2007).

Another group of carbon nanotubes companies sells modified carbon nanotubes (e.g., for sensor applications):

1. Carbon Nanotechnologies, the late Richard Smalley's company ($15 million—from angel investors)—single-walled and functionalized tubes.
2. Nantero (>$20 million—DFJ and other venture capital firms)—NRAM™, a high-density, nonvolatile random access memory chip.
3. Molecular Nanosystems Stanford ($2 million—band of angel investors)—nanotube sensors.
4. C Sixty ($4 million—CNI angel investors)—protease inhibitor, anti-HIV (merged with CNI).

Carbon nanotubes and fullerenes have very broad applications, and from the above it is seen that large manufacturing companies are stepping in, with the price of product dropping rapidly. There is also a clear trend of specializations developing (e.g., modified carbon nanotubes). There have not been any IPOs in this area yet. Applications of fullerenes include clothes (carbon nanotubes form waterproof tear-resistant cloth fibers), concrete (carbon fibers increase the tensile strength, and halt crack propagation), sports equipment (stronger and lighter tennis rackets, bike parts, golf balls, golf clubs, golf shaft, and baseball bats), ultrahigh-speed flywheels (the high strength/weight ratio enables very high speeds to be achieved), and bridges.

Other nanoparticles are used in catalysis, ceramics, dug delivery and functional materials, and coatings. Existing companies are Degussa, Cabot Corp, Johnson Matthey BASF, and Altana; start-up companies include the following:

1. NanoPhase Technologies (Nasdaq). Patents licensed from Argonne in 1990. This was the first Nanotech IPO (1998). It was overhyped and had to pay compensation of $4.8 (January 2002). Altana Chemie AG acquired a 7% share for $10 million in 2004. Products include ZnO nanoparticles for sunscreens and catalysts. Losing $1.3 million per quarter.
2. Ntera ($30 million)—was lithium-ion batteries, now nanochromic displays.
3. Nanogram ($6.7 million)—nanomaterials.
4. Aveka (buyout from 3M)—particle processing.
5. Oxonica ($4 million)—sunscreens, bio tags, catalysts. Floated on Aim 2005 (overdiluted).
6. Quantum Dot Corp ($37 million)—bio-assays. Acquired by Invitrogen in October 2005.
7. Nanosphere ($10 million). Gold nanoparticle DNA probes for biothreat detection (anthrax).

8. Qinetiq—nanomaterials.
9. Nanomagnetics—was data storage, now water purification, Apaclara.
10. Nanosys—the hyped nanotech IPO (oversold; see above). A Larry Bock company. Sells as much the technology as brand name scientists (Charles Lieber, Hongkun Park, Paul Alivisatos, and many others over the years). Products include biosensors (functionalized inorganic semiconductor nanowires, 1D electron flow; highly sensitive and selective), photovoltaics, nanocrystals in host matrix, solar cells as roof tiles (Matsushita contract), macroelectronics; flexible, low-cost substrate, nanostructured surfaces, hydrophobic and antimicrobial surfaces.

A worthwhile book to consult is *Nanotechnology Applications and Markets* by L. Gasman.[105]

Example 10.2: Cosmetics

The market for cosmetics using nanotechnology is expected to grow significantly over the coming years despite continued concern about the safety of the technology. Nanotechnology is being used increasingly in the personal care industry to develop sunscreens and antiaging formulations. The global market for cosmetics using nanotechnology was valued at $62 million as of 2007 and is forecast to grow annually by 16.6%, reaching $155.8 million by 2012, according to a BCC Research report (NANO17D). This market research company listed sunscreens as one of three applications in the biomedical, pharmaceutical, and cosmetics industries that are expected to account for 95% of the market in 2012. Nanoparticles of titanium dioxide and zinc dioxide are used in sunscreens to increase UV light absorption and make formulations transparent—ultimately making them more appealing to the consumer.

Example 10.3: Textiles

Nanotechnologies are providing incremental improvements to existing textile sectors, but they offer the highest growth in the nontraditional sectors. In the near term, the opportunity for nanotechnology in the textiles industry is in product innovation, not process innovation. Nanotechnology is more likely to be used to produce new materials or enhance the properties of existing materials than to reduce the production cost or improve quality. A number of nanotechnology innovations are already commercially available. These include:

- Stain-, wrinkle-, and liquid-resistant fabrics (see NanoTex)
- Clothing that can absorb body odors
- Clothing that emits deodorant by slow release
- Clothing that changes color with change in light
- Clothing that changes color with external or body heat

There are a number of opportunities that will be explored as nanotechnology develops further:

- New blended fabrics for specific applications (e.g., sportswear, mountainwear, military applications), including the incorporation of carbon nanotubes (CNTs) into fabrics. It is expected that composites with CNTs or interweaving with extruded CNT fibers will introduce higher conductivity as well as high strength.
- Property enhancement or alterations (UV blocking, durability, breathability, flexibility, recyclability, color retention, self-repair etc.); included in this is the introduction of electronic properties into fabrics by treatment with inherently conducting polymers. These provide capabilities such as sensing (chemical and mechanical), energy generation (photovoltaics), energy storage (batteries, supercapacitors), and charge dissipation (antistatic).
- In addition, controlled-release polymers may replenish and/or trigger release of antifungal, surface finish, or medical growth aids to the polymer surface. The triggered release systems may be made to be responsive to stimuli such as changes in temperature, humidity pH, and/or dissolved oxygen.
- Development of specific aesthetic properties (e.g., glow in the dark, color change with angle of light, color change with applied electric field).

The global market for nanotechnology in the textiles industry was around $480 million in 2007. The forecast is that nanobased products and processes will be worth $4.9 billion to the textiles industry by 2015. It is estimated that 24.6% of all textiles products available in 2015 will incorporate some form of nanotechnology, with the greatest level of penetration in the hygiene area, which covers numerous markets from household care to leisure and consumer goods to health care. Nanotechnology-based applications such self-cleaning textiles as will start to make a major impact in 2011. There will likely be a significant revenue growth in nanotechnology products in the sector from 2011 onward.

Questions

10.1: Suggest an array of sensors that could be used for an electronic tongue (5 tastes).

10.2: Why do you think that chemical microsensors have not enjoyed the same commercial success as physical microsensors? How might these deficiencies be overcome in the future?

10.3: Instrumentation miniaturization is a fast developing field. The Agilent 2100 Bioanalyzer fabricated by Agilent & Caliper Technologies Co. has been awarded the Pittcon Editors' Gold Award for 2000. How does this instrument benefit from miniaturization? What are its drawbacks? In miniaturized GC, MS, and IMS, what components have been micromachined?

10.4: At Pittcon 2000, Cyrano showcased a portable electronic nose, the Cyranose 320, which distinguishes odors using an array of 32 sensors. Each sensor is made of a composite of conductive carbon black and a non-conductive polymer. Describe how the sensors work and explain how one "trains" an electronic nose to qualitatively and quantitatively analyze unknown samples.

10.5: *Nanotechnology* is currently the buzzword in the scientific community. What does the term mean to you? Provide examples of nanoscale building blocks. What are the characterization tools (metrology) appropriate for this technology? What strategy would you consider for assembling a nanodevice and why?

10.6: Over the last five years mechanical MEMS applications saw a major change in market acceptance. In what applications and why did this occur?

10.7: The United States seems ahead in lithography-based MEMS and NEMS but is perhaps behind in non-lithography-based MEMS. Explain.

10.8: What are the three market components of NEMS today? How do they compare in size?

10.9: What are the largest MEMS applications in the automotive industry?

10.10: What was behind the infamous optical MEMS bubble bust?

10.11: Explain the concept behind the 1st level packaged MEMS and smallest commercialized unit.

10.12: How does the MEMS market compare in size to the IC market?

10.13: What is the difference between an angular rate sensor, a gyro, and an accelerometer?

10.14: Compare manufacturing options for Si-based pressure sensors and accelerometers. Give pros and cons.

10.15: Give a list of reasons why the ISFET did not become a commercial success.

References

1. NEXUS. 2006. *NEXUS market analysis for MEMS and microsystems III, 2005–2009*. Neuchâtel, Switzerland: NEXUS.
2. Bryzek, J., S. Roundy, B. Bircumshaw, C. Chung, K. Castellino, J. R. Stetter, and M. Vestel. 2006. Marvelous MEMS: Advanced IC sensors and microstructures for high volume applications. *IEEE Circuits Devices Mag* 22:8–28.
3. Ehmann, K. F., D. Bourell, M. I. Culpepper, T. J. Hodgson, T. Kurfess, M. Madou, K. Rajurkar, and R. DeVor. 2007. *International assessment of research and development in micromanufacturing*. New York: Springer.
4. Maluf, N. 2000. *An introduction to microelectromechanical systems engineering*. Boston: Artech House, Boston.
5. Wechsung, R., N. Ünal, J. C. Eloy, and H. Wicht. 1998. Market analysis for micro systems 1996–2002. NEXUS Task Force Report. Berlin, p. 136.
6. Fluitman, J. H. J. 1994. Basic technologies and functions; microsystem technology and silicon micromachining. In *Microsystem technology: Exploring opportunities*. The Hague, The Netherlands: Stichting Toekomst der Techniek, pp. 35–40.

7. National Research Council. 1995. National Research Council.
8. Busch-Vishniac, I. J. 1992. The case for magnetically driven microactuators. *Sensors Actuators A* A33:207–20.
9. Zdeblick, M. J., R. Anderson, J. Jankowski, B. Kline-Schoder, L. Christel, R. Miles, and W. Weber. 1994. Thermopneumatically actuated microvalves and integrated electro-fluidic circuits. *Technical Digest: 1994 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC , June 13–16, 1994, pp. 251–55.
10. Bryzek, J., K. Petersen, J. R. Mallon, L. Christel, and F. Pourahmadi. 1990. *Silicon sensors and microstructures.* Fremont, CA: Novasensor.
11. Ravi, T. S., and R. B. Marcus. 1991. Oxidation sharpening of silicon tips. *J Vac Sci Technol B* 9:2733–37.
12. Editorial. 1991. Menlo Park, CA: TSDC (Teknekron).
13. Feynman, R. P. 1992. There's plenty of room at the bottom. *J Microelectromech Syst* 1:60–66.
14. Feynman, R. P. 1993. Infinitesimal machinery. *J Microelectromech Syst* 2:4–14.
15. Yole. 2007. Evolution of the MEMS markets and evolution of the MEMS foundry business. *Transducers*.
16. Collins, A. J. 1992. Problems associated with bringing a sensor technology to the market place. *Sensors Actuators A* A31:77–80.
17. Greenwood, J. C. 1988. Silicon in mechanical sensors. *J Phys E Sci Instrum* 21:1114–28.
18. Ikeda, K., H. Kuwayama, T. Kobayashi, T. Watanabe, T. Nishikawa, T. Yoshida, and K. Harada. 1990. Silicon pressure sensor integrates resonant strain gauge on diaphragm. *Sensors Actuators A* A21-A23:146–50.
19. Petersen, K., F. Pourahmadi, J. Brown, P. Parsons, M. Skinner, and J. Tudor. 1991. Resonant beam pressure sensor fabricated with silicon fusion bonding. *6th International Conference on Solid-State Sensors and Actuators (Transducers '91)*. San Francisco, CA, June 24–27, 1991, pp. 664–67.
20. Zook, J. D., D. W. Burns, H. Guckel, J. J. Sniegowski, R. L. Engelstad, and Z. Feng. 1992. Characteristics of polysilicon resonant microbeams. *Sensors Actuators A* 35:51.
21. Ikeda, K., H. Kuwayama, T. Kobayashi, T. Watanabe, T. Nishikawa, and T. Yoshida. 1988. Three dimensional micromachining of silicon resonant strain gauge. *7th Sensor Symposium*. Tokyo, Japan, May 30–31, 1988, pp. 193–96.
22. Burns, D. W., R. D. Horning, W. R. Herb, J. K. Zook, and H. Guckel. 1996. Sealed-cavity resonant microbeam accelerometer. *Sensors Actuators A* 53:249–55.
23. Bryzek, J., K. Petersen, L. Christel, and F. Pourahmadi. 1992. New technologies for silicon accelerometers enable automotive applications. *SAE Technical Papers Series* 920474:25–32.
24. Editorial. 1995. *Commercial brochure*.
25. Chang, S. 1998. An electroformed CMOS integrated angular rate sensor. *Sensors Actuators A* 66:138–43.
26. Lutz, M. G., W. Golderer, J. Gerstenmeier, J. Marek, B. Maihofer, S. Mahler, H. Munzel, and U. Bischof. 1997. A precision yaw rate sensor in silicon micromachining. *Proceedings: 1997 Int. Conf. on Solid-State Sensors and Actuators*. Chicago, June 1997, pp. 847–50.
27. Shkel, A., C. Acar, and A. Shkel. 2005. An approach for increasing drive-mode bandwidth of MEMS vibratory gyroscopes. *J Microelectromechan Syst* 14:520–28.
28. Roe, J. N. 1992. Biosensor development. *Pharm Res* 9:835–44.
29. Editorial. 1993. *Sensor Business Digest*, 1–11.
30. Battaglia, E. A. 1980. *Ion-selective electrode*. US Patent 4214968.
31. Editorial. 1994. Microminiature tools affecting medical products. *BBI Newsletter* 17:88–91.
32. Lauks, I. 1983. Multielement thin film chemical microsensors. *SPIE* 138–50.
33. Turner, A. P. F. 1994. *Third World Congress on Biosensors.* New Orleans, LA, June 1–3, 1994. Amsterdam: Elsevier Advanced Technology.
34. Tierney, M. J., H. L. Kim, M. D. Burns, J. A. Tamada, and R. O. Potts. 2000. Electroanalysis of glucose in transcutaneously extracted samples. *Electroanalysis* 12:1–6.
35. Gross, T. M., B. W. Bode, D. Einhorn, D. M. Kayne, J. H. Reed, N. H. White, and J. J. Mastrototaro. 2000. Performance evaluation of the MiniMed continuous glucose monitoring system during patient home use. *Diabetes Technology and Therapeutics* 2:49–56.
36. Roe, J. N., and B. R. Smoller. 1998. Bloodless glucose measurements. *Therapeutic Drug Carrier Systems* 15:199–241.
37. Klonoff, D. C. 1997. Noninvasive blood glucose monitoring. *Diabetes Care* 20:433–37.
38. Heller, A., and B. Feldman. 2008. Electrochemical glucose sensors and their applications in diabetes management. *Chem Rev* 108:2482–505.
39. Burbaum, J. J. 2000. The evolution of miniaturized well plates. *J Biomol Screen* 5:5–8.
40. Shapiro, H. M., and M. Hercher. 1986. Flow cytometers using optical waveguides in place of lenses for spectroscopic illumination and light collection. *Cytometry* 7:221–23.
41. Miyake, R., H. Ohki, I. Yamazaki, and R. Yabe. 1991. A development of micro sheath flow chamber. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '91)*. Nara, Japan, pp. 259–64.
42. Sobek, D., A. M. Young, M. L. Gray, and S. D. Senturia. 1993. An investigation of micro structures, sensors, actuators, machines, and systems. *Proceedings: IEEE Micro Electro Mechanical Systems (MEMS '93)*. Fort Lauderdale, FL, pp. 219–24.
43. Sobek, D., S. D. Senturia, and M. L. Gray. 1994. Microfabricated fused silica flow chambers for flow cytometry. *Technical Digest: 1994 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC, June 13–16, 1994, pp. 260–63.
44. Sutherland, R. M., and C. Dahne. 1987. In *Biosensors: Fundamentals and applications,* eds. A. P. F. Turner, I. Karube, and G. S. Wilson, 655–78. Oxford, United Kingdom: Oxford University Press.
45. Garabedian, R., C. Gonzalez, J. Richards, A. Knoesen, R. Spencer, S. D. Collins, and R. L. Smith. 1995. Microfabricated surface plasmon sensing system. *Sensors Actuators A* A43:202–07.
46. Jorgenson, R. C., and S. S. Yee. 1993. A fiber-optic chemical sensor based on surface plasmon resonance. *Sensors Actuators B* B12:213–20.
47. Melendez, J., R. Carr, D. U. Bartholomew, K. Kukanskis, J. Elkind, S. Yee, C. Furlong, and R. Woodbury. 1996. A commercial solution for surface plasmon sensing. *Sensors Actuators B* 35:1–5.
48. Melendez, J., R. Carr, D. U. Bartholomew, H. Taneja, S. Yee, C. Jung, and C. Furlong. 1997. Development of a surface plasmon resonance sensor for commercial applications. *Sensors Actuators B* 38–39:375–79.

49. Ballerstadt, R., and J. S. Schultz. 1998. Kinetics of dissolution of concanavalin a/dextran sols in response to glucose measured by surface plasmon resistance. *Sensors Actuators B* 46:50–55.
50. Madou, M. J., and T. Otagawa. 1989. Micron and submicron electrochemical sensors. *AICHE Symposium Series* 267:7–14.
51. Madou, M. J., and T. Otagawa. 1989. *Microelectrochemical sensor and sensor array*. US Patent 4874500.
52. Wilson, G. S., and Y. Hu. 2000. Enzyme-based biosensors for in-vivo measurements. *Chem Rev* 100:2693–704.
53. Schauer-Vukasinovic, V., L. Cullen, and S. Daunert. 1997. Rational design of a calcium sensing system based on induced conformational changes of calmodulin. *J Am Chem Soc* 119:11102–03.
54. Salins, L. L. E., V. Schauer-Vukasinovic, and S. Daunert. 1998. Optical sensing systems based on biomolecular recognition of recombinant proteins. *SPIE* 3270:16–24.
55. Salins, L. L. E., J. Lundgren, and S. Daunert. 1999. A dynamical investigation of acrylodan-labeled mutant phosphate binding protein. *Anal Chem* 71:589–95.
56. Salins, L. L. E., R. A. Ware, C. Ensor, and S. Daunert. 2001. A novel reagentless sensing system for measuring glucose based on the galactose/glucose-binding protein. *Anal Biochem* 294:19–26.
57. Tolosa, L., I. Gryczynski, L. R. Eichhorn, J. D. Dattelbaum, F. N. Castellano, Rao, and J. R. Lakowicz. 1999. Glucose sensor for low-cost lifetime-based sensing using a genetically engineered protein. *Anal Biochem* 267:114–20.
58. Petersen, K. E., W. A. McMillan, G. T. A. Kovacs, M. A. Northrup, L. A. Christel, and F. Pourahmadi. 1998. Towards next generation clinical diagnostic instruments: Scaling and new processing paradigms. *J Biomed Microdevices* 1:71–79.
59. Manz, A., N. Graber, and H. M. Widmer. 1990. Miniaturized total chemical analysis systems: A novel concept for chemical sensing. *Sensors Actuators B* B1:244–48.
60. McMillan, W. A., K. E. Petersen, L. A. Christel, and M. A. Northrup. 1999. Application of advanced microfluidics and rapid PCR to analysis of microbial targets. *Proceedings of the 8th International Symposium on Microbial Ecology,* eds. C. R. Bell, M. Brylinsky, and P. Johnson-Green. Halifax, NS, Canada: Atlantic Canada Society for Microbial Ecology.
61. Li, P. C. H., and D. J. Harrison. 1997. Transport, manipulation, and reaction of biological cells on-chip using electrokinetic effects. *Anal Chem* 69:1564–68.
62. Kuske, C. R., K. L. Banton, D. L. Adorada, P. C. Stark, K. K. Hill, and P. J. Jackson. 1998. Small-scale DNA sample preparation method for field PCR detection of microbial cell and spores in soil. *App Environ Microbiol* 64:2463–72.
63. Belgrader, P., W. Benette, D. Hadley, J. Richards, P. Stratton, R. Mariella, Jr., and F. Milanovich. 1999. PCR detection of bacteria in seven minutes. *Science* 284:449–50.
64. Cheng, J., E. L. Sheldon, L. Wu, A. Uribe, L. O. Gerrue, J. Carrino, M. J. Heller, and J. P. O'Connell. 1998. Preparation and hybridization analysis of DNA/RNA from *E. coli* on microfabricated bioelectronic chips. *Nat Biotechnol* 16:541–45.
65. Lee, S. W., and Y.-C. Tai. 1999. A micro cell lysis device. *Sensors Actuators A* 73:74–79.
66. Wilding, P., L. J. Kricka, J. Cheng, G. Hvichia, M. A. Shoffner, and P. Fortina. 1998. Integrated cell isolation and polymerase chain reaction analysis using silicone microfilter chambers. *Anal Chem* 257:95–100.
67. Anderson, R. C., X. Su, G. J. Bogdan, and J. Fenton. 2000. A miniature integrated device for automated multistep genetic assays. *Nucleic Acids Res* 28:e60.
68. Han, F., and S. J. Lilard. 2000. *In situ* sampling and separation of RNA from individual mammalian cells. *Anal Chem* 72:4073–79.
69. Khandurina, J., T. E. McKnight, S. C. Jacobson, L. C. Waters, R. S. Foote, and J. M. Ramsey. 2000. Integrated system for rapid PCR-based DNA analysis in microfluidic devices. *Anal Chem* 72:2995–3000.
70. Pourahmadi, F., M. Taylor, G. Kovacs, K. Lloyd, S. Sakai, T. Schafer, B. Helton, L. Western, S. Zaner, J. Ching, B. McMillan, P. Belgrader, and M. A. Northrup. 2000. Toward a rapid, integrated, and fully automated DNA diagnostic assay for *Chlamydia tranchomatis* and *Neisseria gonorrhoeae*. *Clin Chem* 46:1511–13.
71. Lagally, E. T., I. Medintz, and R. A. Mathies. 2001. Single-molecule DNA amplification and analysis in an integrated microfluidic device. *Anal Chem* 73:565–70.
72. Krishnan, M., V. Namasivayam, R. Lin, R. Pal, and M. A. Burns. 2001. Microfabricated reaction and separation systems. *Anal Biotechnol* 12:92–98.
73. Yang, J. M., J. Bell, Y. Huang, M. Tirado, D. Thomas, A. H. Forster, R. W. Haigis, P. D. Swanson, R. B. Wallace, B. Martinsons, and M. Krihak. 2001. An integrated, stacked microlaboratory for biological agent detection with DNA and immunoassays. *Biosens Bioelectron* 17:605–18.
74. Kellogg, G. J., T. E. Arnold, B. L. Carvalho, D. C. Duffy, and N. F. Sheppard. 2000. Centrifugal microfluidics: Applications. *Micro total analysis systems,* eds. A. van den Berg, W. Olthuis, and P. Bergveld. Enschede, The Netherlands: Kluwer Academic Publishers, pp. 239–42.
75. Yuen, P. K., L. J. Kricka, P. Fortina, N. J. Panaro, T. Sakazume, and P. Wilding. 2001. Microchip module for blood sample preparation and nucleic acid amplification reactions. *Genome Res* 405–12.
76. Wilding, P., M. A. Shoffner, and L. J. Kricka. 1994. PCR in a silicon microstructure. *Clin Chem* 40:1815–18.
77. Cheng, J., M. A. Shoffner, G. E. Hvichia, L. J. Kricka, and P. Wilding. 1996. Chip PCR: Investigation of different PCR amplification systems in micro-fabricated silicon-glass chips. *Nucleic Acids Res* 24:380–85.
78. Brody, J. P., T. D. Osborn, F. K. Foster, and P. Yager. 1995. A planar microfabricated fluid filter. *8th International Conference on Solid State Sensors and Actuators, and Eurosensors IX*. Stockholm, Sweden, June 1995, pp. 779–82.
79. Fluitman, J. H., A. van den Berg, and T. S. Lammerink. 1995. In *Micro total analysis systems,* eds. A. van den Berg, and P. Bergveld, 73–83. Norwell, MA: Kluwer Academic Publishers.
80. Pipper, J., Y. Zhang, P. Neuzil, and T-M. Hsieh. 2008. Clockwork PCR including sample preparation. *Angew Chem Int Ed* 47:3900–04.
81. Kido, H., M. Micic, D. Smith, J. Zoval, J. Norton, and M. Madou. 2007. A novel, compact disk-like centrifugal microfluidics system for cell lysis and sample homogenization. *Colloids Surf B: Biointerfaces* 58:44–51.
82. Scott, C. D., and J. C.Mailen. 1972. Dynamic introduction of whole-blood samples into fast analyzers. *Clin Chem* 18:749–52.
83. Brenner, T., R. Zengerle, and J. Ducrée. 2003. A flow-switch based on coriolis force. In *Proceedings of μTAS*. Lake Tahoe, CA, October 2003. Norwell, MA: Kluwer Academic.
84. Angell, J. B., S. C. Terry, and P. W. Barth. 1983. Silicon micromechanical devices. *Sci Am* 248:44–57.

85. Editor. 1998. MicroChemical Systems.
86. Madou, M. J. 1994. Compatibility and incompatibility of chemical sensors and analytical equipment with micromachining. *Technical Digest: 1994 Solid State Sensor and Actuator Workshop*. Hilton Head Island, SC, June 13–16, 1994, pp. 164–71.
87. Lyle, R. P., and D. Walters. 1997. In *1997 International Conference on Solid-State Sensors and Actuators*, 975–78. Chicago.
88. Madou, M. J., and S. R. Morrison. 1989. *Chemical sensing with solid state devices.* New York: Academic Press.
89. Buck, L., and R. Axel. 1991. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* 65:175–87.
90. Madou, M. 1994. *NIST Special Publication 865*. Gaithersburg, MD: National Institute of Standards and Technology.
91. Goldstein, M. K. 1991. *Biomimetic sensor that simulates human response to airborne toxins, Quantum Group*. US Patent 5063164.
92. Sandifer, J. R., and J. J. Voycheck. 1999. A review of biosensor and industrial applications of pH-ISFETs and an evaluation of Honeywell's"DuraFET". *Mikrochim Acta* 131: 91–98.
93. Kolesar, E. S., and R. R. Reston. 1995. Separation and detection of toxic gases with a silicon micromachined gas chromatography system. *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications, National Aeronautics and Space Administration (NASA) and The Aerospace Corporation*. Houston, TX, October 30–November 3, 1995, pp. 102–14.
94. Terry, S. C., J. H. Jerman, and J. B. Angell. 1979. A gas chromatographic air analyzer fabricated on a silicon wafer. *IEEE Trans Electron Dev* ED-26:1880–86.
95. Overton, E. B. Ohio State University. Personal communication.
96. Leeuwis, H., and J. a. v. V. 1994. In *Microsystem technology*, ed. G. K. Lebbink. Alphen aan den Rijn/Zaventem, The Netherlands: Samsom BedrijfsInformatie, pp. 90–94.
97. Goldner, H. 1993. 'Honey, they shrunk the GC!' Gas chromatographs go mini. *Res Dev* June: 45.
98. Stalder, R. E., S. Boumsellek, T. R. Van Zandt, T. W. Kenny, M. H. Hecht, and F. E. Grunthaner. 1993. Micromachined array of electrostatic energy analyzers for charged particles. *J Vac Sci Technol A* 12:2554–58.
99. Bryzek, J., S. Roundy, B. Bircumshaw, C. Chung, K. Castellino, J. R. Stetter, and M. Vestel. 2006. Marvelous MEMS: Advanced IC sensors and microstructures for high volume applications. *IEEE Circuits Devices Mag* 22: 8–28.
100. Leroy, C., M. B. Pisani, C. Hibert, D. Bouvet, M. Puech, and A. M. Ionescu. 2006. High Quality Factor Copper Inductors Integrated in Deep Dry-Etched Quartz Substates. *Symposium on Design, Test, Integration and Packaging of MEMS/MOEMS*, Stresa, Italy, pp. 26–28.
101. Albert, P., J. Costello, and P. Salomon. 2000. RF/microwave applications of MEMS integrated in CMOS technology. *MST News* June: 12–13.
102. Ehmann, K. F., D. Bourell, M. I. Culpepper, T. J. Hodgson, T. Kurfess, M. Madou, K. Rajurkar, and R. DeVor. 2007. *International assessment of research and development in micromanufacturing*. New York: Springer.
103. Edwards, S. A. 2006. *The nanotech pioneers: Where are they taking us?* Weinheim, Berlin: Wiley-VCH.
104. Petersen, K. E. 1982. Silicon as a mechanical material. *Proc IEEE* 70:420–57.
105. Gasman, L. 2006. *Nanotechnology applications and markets.* Norwood, MA: Artech House Publishers.

Index

A

Accelerometers
  automotive
    airbag accelerometer, 560–561
    suspension accelerometers, 561–563
  in consumer electronics, 610–611
  manufacturing options, 563–564
  market perspectives, 559–560
  performance requirements for, 567
Acoustic imaging, 420–421
Acoustic sensing, of mass, 417–420
Acoustic streaming, 31–32
Acoustic waves
  devices
    bulk, 408
    bulk acoustic wave, 408
    flexural plate waves, 409–410
    SAW, 408–409
  types, 407–408
Active sensors, 390
Actuators
  vs. chemical/electrochemical, 461–462
  design principles, 388
  electrostatics
    dynamics of, 395–397
    modeling, 394–395
    parallel plate capacitor (*see* Parallel plate capacitors)
  energy conversion, 387–388
  force exerted *vs.* displacement, 463, 465
  frequency *vs.* displacement, 463, 465
  magnetic
    bead array counter (BARC), 436–437
    capacitive motor, 430
    vs. electrostatic actuation, 437–438
    force, translational motion, 432
    levitation valve, 433–434
    micromotors, 434–436
    scaling, solenoid, 432
  mass sensitivity, 465
  microbalance *vs.* cantilever, size, 467
  micromachined single-crystal silicon cantilevers, 466–467
  nanomechanical resonators, 468
  vs. quantum, 462–463
  self-generating *vs.* modulating transducers, 390
  smart materials, 391
  stresses, cantilever beam, 466
  thermal (*see* Thermal actuators)
  transduction principles, 389–390
Additive process
  comparison of, microsensors and micromachining, 20, 21
  selection criteria for, 20
Aging, 77–78
Airbag accelerometer, 560–561
Airy patterns, 294
Allometry, 363
Amino acids, 112
Amorphous silicon
  vs. amorphous polysilicon, 262
  behavior, 261–262
  polycrystalline-based solar cells *vs.* ultrathin, 262–263
Amperometric invasive glucose sensor, 279, 280
Angular rate sensors
  Coriolis force, 567–568
  cost, 566
  Delphi-Delco, 571
  design principle, 569–570
  double-ended tuning fork, 568–569
  drive-mode oscillators, 572
  electronic gyro, 564
  fiber optic gyroscope, 566
  market perspectives, 566–567
  mechanical gyroscope, 565
  mechanism, 564–565
  performance requirements for, 567
  ring-laser gyroscope, 565–566
  rotating ring gyroscope, 570
  Shkel angular rate sensors, 571–572
  vibrating ring gyroscope, 570–571
Anodisc™ membrane filters, 255
Argonne method, 251
Artificial intelligence, 516–518
Artificial life, 519–520
Artificial neural network software, 520–521
Assembly processes
  capillary tension force, ball *vs.* plate, 222–223
  cilia arrays, 224
  deterministic *vs.* stochastic, 220
  electrostatic force, spherical particle *vs.* wall, 222
  issues, 224–225
  Lamb *vs.* Rayleigh waves, 223–224
  micro-and nanoassembly approaches, 220–221
  microhouse, 223
  part size and assembly accuracy, 220
  scaling effects, 221–224
  surface forces *vs.* gravitational force, 221–222
  vacuum gripper, 223–224
Atomic force microscopy (AFM)
  cantilever model and examples, 317–319
  contact mode, 319
  feedback loop, 319–320
  non-contact mode, 319
  surface force measurement, 317
  tapping mode, 319
  Veeco instruments, 320
Auger process, 329–330
Automotive sensors
  accelerometers
    airbag accelerometer, 560–561
    manufacturing options, 563–564
    market perspectives, 559–560
    suspension accelerometers, 561–563
  angular rate sensors (*see* Angular rate sensors)
  in car, 552
  migration status of, 550, 552
  pressure sensor
    direct tire pressure monitoring system, 555–556
    manifold air pressure measurement, 554
  sensors in passenger vehicles, 549–550
  solid state oxygen sensor, 550
  in transportation context, 550, 551

B

Batch fabrication techniques, 10–11
Batteries
  3D architectures, charge-insertion, 479–480
  primary and secondary, 477
  reversible, 479
  specific energy and energy density, 477
  thin-film, 479
BCS model
  coherence length, 455
  Cooper pair formation, 454
  energy bandgap, 453
  impurity and defects, 453
  lattice deformation, 454
  origins, 452
  pairing, phonons, 454
  phase transition, 454
  theory of superconductivity, 453

Zeeman splitting, 455
zero-spin particles, 455
Bead array counter (BARC), 436–437
Bimetallic actuators, 456–457
Biocompatibility
definition, 277
mechanism
amperometric invasive glucose sensor, 279, 280
biological system perspective, 278
fibrinogen, 278
implant perspective, 278
plasma protein adsorption, 278
surface modifications, 279–280
Biofuels, 487–490
Biogenesis
Archaea, 58–60
clay platelets, 57–58
common terms used, 57
complexity theory, 60–61
DNA, RNA, and proteins, 56–57
information and communication theory, 61
Biomimetics
amino acids, biological systems, 112
biogenesis
Archaea, 58–60
clay platelets, 57–58
common terms used, 57
complexity theory, 60–61
DNA, RNA, and proteins, 56–57
information and communication theory, 61
biological cells, 83–84
biotechnology and
definition, 102–103
genetically engineered proteins, 103–104
polymerase chain reaction, 104–107
bottom-up manufacturing, 51
cloning
recombinant DNA technology, 78–79
reproductive, 79–80, 82–83
therapeutic/biomedical, 80–81
DNA
building blocks, 61
double helix, 62–63
hybridization, 63
replication process, 64–65
Watson-Crick pairing, 63–64
DNA chips
DNA arrays, applications, 108–109
sequencing and hybridization, 108
energy efficiency, 109
extraterrestrial intelligence
biological elements, 55
cultural elements, 55
Drake equation, 54
physical elements, 54
extremophiles, 59, 60, 110–111
Fall of Icarus, 49, 50
genetic code
codon, 69
death and aging, 77–78
DNA and RNA, 69
evolution, 74–75
exons, 70
gene copying errors, 72
genes and chromosomes, 72–74
golgi apparatus, 71
human genome project, 76–77
introns, 70
protein-building process, 69
retroviruses, 71–72
ribozymes, 70–71
transcription, 69
transport RNA, 71
transposomes, 75–76
genetic engineering, 110
human and natural mechanical designs, 51
hydrophobic and hydrophilic forces
amphiphilic molecules, 96
bicontinuous surface, 98–99
clathrate compound, 97
definition, 96
Gibbs free energy, 97
hydrocarbon groups, 96–97
micelle, 97–98
liposomes, 99–100
micelles, 99–100
nanomanufacturing, 126–127
natural biopolymers
coil entropy, 100–101
intrinsic self-assembling properties, DNA, 101
protein folding and unfolding, 102
nature's engineering feats
complex systems, 54
life, definition, 52–53
living things, characteristics, 53
thermodynamics, 53–54
organelles
functions, 84–85
mitochondria and phylogenetic trees, 85–86
mitochondria and power, 86–91
molecular motors, 91–96
prokaryotes and eukaryotes, 83–84
proteins
emergence, 66
examples, 66
making, 65–66
roles, 65
self-repair and recycling, 67–68
structural organizational levels, 68
structure, 67, 68
water release in, 66
stem cells
bioethics, 82–83
cell-reprogramming, 81–82
characteristics, 82
types, 81
synthetic cell, 52
universal genetic code, 113
Biopolymers
coil entropy *vs*. bond enthalpy, DNA, 101
intrinsic self-assembling properties, DNA, 101
noncovalent interactions, 100–101
protein folding and unfolding, 102
Biosensors, polymer/metal-based
biomodules, 43
buried SU-8 microchannel, 44–45
chemical sensor array, 42, 43
integrated sensor array fabrication method, 42, 43
ion selective electrode array element fabrication, 43, 44
photoresist material properties, 44
Biotechnology, *see also* Medical and biotechnology, MEMS in
definition, 102–103
genetically engineered proteins, 103–104
polymerase chain reaction, 104–107
Blister test, 337
Blood glucose monitor
specifications, 24–25
types, 23
BMGs, *see* Bulk metallic glasses (BMGs)
Brains, miniature devices
artificial neural networks, 512
entanglement, 514
MOSFET switch, 513
from neurons, 514–516
qubit, 514
spintronics devices, 513
Buckingham P theorem, 365–366
Buckminsterfullerene, *see* Buckyballs, 0D
Buckyballs, 0D, 160–162
Bulk acoustic wave (BAW) devices, 408
Bulk metallic glasses (BMGs)
advantages, 273–274
crystallization, 270–271
mechanical properties and processing temperature, 272, 273
supercooled liquid region (SCLR), 271
thermoplastic forming (TPF), 271

C

Calmodulin, 103–104
Capacitors, 480–481
Cap-on carrier packaging method, 214–215
Carbon-based MEMS (C-MEMS)
carbon, advantages, 265
carbon microfabrication technique, 266
3D carbon electrodes, 270
Li-ion batteries, 267–268
photolithography, 266–267
photoresist-derived architectures, 267
principle, 266
self-organized groups, carbon posts, 269

suspended carbon fibers, 268, 269
three level structure, 268
washcloth line sensors, 269
Carbon nanostructures
buckyballs, 0D, 160–162
fullerenes, 159–160
nanotubes, 1D, 162–165
Carbon nanotubes
applications
chemical sensors and nanoreactors, 172
in composites, 172–173
electronic properties, 169–172
mechanical properties, 172
manufacture, 165–169
Casimir forces, 462–464
Cavity sealing
epitaxial deposition, 217–218
hexagonal honeycomb poly-Si (HEXSIL), 218–219
polysilicon and silicon nitride, 215–217
CCD image sensors, 11
CD optical pickup, 7, 8
Centrifugal pumping, 33
Ceramics
advantages, 254
polycrystalline TiO_2 anodization, 257
porous anodic alumina (*see* Porous anodic alumina)
traditional and advanced, 254
Chemical machining (CHM), 22
Chemical sensors, 172
Chip scale packaging (CSP), 214
Cloning
recombinant DNA technology, 78–79
reproductive, 79–80, 82–83
therapeutic/biomedical, 80–81
CLSM, *see* Confocal laser scanning microscopy (CLSM)
C-MEMS fractals, 511–512
Coanda effect, 374
Complexity theory, 60–61, 518–519
Compositional metrology
Auger electron spectroscopy and scanning Auger microscopy, 329–330
EDAX, 329
Computer-aided design (CAD), MEMS
ALGOR, 350
anisotropic crystalline etch simulation (ACES™), 348
AnSoft, 347
CAEMEMS software package, 347
CoventorWare™, 348
design, simulation, and verification, 342, 343
early development, 346
evolution, 346
FEMLAB, 350
future of, 350–351
IntelliCAD architecture, 348, 349
mask-to-model approach, 346
mechanical drafting tools, 347
MEMCAD architecture, 348, 349
NiTi-based valve, 3D visualization, 342, 343
process simulators, 347
SUGAR V3.1, 348
Confocal laser scanning microscopy (CLSM), 298
Confocal microscopy, 297–298
Continuous manufacturing
in-line lithography equipment, 12
printed electronics engineering, 12–13
printed products on flexible substrates, 11, 12
printing industry, 11
roll-to-roll manufacturing, 11, 12
Cooper pairs, 454
Cyborg beetle system, 493–494

D

Darwin's theory, 75
Dew point sensor, 419–420
Diamond
amorphous, 252
film synthesis
CVD, 250, 251
lithographic patterning, 252
selective deposition, 252
p-i-p structure fabrication process, 252
vs. semiconductors, properties, 250
surface micromachining, 251
wide bandgap semiconductor material, 249
Diamond-based MEMS, 252
Dielectric induction, 421–422
Dielectrophoresis, 226
Differential scaling, 363
Digital mirror device, 215
Dimensional homogeneity theorem, 365
Dimensional metrology
atomic force microscopy (AFM), 317–320
confocal microscopy, 297–298
ellipsometry (*see* Ellipsometry)
fluorescence microscopy, 296–297
interferometry (*see* Interferometry)
light microscopy, 289–296
NSOM/SNOM, 310–313
scanning electron microscopy, 304–309
scanning proximal microscopes, 309–310
scanning tunneling microscope (STM), 313–317
scatterometry, 287, 289
two-photon microscopy, 298–299
Disk method, 333–334
DNA
building blocks, 61
chips
DNA arrays, applications, 108–109
sequencing and hybridization, 108
double-stranded, 62–63
hybridization, 63
replication process, 64–65
Watson-Crick pairing, 63–64
DNA arrays
applications of, 108–109
molecular diagnostics, 580–581
DNA-mediated self-assembly, 229–230
Dog-bone method, 334–337
Instron sample, 335
micro-Instron specimen fabrication, 335–336
micromachined test structure, ultimate strain, 334, 335
Young's modulus measurements, 334, 335
Dual inline package (DIP), 193–194
Dye-sensitized Grätzel design, 502–503

E

EDAX, *see* Energy dispersive analysis of X-rays (EDAX)
EFM, *see* Electrostatic force microscopy (EFM)
Electrets, 429
Electric actuators, 391
Electrical connection methods
flip-chip bonding, 198–200
tape automated bonding (TAB), 198
wire bonding, 197–198
Electrochemical actuators, 461–462
Electrochemical double-layer (ECDL) capacitors, 481
Electrochemical machining (ECM), 22
Electrochemical sensors
blood electrolyte and blood gas sensors, 577–578
gas sensor, 600–601
tabletop and handheld in vitro diagnostics, 576–577
Electrochemistry, fractals
2D *vs.* 3D scaling, 511
fractal electrode, 510
miniaturization, 508
networks, 509–510
scaling of, cube and sheet, 508–509
Electronic noses, 598–600
Electronic tongue, 600
Electro-osmosis, 32–33, 226
Electrophoresis, 226
Electrorheological fluids, 429–430
Electrostatic actuators
dynamics of, 395–397
electrical and mechanical circuits, 395
merits, 398–400
parallel plate capacitor (*see* Parallel plate capacitors)
performance and modeling, 400–402
Electrostatic force microscopy (EFM), 321
Electrostatic resonators
fundamental frequency, 397
quality factor, 398

Electrostatics
continuum theory, breakdown of, 380–381
scaling laws
breakdown voltage, 379
isometric scaling, 378
mean free path, 380
parallel plate capacitor, 379
Paschen curve, 380
stored energy, 380
Electrostriction
ceramic piezoelectrostrictor, ultrasonic motor, 423
mechanical deformations, 422
progressive wave, 423
Ellipsometry
ellipsometer setup, 301, 302
noncontact optical technique, 301
polarization change, 301
theory, 302–304
Energy dispersive analysis of X-rays (EDAX), 329
Energy harvesting
vs. batteries, fuel cells, and nuclear power, 505–506
energy sources, 505
vibrational types, 505–506
Environmental SEM (ESEM), 309
Eukaryotes, 57, 83–84
Everhart-Thorny (E-T) detector, 308
Exons, 70
External-combustion engines, 494
Extremophiles, 59, 60, 110–111

F

Fast microfluidic CD prototyping process
poly-dimethylsiloxane (PDMS), 275–276
PSA layers, 274–275
Ferrofluids
advantages and disadvantages, ferrowax, 447–448
coercivity *vs.* particle size, 444
ferrowax plug, 448
fractals, 446
hysteresis loop, 444
magnetic force, 445
normal-field instability, 445–446
positioning systems, multiaxle movement, 445–446
wax valves, 447
Fiber optic gyroscope, 566
Film bulk acoustic resonator (FBAR), 408
Finite element analysis
boundary and loading conditions, 345
data correlation and design optimization, 346
geometry development, 343–344
material property assignment, 344
mesh generation, 344–345
run analysis, 345
von Mises stress plot, 345–346
Flexural plate wave (FPW) devices, 409–410
bimorph and stack actuator, 410
construction and operation, 409–410
Lamb waves, 409
vs. SAW, 410
Flip-chip bonding method, 198–200
Flow cytometry, 582–583
Fluorescence microscopy, 296–297
Fractals, 364–365
C-MEMS, 511–512
definition, 506
dissipative structure, 508
in electrochemistry, 508–511
Hausdorff dimension, 507
in nature, 508
Front-end and back-end processes, 189–191
Fuel cells
microfuel cells (*see* Microfuel cells)
schematic representation, 482
stationary products, 483
type of, 482
Fullerenes, 159–160

G

Gas chromatography
advantages, 602
functional block diagram, 601
Hewlett-Packard hand held, 602
integrated, 602
mass spectrometry system, 603–604
vapor separation, 601–602
Gas phase synthesis, nanotechnology
liquid and vapor precursors, 158–159
solid-state precursors, 157–158
Genetically engineered proteins, 103–104
Genetic code
codon, 69
death and aging, 77–78
DNA and RNA, 69
evolution, 74–75
exons, 70
gene copying errors, 72
genes and chromosomes, 72–74
golgi apparatus, 71
human genome project, 76–77
introns, 70
protein-building process, 69
retroviruses, 71–72
ribozymes, 70–71
transcription, 69
transport RNA, 71
transposomes, 75–76
GIXRD, *see* Gracing incidence X-ray diffraction (GIXRD)
Glucose sensors
amperometric, 579
diabetes mellitus, 579
GlucoWatch® biographer, 32, 33, 580
manufacturing techniques, 579
MEMS microneedles, 580
Gracing incidence X-ray diffraction (GIXRD), 329

H

Hard superconductors, 451
Heat engines, MEMS
digital propulsion, 496
external and internal, 494
gas turbine, 495
megapixel microthruster array, 496
silicon/glass thruster structure, 495
in solar cells, 501–502
High-critical-temperature superconductors (HTS), 451
Hot-point probe operation, 323, 324
Human genome project, 76–77
Humidity sensor, 419–420
Hybridization, 63, 108
Hydrogenated amorphous silicon (a-Si:H)
active semiconductor device fabrication, 262
optoelectronic parameters, 263, 264
plasma-enhanced chemical vapour deposition (PECVD), 263
Hydrogen storage, 485–487
Hydrophobic and hydrophilic forces
amphiphilic molecules, 96
bicontinuous surface, 98–99
definition, 96
Gibbs free energy, 97
hydrocarbon groups, 96–97
hydrophobic effect, 97
micelle, 97–98

I

Industrial automation
electrochemical and optochemical gas sensors, 600–601
electronic noses and tongues, 598–600
gas chromatography, 601–603
ISFET technology, 601
mass spectrometry, 603–604
micromachine analytical instruments, 597
Nexus III report, 596
optical absorption instruments, 604–605
solid-state gas sensors, 597–598
Ink-jet heads, 606–607
In situ fabrication, nanoparticles
liquid phase synthesis, 154–157
lithography techniques, 146–147
self-organized epitaxial growth (*see* Self-organized epitaxial growth, nanotechnology)
solid-state processes, 153–154

In situ stress-measuring techniques, thin films
biaxial measurements, 333–334
local, intrinsic and residual stress, 333
substrate bending principle, 333
uniaxial measurements, 334–337
Instron sample, 335
Integrated sample preparation
DNA analysis devices, 592, 593
OrbiPrep system, 594–596
Tecan Boston prototype, 592–594
Interferometry
definition, 299
laser Doppler vibrometer, 323–324
Michelson interferometer, 299–300
noncontact profilometry, 299
stroboscopic LED, 323–324
SWLI, 299, 300
Internal-combustion engines, 494
Introns, 70
In vitro diagnostics, 575–576
In vivo MEMS biosensor
glucose sensing, 586–587
pH and blood gases, 586
Ion mobility spectrometer (IMS)
fabrication, microvolcanoes, 383
field ionization (FI) technique, 382
micromachined volcanoes, 382
microvolcano ionization sources, 382–383
Ni ionization source, 381
Ion-sensitive field-effect transistors (ISFETs), 601
Iontophoresis, 32–33
Isometric scaling, 357
IT peripherals and telecommunications markets, MEMS in
accelerometers, 610–611
definition, 605
ink-jet heads, 606–607
MEMS foundries, 613
MEMS microdisplays, 607–608
optical switches, 608–610
read-write heads, 606
RF-MEMS, 611–613

J

Johnson & Johnson blister pouch HIV test, 30–31

L

Lab-on-a-Chip, 539–540
Laser Doppler vibrometer, 324
Layer bonding, MEMS
alignment, 210–212
field-assisted thermal bonding, 201–204
imaging and bond strength, package hermeticity tests, 212–214
organic layers, 208–209
plastic-plastic bonding, 209–210
silicon fusion bonding, 204–207
thermal bonding, intermediate layers, 207–208
Light microscopy
airy patterns, 294
divisions, 294–295
history, 289, 292, 293
Nomarski microscope, 295
oil immersion microscopy, 294
optical microscopy, 289, 295, 296
phase-contrast microscopy, 295
resolution, 293–294
uses, 293
Liposomes, 99–100
Liquid-junction solar cell, 502–503
Liquid phase synthesis
acoustic cavitation process, 156
apparatus, quantum dots, 155
colloidal processing, 154
microemulsion nanoparticle, 156
nanoparticle formation, solution, 155
Lithographic diamond patterning, 252
Lithography-based MEMS
all-optical switching, 530–531
definition, 529
digital micromirror devices, 532
high-throughput screening, 531
mechanical and MOEMS devices, 532–533
MEMS market, 529–530
microfluidic application, 531–532
security applications, 533
Low-critical-temperature superconductor (LTS), 450
Low-pressure chemical vapor deposition (LPCVD) silicon nitride
etch characteristics, 261, 262
vs. PECVD, properties, 260, 261
Lysing
mechanical methods, 590
nonmechanical methods, 590–592

M

Magnetic actuators
bead array counter (BARC), 436–437
capacitive motor, 430
vs. electrostatic actuation, 437–438
force, translational motion, 432
levitation valve, 433–434
micromotors, 434–436
scaling, solenoid, 432
solenoid inductor, 431
Magnetic force microscopy (MFM), 320–321
Magnetic micromotors, 434–436
Magnetohydrodynamic (MHD) pumping, 441–443
Magnetorheological fluids, 443–444
Magnetostriction, 438–440
Magnetostrictive materials, 440–441
Manifold absolute pressure (MAP) sensor, 554
Mass-sensing devices
BAW, 418
FPW, 420
SAW, 418–420
Mass spectrometry (MS), 603–604
Matteucci effect, 440
Mechanical gyroscope, 565
Mechanical pumps, microfluidic propulsion
Animas-Debiotech Si-based micropump, 29–30
piezopump design, 30
pouch HIV test, 30–31
Medical and biotechnology, MEMS in
applications, 574
biomedical microinstrumentation, 574
biotechnology market, 573
DNA arrays, 580–581
electrochemical sensors, 576–578
flow cytometry, 582–583
global health care market, 573
glucose sensors, 578–580
in vitro diagnostics, 575–576
in vivo applications, 585–587
market character, 574–575
miniaturized well plates, 581–582
silicon-based bio-MEMS, 573
surface plasmon sensor, 583–585
MEMS, *see* Microelectromechanical systems (MEMS)
MFM, *see* Magnetic force microscopy (MFM)
MGS1100 carbon monoxide sensor, 598
Micelles, 99–100
Michelson interferometer, 299–300
Microactuators, 541
Microassembly
parallel
circuit integrity and LED activation, 229
electric field directed placement, LEDs, 228
fluid self-assembly, 226
microLEDs, 228
self-assembly, electrostatic levitation, 227
serial
AC electrokinetic effect, 226
bioparticles, 226
manipulation methods, 221, 224
optical tweezers, 224–225
single-beam gradient trap, 225
Microcomponents, 541–542
Microdisplays, 607–608
Microelectromechanical systems (MEMS)
applications
accelerometers, 559–560, 563–564
automotive accelerometer, 560–563
automotive market, 549–550
automotive pressure sensor, 554–556

consumer electronics sensor, 572–573
first-level packaged MEMS, 542–543
general sample preparation protocols, 588–592
in industrial automation and environmental monitoring markets, 596–605
ink-jet printer head, 527
integrated sample preparation, 592–596
in IT peripherals and telecommunications markets, 605–613
lithography-based MEMS, 529–433
market perspectives, 566–567
market studies, 535–537
MEMS angular rate sensors, 564–572
microactuators, 541
micromachine market, 543–549
micromachining, 538
micromechatronics, 540
microsensors, 540–541
microstructures or microcomponents, 541–542
microsystems technology, 539
micrototal analysis systems/ lab-on-a-chip, 539–540
non-lithography-based MEMS, 533–535
packaged digital micromirror devices, 527
power source (*see* Power, miniature devices)
pressure measurements, 550–554
pressure sensing, manufacturing approach, 556–557
resonant beam sensors, 557–559
silicon-based MEMS devices, 537–538
smallest commercialized units, 542–543
μ-TAS and LOC, molecular diagnostics, 587–588
brains (*see* Brains, miniature devices)
definition, 538–539
heat engines
digital propulsion, 496
external and internal, 494
gas turbine, 495
megapixel microthruster array, 496
silicon/glass thruster structure, 495
in solar cells, 501–502
in medical and biotechnology
applications, 574
biomedical microinstrumentation, 574
biotechnology market, 573
DNA arrays, 580–581
electrochemical sensors, 576–578
flow cytometry, 582–583
global health care market, 573
glucose sensors, 578–580
market character, 574–575
miniaturized well plates, 581–582
silicon-based bio-MEMS, 573
surface plasmon sensor, 583–585
in vitro diagnostics, 575–576
in vivo applications, 585–587
metrology
compositional, 327–330
dimensional, 287–324
EFM, 321
equipment classification, 286–287
measurement tools, 287–289
MFM, 320–321
nanoindentation, 321
in situ stress-measuring techniques, 333–338
structural (*see* Structural metrology)
thermal microscopy, 321–322
modeling, 341–343
CAD for MEMS, 346–351
finite element analysis, 343–346
processes
biocompatibility and surface modification, 277–280
BMGs, 270–274
C-MEMS, 265–270
fast microfluidic CD prototyping, 274–276
micro glass blowing, 276–277
substrate materials
amorphous silicon, 261–263
ceramics, 254–257
diamond, 249–252
gallium arsenide (GaAs), 245–249
hydrogenated amorphous silicon, 262–264
miniaturization, 240–241
polymer MEMS-polyimides, 257–259
quartz, 241–245
silicon carbide (SiC), 252–254
silicon nitride (Si_xN_y), 259–261
Microfluidics propulsion techniques
acoustic streaming, 31–32
centrifugal pumping, 33
comparison of, 29
electro-osmosis and iontophoresis, 32–33
mechanical pumps
Animas-Debiotech Si-based micropump, 29–30
piezopump design, 30
pouch HIV test, 30–31
vacuum pressure reservoir, 33–34
Microfuel cells
fuels and fuel storage, 485
methanol crossover, 485
micromachining, 490–494
Toshiba's miniature direct methanol fuel cell, 484
Micro glass blowing process, 276–277
Micro-Instron specimen fabrication, 335–336
Micromechatronics, 540
Micro-origami, 230–232
GaAs corner plates, strain, 231
micromirror, 230
microtubes and microspirals, 231–232
schematic, process, 230–231
self-assembled microstage fabrication, hinges, 231
three-dimensional structures, 232
Microsensors, 540–541
Microstructures, 541–542
Microsystems technology (MST), 539
Micrototal analysis (μ-TAS), 539–540
Miniaturized well plates, 581–582
Mitochondria
anabolic and catabolic process, 87–89
anaerobic respiration, 90
ATP
energizes process, 90
molecular structure, 87
Calvin-Benson cycle, 91, 92
cellular aerobic respiration, 89–90
glycolysis, 88–89
hydrolysis, ATP to ADP, 87
inner structure, 87
Krebs cycle, 89
metabolism, 90
photosynthesis, plants, 87–90, 92
and phylogenetic trees, 85–86
pigments, 91
pyruvate pathways, 89–90
Molecular assembler, 135–136
Molecular motors
ATPase, 92–94
bacterial flagellum, 93–94
dynein, 92, 95–96
kinesin motor, 93
myosin, 94
optical tweezers, 92, 93
tracks and functions, 92
types, 92
Molecular nanotechnology, 118
Monolithic *vs.* hybrid MEMS, 26–27
Multichip module (MCM)
categories, 195
four-chip, 196
two-chip, 195–196
wiring *vs.* flip-chip bonding, 196

N

Nanocomposites, 127
Nanoelectromechanical system (NEMS), *see also* Microelectromechanical systems (MEMS)
brains (*see* Brains, miniature devices)
CAD for, 350–351

materials, 240–241
metrology
compositional, 327–330
dimensional, 287–324
EFM, 321
equipment classification, 286–287
measurement tools, 287–289
MFM, 320–321
nanoindentation, 321
in situ stress-measuring techniques, 333–338
structural (*see* Structural metrology)
thermal microscopy, 321–322
modeling, 341–343
CAD for MEMS, 346–351
finite element analysis, 343–346
onboard energy sources, 476–477
power source (*see* Power, miniature devices)
solar cells, 501–504
Nanofabrication, 118
Nanomaterials, 620–622
Nanoparticles
blue shift, semiconductor, 141–143
magnetic properties, 145
manufacture, 145–146
plasmonic effects, metal, 143
superhydrophobicity and lotus effect, 144–145
transparent ceramics, 143–144
Nanoreactors, 172
Nanostructure properties
chemical effects, 140–141
heat capacity, 137–138
mechanical properties, 139–140
melting temperature, 137
vs. microstructures, 136
thermal conductivity, 137
thermoelectric power, 138–139
Nanotechnology
ATP, hydrolysis of, 129
biomimetic nanomanufacturing, 126–127
carbon nanostructures
buckyballs, 0D, 160–162
fullerenes, 159–160
nanotubes, 1D, 162–165
carbon nanotube
chemical sensors and nanoreactors, 172
in composites, 172–173
electronic properties, 169–172
manufacture, 165–169
mechanical properties, 172
carbon usage, 128–129
compartmentalization, cells, 130
cosmetics, 622
definition, 118
gas phase synthesis
liquid and vapor precursors, 158–159
solid-state precursors, 157–158
hierarchical organization and manufacturing sizes, 127–128
liquid phase synthesis, 154–157
lithography techniques, 146–147
molecular assembler, 135–136
nanochemistry, 118
nanocomposites, 127
nanomaterials
carbon nanotubes, 621
PEN, 620–621
nanoparticles
blue shift, semiconductor, 141–143
magnetic properties, 145
manufacture, 145–146
plasmonic effects, metal, 143
superhydrophobicity and lotus effect, 144–145
transparent ceramics, 143–144
nanostructure properties
chemical effects, 140–141
heat capacity, 137–138
mechanical properties, 139–140
melting temperature, 137
vs. microstructures, 136
thermal conductivity, 137
thermoelectric power, 138–139
overall market and market character
California-based companies, 616–619
initial public offering, 616
Lux research estimate, 614
nanolithographic tools, 615
Nanosys, 616
NGL equipment, 615
technical market research report, 614
self-assembly, 122–124
self-organization, 124–126
self-organized epitaxial growth, 147–153
self-replication, 130–131
solid-state processes, 153–154
stained glass, 118
supramolecular chemistry, 131–135
textiles, 622–623
tools, 620
top-down *vs*. bottom-up manufacturing
absolute tolerances, 121
convergence, 120–121
MEMS *vs*. nanochemistry, 120
micromachining, 121
self-assembly and self-organization, 119
Nanotubes, 1D, 162–165
Natural biopolymers, *see* Biopolymers
Natural selection, *see* Evolution
Near-field scanning optical microscopy (NSOM/SNOM), 310–313
Nomarski microscope, 295
Non-lithography-based MEMS
desk top factory (DTF), 533, 534
examples of, 534
global economy, 535
vs. lithography-based MEMS, 533
micromachining, 535
S100 nanomanipulator system, 534
Nonlithography *vs*. lithography-based manufacturing
additive processes, comparison of, 20, 21
Bacon's idols, 5
batch fabrication, 10–11
CCD image sensors, 11
CD optical pickup, 7, 8
continuous manufacturing
in-line lithography equipment, 12
printed electronics engineering, 12–13
printed products on flexible substrates, 11, 12
printing industry, 11
roll-to-roll manufacturing, 11, 12
decision tree, optimized micromanufacturing
blood glucose monitor, 23
sensor/actuator principle, 23
specifications, 24–25
different sensor fabrication, 19–20
disposable cassette, manufacturing technique, 35–36
fluid propulsion methods and MEMS integration
acoustic streaming, 31–32
comparison of, 29
electro-osmosis and iontophoresis, 32–33
mechanical pumps, 29–31
vacuum pressure reservoir, 33–34
heating and cooling, MEMS integration, 34
interdisciplinary teams, 5–6
Kittyhawk, 7, 8
magnetic read-write heads, 7, 8
mechanical and optical MEMS, 6
metal removal rate, 20, 22
microfluidic element, 8–10
micromechanisms, scaling characteristics, 25–26
miniaturization technology, 4–5
nanomachining and nanochemistry, 6
package inward design, 22–23
partitioning
in microfluidic instrument, 27–29
monolithic *vs*. hybrid MEMS, 26–27
photoformed glass, 8–10
polymer/metal-based biosensors
biomodules, 43
buried SU-8 microchannel, 44–45
chemical sensor array, 42, 43
integrated sensor array fabrication method, 42, 43
ion selective electrode array element fabrication, 43, 44
photoresist material properties, 44

sample injection systems, 34–35
serial manufacturing, 10
substrate and machining approach, checklist, 37–40
substrate materials
engineering guidelines, 37
material properties, 37, 38
performance comparison of, 36
sensor applications, 36
silicon-on-insulator (SOI) micromirror, 37
thin-*vs*. thick-film technology, 19
traditional *vs*. nontraditional, 14–18

O

Objective lenses, 294
Oil immersion microscopy, 294
Optical absorption instruments, 604–605
Optical microscopes (OMs)
dimensional metrology, 289, 295, 296
resolution, 330–331
Optical switches, 608–610
Optical Wyko white-light profilometers, 300, 301
Optochemical gas sensors, 600–601

P

Packaging
cap-on chip, 214–215
chemical etching, 200
dicing, 191–192
DNA, 186–187
front-end and back-end processes, 189–191
in ICs and MEMS, 187–189
laser drilling., 200
multichip (*see* Multichip module (MCM))
novel 3D structures, 201
single-device, 192–195
temperature gradient zone melting (TGZM), 200
Parallel microassembly
circuit integrity and LED activation, 229
electric field directed placement, LEDs, 228
fluid self-assembly, 226
microLEDs, 228
self-assembly, electrostatic levitation, 227
Parallel plate capacitors
dielectric/metallic slabs, 394
gravitational and electrostatic forces, 394
laterally driven actuator, 392
linear and side-drive motors, 393
parallel-and vertical-driven actuators, 392
stored energy density, 391
vertical and lateral resonators, 392–393
vertically driven polysilicon microbridge, 392
Paraxial approximation, 305
Passive sensors, 390
Phase-contrast microscopy, 295
Phosphatidylcholine, 99, 100
Photobleaching, 297
Piezoelectricity
acoustic wave sensors, 402
applications
acoustic imaging, 420–421
acoustic sensing, of mass, 417–420
motors, 414–416
pumping, 416–417
direct effect, 402
electrostriction, 403
ferroelectric materials, 403
ferroic materials, 403
inverse effects, 402
materials
aluminum nitride, 411
barium titanate, 411
physical properties, piezofilm, 411–412
polyvinylidene fluoride (PVDF), 411
properties, comparison, 413
PZT, 411
quartz, 410
Rochelle salt, 410
zinc-oxide, 412
mechanism of
constitutive equations and constants definitions, 405
coupling coefficient, 407
crystal systems, 404
elastic deformation, 406
electromechanical coupling coefficient, 405
interlocking behavior, 404
quartz, 403
Piezoelectric motors
advantages, 415–416
energy density, rotor/stator gap, 415
Panasonic ultrasonic motor, 414, 416
thin film PZT linear and rotary motors, 414, 416
Piezoelectric pumping, 416–417
Plasma arc machining (PAM), 20, 22
Plasma-enhanced chemical vapor deposition (PECVD) silicon nitride
advantage, 260
vs. LPCVD, properties, 260, 261
process conditions, 260, 261
Polyimide
cilia motion pixel, SEM, 258
nonphotosensitive macro applications, 257
photoresists, 258–259
photosensitive, 258, 259
planarization, 258
positive and negative aspects, 257
special perfluorinated, 259
Polymerase chain reaction (PCR)
chambers, 106–107
SmartCycler system, 107
temperature profile, 105–106
Thermus aquaticus, 105
Polymer MEMS
advantages, 257
polyimide (*see* Polyimide)
Porous anodic alumina
applications, 256
high-resolution micromachining, 255
humidity sensors, 256
lower-resolution microstructure preparation, 255
mechanical filters, 255–256
nanoporous alumina templates, CNT growth, 256
pore diameter, 254
Power, miniature devices
batteries
3D architectures, charge-insertion, 479–480
primary and secondary, 477
reversible, 479
specific energy and energy density, 477
thin-film, 479
capacitors, 480–481
combustion engines (*see* Heat engines, MEMS)
energy scavenging (*see* Energy harvesting)
energy storage, 476–477
fractals
C-MEMS, 511–512
definition, 506
dissipative structure, 508
in electrochemistry, 508–511
Hausdorff dimension, 507
in nature, 508
fuel cells (*see* Fuel cells)
solar cells (*see* Solar cells)
Pressure-sensitive adhesive (PSA) layers, 274–275
Pressure sensors
automotive
manifold absolute pressure sensor, 554
tire pressure monitoring systems, 555–556
in automotive industry, 551
comparison of, 553
diaphragm, 550
manufacturing of, 556–557
market revenue, 550–551
piezoresistive and capacitive, 553
resonant beam sensors
basic building blocks, 557, 558
digital pressure sensor, 558–559
features, 558
resonant silicon sensor, 559
structure, 557–558

silicon-MEMS, 551–552
suppliers of, 553–554
Prokaryotes, 57, 83–84
Proteins
emergence, 66
examples, 66
folding and unfolding, 102
making, 65–66
roles, 65
self-repair and recycling, 67–68
structural organizational levels, 68
structure, 67, 68
water release in, 66
Proton exchange membrane fuel cell, 357
PSA layers, *see* Pressure-sensitive adhesive (PSA) layers
Pyroelectricity
CMOS, 425
definition, 423
growth processes, 424
single-and multiple-element detectors, 424–425

Q

Quad flat pack (QFP), 185, 192–193
Quality factor, 398
Quantum actuators, 462–463
Quartz
machining methods, 242
oscillators, accuracies, 243
physical properties, 241–242
Quartz crystal microbalance (QCM), 408, 418

R

Radiofrequency-MEMS
acoustics, 612–613
compatibility, 611
consumer and IT applications, 613
electrostatic switches, 611
fabrication, 611–612
market, 612
quality factor, 612
Rayleigh-Bénard instability, 124
Rayleigh waves, *see* Surface acoustic waves (SAW) devices
Read-write heads, 606
Recombinant DNA technology, 78–79
Reproductive cloning, 79–80, 82–83
Reversible batteries, 479
Ribozymes, 70–71
Rotating ring gyroscope, 570
Rotaxene switch, 132

S

Salt-assisted spray pyrolysis (SASP), 159
Scaffolded DNA origami, 122–124
Scaling laws
allometry, 363
and diffusion, 376–377
dimensional analysis, 365–368
in electrostatics
breakdown voltage, 379
higher-energy densities, 381
isometric scaling, 378
mean free path, 380
parallel plate capacitor, 379
Paschen curve, 380
stored energy, 380
fractals, 364–365
mathematical approach, 362
metabolic heat production *vs*. body weight, 363–364
microintuition
NEMS devices, 361
objects and their linear size, 361–362
radioactive particle halflife, 361
Tom Thumb and Gulliver's Travels, 360–361
similarity/isometry, 362
of strength-weight ratio and inertia, 377–378
surface area *vs*. body mass, 363
in surface swimming, 371–374
in surface tension and adhesion, 374–376
surface-volume ratio, 369–371
thermal actuators, 456
trimmer's vertical bracket notation, 368–369
Scanning electron microscopy (SEM)
dimensional metrology
backscattered electrons (BSEs), 308
e-beam radiation, relative intensity *vs*. emission phenomena, 307
effective F-number, 305, 306
electron focusing, 305
electron-solid interaction, 306, 307
ESEM, 309
Everhart-Thorny (E-T) detector, 308
insulating and nonconductive samples, 304–305
magnification, 305
vs. OM, *Trochodiscus longispinus*, 306
paraxial approximation, 305
schematic drawing, 305
secondary electrons, 306
surface/microvolume examination, 304
vacuum system requirement, 304
resolution, 331
Scanning force microscopy, *see* Atomic force microscopy (AFM)
Scanning proximal microscopes, 309–310
Scanning tunneling microscope (STM)
dimensional metrology
electron density clouds, 315
inventors, 316–317
local density of states (LDOS), 314
operation modes, 316
piezoelectric material, 315
quantum mechanical uncertainty principle, 313
resolution, 315–316, 331–332
tungsten, 314–315
tunneling, 313–314
resolution, 331–332
Scanning white light interferometry (SWLI), 299, 300
Scatterometry, 287, 289
Selective deposition, diamond film synthesis, 252
Self-assembly
carbon flowers, 269
of dendrimer, 134
DNA-mediated assembly, 229–230
micro-origami, 230–232
nanotechnology, 122–124
scaffolded DNA origami, 122–124
vs. self-organization, 122
stochastic, 220
Self-organization, nanotechnology
bifurcation points, 125
flock formation, 126
fluctuations/noise, 125
hexagonal Bénard cell patterns, 124
Rayleigh-Bénard instability, 124
Self-organized epitaxial growth, nanotechnology
elastic strain, 150–151
film growth mechanisms, 150
Ge-Si, 150–151
Gibbs free energy, 148
heteroepitaxy technique, 147
InAs quantum dots, GaAs surface, 152
island formation, flat solid surfaces, 148–149
nanoparticle growth, vapor, 148
Pt growth, Pt(111), 151–152
quantum dots, pyramid-shaped recesses, 153
reflection high-energy electron diffraction (RHEED), 153
step flow growth, 150
Stranski-Krastanov growth mode, 151–152
surface tension, 149
Self-replicating robots, 130–131
Senscompatibility, 278–279
Sensors, *see* Actuators
Serial manufacturing, 10
Serial microassembly
AC electrokinetic effect, 226
bioparticles, 226
manipulation methods, 221, 224
optical tweezers, 224–225
single-beam gradient trap, 225
Shape memory alloy (SMA) actuators, 457–459

Silicon carbide (SiC)
films, 253
poly-SiC resonator fabrication process, 253
polytypes, 252–253
Silicon MEMS (Si-MEMS) resonators
high Q argument, 245
integration argument, 245
low cost argument, 244–245
low power consumption, 245
MEMS oscillator analysis, silicon *vs.* quartz, 244
multifrequency argument, 245
size argument, 244
Silicon nitride (Si_xN_y)
applications, 259
LPCVD nitride
etch characteristics, 261, 262
vs. PECVD, properties, 260, 261
PECVD nitride
advantage, 260
vs. LPCVD, properties, 260, 261
process conditions, 260, 261
uses, 259
Single-device packaging
chip scale packaging (CSP), 193–194
dual inline package (DIP), 193–194
leadframes, metal sheets, 192–193
methods, 195
pin grid arrays (PGAs), 193–194
quad flat pack (QFP), 192–193
stress release bellows, 195
TO-5 packaged accelerometer, 194
Single-stranded RNA, 62
Smart materials, 391
Soft superconductor, 450
Solar cells
bandgaps and efficiencies, 499–500
development, 497
energy diagram, 497–498
MEMS and NEMS, 501–504
multijunction, 498
photovoltaic (PV) effect, 496
polymer, 500
quantum efficiency, 497
research efficiencies, 500
silicon-based types, efficiency, 499
Solid-state gas sensors
arrays, electronic noses and tongues, 598–600
market, 597
MGS1100 carbon monoxide sensor, 598
operating principle, 597
Taguchi gas sensor, 597–598
thin-film micromachined gas sensors, 598
Solid-state processes, nanotechnology, 153–154
Steam reforming, 483
Stem cells
bioethics, 82–83
cell-reprogramming, 81–82
characteristics, 82
types, 81
Structural metrology
stylus profilometry, 324–325
transmission electron microscopy
bright-field (BF) mode, 326
dark-field imaging mode, 326
development, 325
diffraction mode, 326–327
electron path, thin specimen, 326
vs. slide projector, 326
spot location, Bragg's law, 327, 328
spot location, Bragg's law, 327, 328
Stylus profilometry, 324–325
Supercapacitors, 481
Superconductivity
ceramic materials, 451–452
critical temperature, 451
definition, 449
Dutch physicist, Heike Kammerlingh-Onnes, 449
lanthanium-barium-copper oxide, 450–451
LTS *vs.* HTS, 451–452
magnetic levitation, 452
Meissner effect, 450
mercury, 449–450
metal superconductors, 450–451
model (*see* BCS model)
partial Meissner effect, 451–452
single particle, 449
type I, 450
type II, 451
Supramolecular chemistry
aggregation, 134–135
dendrimer synthesis, 133
photocurrent generation, 132–133
rotaxene switch, 132
self-assembled dendritic structures, 134
unimolecular micelles, 135
weak interactions, 131
Surface acoustic waves (SAW) devices, 408–409
Surface mount technology (SMT), 193
Surface plasmon sensor
immunosensor, 584
measurements with, 584
Spreeta™, 584–585
surface plasmon resonance, 583
Surface swimmers, scaling, 371–374
Surface tension and adhesion, 374–376
Suspended membrane methods, 337–338
Suspension accelerometers, 561–563
SWLI, *see* Scanning white light interferometry (SWLI)

T

Taguchi-type sensor, 598
Tape automated bonding (TAB) method, 198
Telomeres, 77–78
Thermal actuators
parallel beam thermomechanical actuator, 455
scaling of, 456
types of (*see* Thermomechanical actuators)
Thermoelectricity, 425–428
Thermomechanical actuators
bimetallic, 456–457
shape memory alloy (SMA), 457–459
thermopneumatic, 456
Thermopneumatic actuators, 456
Thin-film batteries, 479
Thin films
conductivity and carrier type
four-point probe measurement arrangement, 323
hot-point probe operation, 323, 324
material resistors, 322
semiconductor resistors, 322
sheet resistance, 322–323
in situ stress-measuring techniques
disk method, 333–334
dog-bone method, 334–337
Poisson ratio, 338
suspended membrane methods, 337–338
stress nonuniformity measurement
cantilever beams, 340–341
cantilever spirals, 341, 342
surface micromachined structures
clamped-clamped beams, 338
lateral resonators, 340
ring crossbar structures, 338
vernier gauges, 338–340
Thin-*vs.* thick-film technology, 19
Tire pressure monitoring systems (TPMS), 555–556
TO-5 packaged accelerometer, 194
Transparent ceramic nanoparticles, 143–144
Transposomes, 75–76
Two-photon microscopy, 298–299
Type II superconductors, 451
Type I superconductor, 450

U

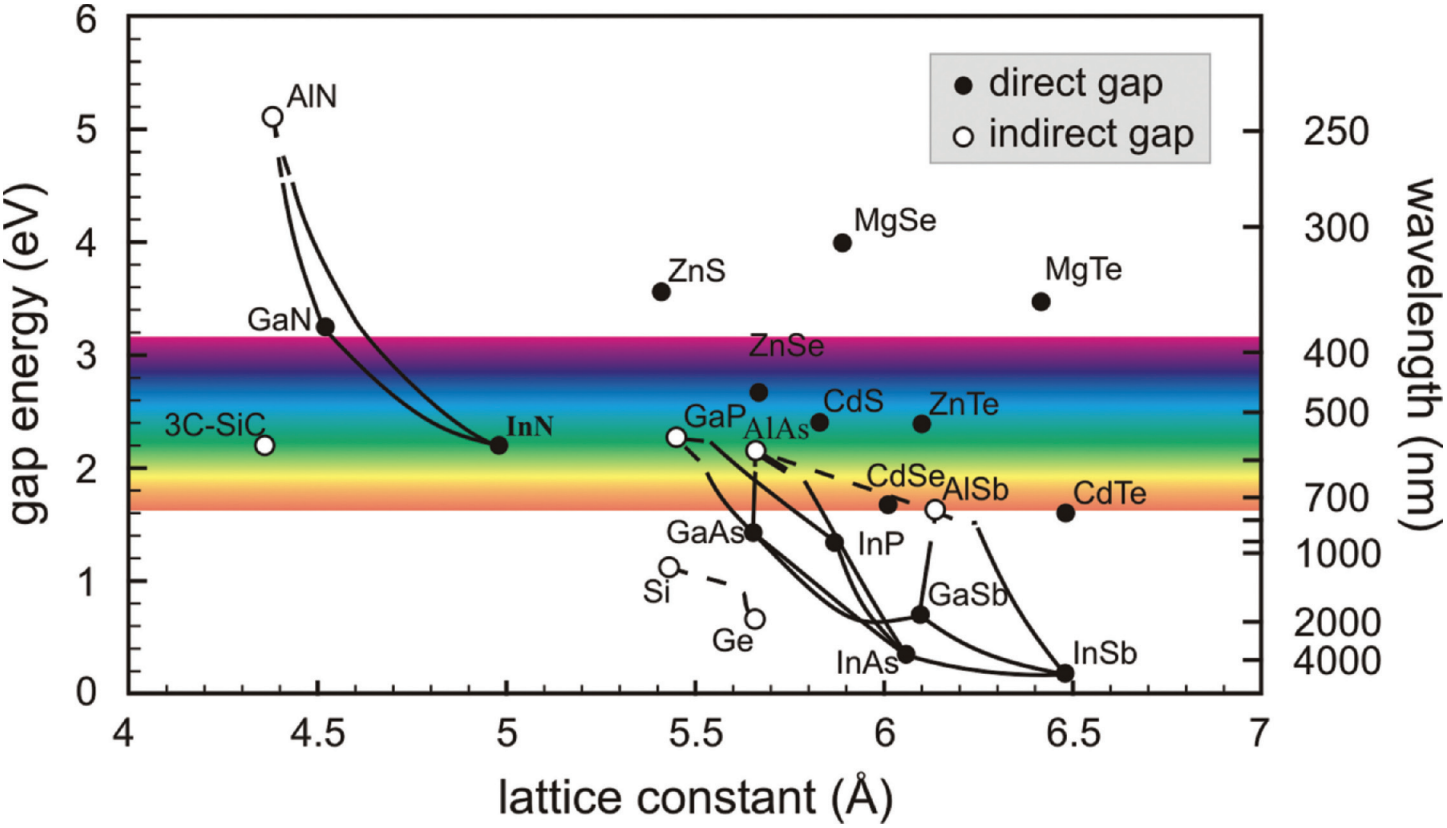
Ultracapacitors, 481
Universal genetic code, 113

V

Vacuum gripper, 223–224
Vacuum pressure reservoir, 33–34
Vibrating ring gyroscope, 570–571

W

Watson-Crick DNA base pairing, 63
Wire bonding method, 197–198


direct gap
indirect gap
AlN
GaN
3C-SiC
InN
ZnS
MgSe
MgTe
ZnSe
CdS
ZnTe
GaP
AlAs
CdSe
AlSb
CdTe
GaAs
InP
Si
Ge
GaSb
InAs
InSb
gap energy (eV)
lattice constant (Å)
wavelength (nm)
0
1
2
3
4
5
6
4
4.5
5
5.5
6
6.5
7
250
300
400
500
700
1000
2000
4000

Frequently used constants

Quantity	Symbol	Value	Unit	Relative std. uncert. u_r
speed of light in vacuum	c, c_0	299 792 458	m s^{-1}	(exact)
magnetic constant	μ_0	$4\pi \times 10^{-7}$	N A^{-2}	
		$= 12.566\,370\,614\ldots \times 10^{-7}$	N A^{-2}	(exact)
electric constant $1/\mu_0 c^2$	ϵ_0	$8.854\,187\,817\ldots \times 10^{-12}$	F m^{-1}	(exact)
Newtonian constant of gravitation	G	$6.67428(67) \times 10^{-11}$	m^3 kg^{-1} s^{-2}	1.0×10^{-4}
Planck constant $h/2\pi$	h	$6.626\,068\,96(33) \times 10^{-34}$	J s	5.0×10^{-8}
	$\hbar$	$1.054\,571\,628(53) \times 10^{-34}$	J s	5.0×10^{-8}
elementary charge	ϵ	$1.602\,176\,487(40) \times 10^{-19}$	C	2.5×10^{-8}
magnetic flux quantum $h/2e$	Φ_0	$2.067\,833\,667(52) \times 10^{-15}$	Wb	2.5×10^{-8}
conductance quantum $2e^2/h$	G_0	$7.748\,091\,7004(53) \times 10^{-5}$	S	6.8×10^{-10}
electron mass	m_e	$9.109\,382\,15(45) \times 10^{-31}$	kg	5.0×10^{-8}
proton mass	m_p	$1.672\,621\,637(83) \times 10^{-27}$	kg	5.0×10^{-8}
proton-electron mass ratio	m_p/m_e	1836.152 672 47(80)		4.3×10^{-10}
fine-structure constant $e^2/4\pi\epsilon_0\hbar c$	α	$7.297\,352\,5376(50) \times 10^{-3}$		6.8×10^{-10}
inverse fine-structure constant	α^{-1}	137.035 999 679(94)		6.8×10^{-10}
Rydberg constant $\alpha^2 m_e c/2h$	R_∞	10 973 731.568 527(73)	m^{-1}	6.6×10^{-12}
Avogadro constant	N_A, L	$6.022\,141\,79(30) \times 10^{23}$	mol^{-1}	5.0×10^{-8}
Faraday constant $N_A e$	F	96 485.3399(24)	C mol^{-1}	2.5×10^{-8}
molar gas constant	R	8.314 472(15)	J mol^{-1} K^{-1}	1.7×10^{-6}
Boltzmann constant R/N_A	k	$1.380\,6504(24) \times 10^{-23}$	J K^{-1}	1.7×10^{-6}
Stefan-Boltzmann constant $(\pi^2/60)k^4/\hbar^3 c^2$	σ	$5.670\,400(40) \times 10^{-8}$	W m^{-2} K^{-4}	7.0×10^{-6}
Non-SI units accepted for use with the SI				
electron volt: (e/C) J	eV	$1.602\,176\,487(40) \times 10^{-19}$	J	2.5×10^{-8}
(unified) atomic mass unit				
$1\,\mathrm{u} = m_u = \frac{1}{12} m(^{12}\mathrm{C})$	u	$1.660\,538\,782(83) \times 10^{-27}$	kg	5.0×10^{-8}
$= 10^{-3}$ kg mol^{-1}/N$_A$				